

# RuleReasoner: Reinforced Rule-based Reasoning via Domain-aware Dynamic Sampling

Yang Liu<sup>\*</sup>, Jiaqi Li<sup>\*</sup> and Zilong Zheng<sup>✉</sup>

NLCo Lab, Beijing Institute for General Artificial Intelligence

{liuyang, lijiaqi, zlzheng}@bigai.ai

Rule-based reasoning has been acknowledged as one of the fundamental problems in reasoning, while deviations in rule formats, types, and complexity in real-world applications pose severe challenges. Recent studies have shown that large reasoning models (LRMs) have remarkable reasoning capabilities, and their performance is substantially enhanced by reinforcement learning (RL). However, it remains an open question *whether small reasoning models (SRMs) can learn rule-based reasoning effectively with robust generalization across diverse tasks and domains*. To address this, we introduce **Reinforced Rule-based Reasoning**, *a.k.a.* **RULEREASONER**, a simple yet effective method to conduct rule-based reasoning via a wide collection of curated tasks and a novel domain-aware dynamic sampling approach. Specifically, RULEREASONER resamples each training batch by updating the sampling weights of different domains based on historical rewards. This facilitates domain augmentation and flexible online learning schedules for RL, obviating the need for pre-hoc human-engineered mix-training recipes used in existing methods. Empirical evaluations on in-distribution (ID) and out-of-distribution (OOD) benchmarks reveal that RULEREASONER outperforms frontier LRMs by a significant margin ( $\Delta 4.1\%$  average points on eight ID tasks and  $\Delta 10.4\%$  average points on three OOD tasks over OpenAI-o1). Notably, our approach also exhibits higher computational efficiency compared to prior dynamic sampling methods for RL.



**Code:** <https://github.com/bigai-nlco/RuleReasoner>



**Model:** <https://huggingface.co/RuleReasoner>

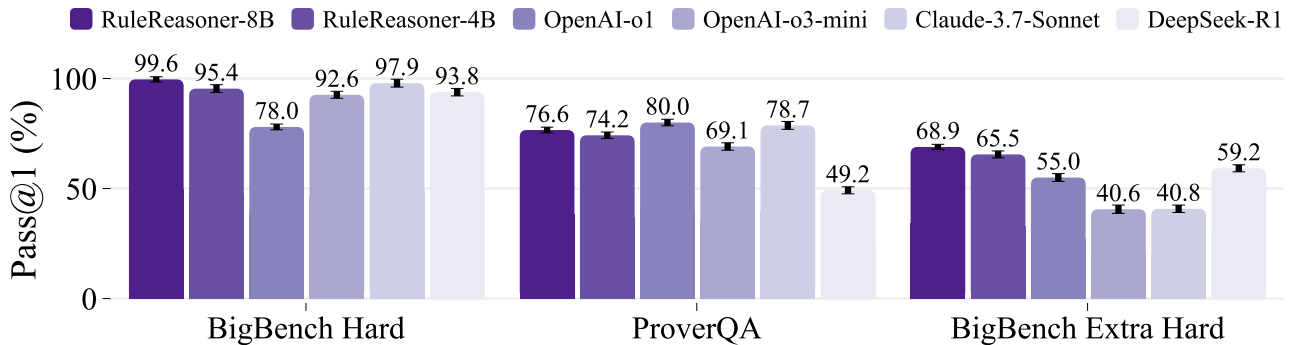


Figure 1: Out-of-distribution performance comparison between RULEREASONER (8B and 4B) and other frontier reasoning models on challenging rule-based reasoning benchmarks.

<sup>\*</sup> Equal Contributions.

<sup>✉</sup> Corresponding author(s): Zilong Zheng.

## 1. Introduction

Rule-based reasoning (Xu et al., 2024a; Wang et al., 2024a; Servantez et al., 2024a; Morishita et al., 2024; Wang et al., 2024b) is an ability to drawing new conclusions or making decisions based on provided facts and predefined logical rules, which requires strong model ability on reasoning. It emulates human reasoning and mirrors the structured deductive processes that humans employs in various domains like mathematics, law, and medical diagnostics (Liu et al., 2023a; Xiong et al., 2024). The needs of rule-based reasoning increasingly grows in scenarios requiring transparency, interpretability, and adherence to domain constraints. Moreover, the deviations of rules for different scenarios lead to significant changes on the reasoning process, which requires more controllable and adaptable reasoning capability under ever-changing circumstances (Saparov et al., 2023; Tang et al., 2023a).

Recent work has demonstrated the remarkable reasoning capabilities of large reasoning models (LRMs) with an intermediate thinking process, chain-of-thought (CoT) (Wei et al., 2022a), notably the long thinking ability stimulated by reinforcement learning (RL) (Xie et al., 2025; Meng et al., 2025; Guo et al., 2025; Hu et al., 2025). However, conventional approaches rely closely on scaling to larger model sizes or supervision distilled from superior models. On the other hand, it is more challenging for small language models (SLMs) with relatively weak in-context learning and instruction-following abilities to understand and associate rules and facts provided in the context (Brown et al., 2020; Srivastava et al., 2025; Li et al., 2025a; Guan et al., 2025). In this work, we explore whether the capability of rule-based reasoning for SLMs can be learnt and enhanced effectively. Furthermore, we are curious whether this ability can be generalized across various tasks and reasoning paradigms with sufficient interpretability.

With this in mind, this work aims to investigate **whether it’s feasible and effective to enhance the rule-based reasoning competence of SLMs using RL**. We also intend to further validate that this incentivized capability can generalize to unseen logical reasoning tasks, forms, and difficulties, achieving comparable performance to or even outperforming proprietary reasoning models. Therefore, in this paper, we propose **RULEREASONER**, which bridges the gap between rule-based reasoning and SLMs. This leverages the advantage of existing RL approaches and further enhances its critical limitations of sample inefficiency, and rigid rule application in dynamic contexts and tasks. More specifically, RULEREASONER leverages the task reward at each training step to estimate domain weights without requiring any human prior knowledge or significant computational overhead from repeated rollouts. First, RULEREASONER initializes to train a language models in a standard Reinforcement Learning with Verifiable Rewards (RLVR) way. Second, within the current training iteration, RULEREASONER updates the domain weights calculated by the associated task rewards (*i.e.*, historical rewards) from prior training steps in current iteration. Finally, RULEREASONER samples a batch based on these domain weights to perform policy optimization using a given policy gradient algorithm.

**Our innovations** are summarized as three folds:

- **Comprehensive Rule-centric Data Curation:** We curate a large and diverse dataset that spans eight rule-based reasoning tasks with explicit or implicit contextual rules tied to each question. These rules vary in format (explicit versus implicit), reasoning forms (deductive versus inductive), and complexity (multiple reasoning depths), enabling systematic training and evaluation in generalizable rule application rather than memorization.
- **RLVR on Rule-based Reasoning:** We design a novel RLVR framework that harnesses rule-based verification to produce outcome rewards, enabling models to learn structural reasoning paths. Unlike supervised learning (SL), RLVR encourages exploring and exploiting valid reasoning steps instead of imitation, improving generalization to unseen rules.
- **Domain-aware Dynamic Sampling:** To harmonize SLMs’ proficiencies across imbalanced domains, we present an adaptive sampling algorithm that dynamically reweights training domains based on their degree of under-optimization. This ensures balanced learning dynamics across tasks, enhancing both in-distribution (ID) and out-of-distribution (OOD) reasoning performance. Our method is intuitively compatible with a variety of RLVR algorithms for enhancement.

We perform extensive experiments and evaluations on RULEREASONER, introducing two best rule-

based reasoning models: RULEREASONER-8B and RULEREASONER-4B. Our empirical results show that RULEREASONER significantly improves both training efficiency and task performance: (1) As shown in Figure 1, RULEREASONER-8B outperforms OpenAI-o1, Claude-3.7-Sonnet, and DeekSeek-R1, achieving higher performance than other strong RLVR methods with fewer training steps. Specifically, RULEREASONER-8B achieves  $\Delta 14\%$  and  $\Delta 49\%$  OOD pass@1 accuracy respectively over o1 and the original base model. (2) RULEREASONER-4B further demonstrates that SLMs can effectively learn rules even with a smaller model size, achieving an average pass@1 of 78.3% on three OOD benchmarks ( $\Delta 7.3\%$  over o1). (3) Surprisingly, RULEREASONER achieves comparable task performance with notably fewer training steps than other RLVR methods. This suggests that RULEREASONER not only enhances ID and OOD performance but also increases sample utilization, leading to improved training efficiency.

## 2. Background and Related Work

Reinforcement learning has played a critical role in improving the reasoning capabilities of large language models (Silver and Sutton, 2025), particularly through approaches such as RLVR (Guo et al., 2025; Li et al., 2025b; Zuo et al., 2025). In this section, we introduce the task formulation of rule-based reasoning (§2.2) and briefly discuss key components of prior RLVR methods (§2.3) and their limitations on training sample efficiency.

### 2.1. Related Work

It has demonstrated that the large language models can conduct rule-based reasoning including rule grounding and implementation with promising performance (Zhu et al., 2023; Servantez et al., 2024b). Early works in rule-based reasoning start from predefined rules in symbolic forms with an emphasis on scalability and compositionality in specific tasks (Tang et al., 2023b; Luo et al., 2024; Jia et al., 2024; Gui et al., 2024; He et al., 2025) while recent works are dedicated to perform rule-based reasoning in natural language that are more applicable for real scenarios (Zhou et al., 2024; He et al., 2024; Tang et al., 2024). It is also worth noting that recent advances in logical reasoning, such as Logic-RL (Xie et al., 2025), are generally considered rule-free for reasoning, which differs from our task formulations and experimental settings. These methods explore the potential of rule learning through diverse prompting method (Diallo et al., 2025; Peng et al., 2024), supervised distillation (Wang et al., 2024c), and external memory augmentation (Wang et al., 2024d;b). However, they spend less effort adapting the reasoning capability of LLMs to unseen tasks with limited task types and formats. Inspired by recent advancements in RLVR methods focused on mathematical reasoning and code generation (Zhang et al., 2025; Chen et al., 2025; Wei et al., 2025; Li et al., 2025c; Zhao et al., 2025; Li et al., 2025d), we further optimize their limitation on data efficiency with dynamic data sampling along with a curated collection of diverse rule-based reasoning training data. Our method improves model performance across both ID and OOD reasoning tasks with higher generalization and computational efficiency.

### 2.2. Preliminaries: Rule-based Reasoning

Given a question, a set of sentences consisting of relevant **facts** and associated **rules** as context, the model is asked to answer the question by reasoning and applying the given rules and facts. We refer the rules in this paper as *contextual logic rules* in this paper, which are expressed in natural language and specifically given for each question (which may differ for each problem). The provided rules can either explicitly or implicitly generated as principles or premises for solving the question. For a grouped domain of datasets  $\mathcal{D}$  consisting of various domains  $\{d_1, d_2, \dots, d_n\}$ , we have  $\mathcal{D} = \{q, r, y\}$ , where  $q$  is a question,  $r$  is a reasoning trajectory, and  $y$  is a verifiable answer.

### 2.3. On-policy Reinforcement Learning

**Reward Shaping.** To teach models to learn reasoning, we design a rule-based exact match (EM) reward function to evaluate the response according to the final answer, ensuring both the correctness of the

answers and the adherence to the format. We define  $(q, \hat{y}) \sim \mathcal{D}$ ,  $y \sim \pi_\theta(\cdot|q)$ , and

$$\mathcal{R}_{\text{EM}}(\hat{y}, y) = \begin{cases} 1 & \text{is\_equivalent}(\hat{y}, y), \\ -1 & \text{otherwise.} \end{cases} \quad (1)$$

**Policy Optimization.** We adopt the basic form of GRPO (Shao et al., 2024) but discard the part of KL loss, encouraging the model to explore various solutions. For each question-answer pair  $(q, y)$ , the policy model  $\pi_{\theta_{\text{old}}}$  samples to generate a group of responses  $\{y_1, y_2, \dots, y_G\}$  and calculates the associated rewards  $\{r_1, r_2, \dots, r_G\}$ , given the oracle answer  $y$ , using the aforementioned reward function  $\mathcal{R}_{\text{EM}}$ .

$$\mathcal{J}(\theta) = \mathbb{E}_{(q, y) \sim \mathcal{D}, \{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[ \frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \left( \min(r_{i,t}(\theta) A_{i,t}, \text{clip}(r_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon) A_{i,t}) \right) \right], \quad (2)$$

where  $r_{i,t}(\theta)$  is the rate of importance sampling for domain  $d_i$  at the  $t$ -th token for  $y_i$ , and  $A_i$  is the advantage as the critic obtained by normalizing the rewards within each group. We strictly follow the on-policy training method, performing only one gradient update after the policy model  $\pi_{\theta_{\text{old}}}$  generates a group of  $G$  rollouts, to enable stable RL training and prevent entropy collapse.

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} | q, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} | q, y_{i,<t})}, \quad A_i := \tilde{r}_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (3)$$

**Limitations of RLVR on Data Efficiency.** Though current RLVR elicits the long chain-of-thought reasoning ability based on the policy gradient RL algorithm like PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024), the efficiency of training data for RLVR remains relatively unexplored. Existing works like DAPO (Yu et al., 2025) oversamples and filters out prompts with the accuracy equal to 1 and 0 to enhance training efficiency. However, it does not push the limits of training efficiency due to the large recompute cost in the rollout stage. SRPO (Zhang et al., 2025a) shows the gains via epoch-level re-sampling with RLVR without exploring the agile sampling methods for fine-grain control. Moreover, ADARFT (Shi et al., 2025) explores an efficient batch-level sampling method using curriculum learning; however, it relies on human priors or an empirical success rate by models on sample difficulty. In the following sections, we expend great effort to further leverage training examples to achieve higher reasoning performance effectively.

### 3. Domain-aware Policy Optimization with Dynamic Sampling

In online data sampling, increasing the number of samples in an individual domain improves its density but reduces the relative proportion of others, potentially harming performance in less-represented domains without timely control (Albalak et al., 2024) or causing obvious tradeoff across domains (Xie et al., 2023).

To address similar issues in RLVR, as shown in Alg. 1 and Figure 2, we propose **Domain-aware Dynamic Sampling (DADS)**, a simple yet effective sampling method for RLVR, aiming to improve the performance of a policy model  $\pi_\theta$  for solving multi-domain rule-based reasoning

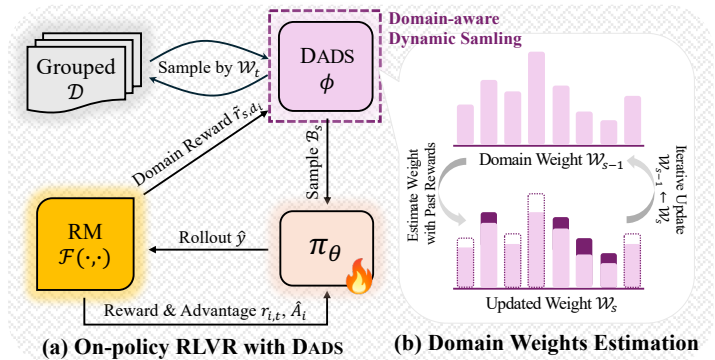


Figure 2: Diagram of RULEREASONER training recipe.



tasks. DADS dynamically adjusts the probability of sampling data from different domains based on their historical rewards. By prioritizing domains that yield lower verifiable rewards or those lagging behind a target reward, DADS enhances sample efficiency to re-sample the training batch  $\mathcal{B}_s$  and mitigates the domain imbalance issue, leading to faster and more stable learning of policies that satisfy reward specifications. We instantiate RULEREASONER with the gradient policy algorithm of GRPO variant in this work to demonstrate its effectiveness and efficiency.

### 3.1. Domain-aware Dynamic Sampling (DADS)

**Domain-aware Rewards.** At each training step  $s$ , to evaluate the proficiency for a domain  $d_i \in \mathcal{D}$ , we define  $\bar{r}_{d_i,s}$  as the algebraic mean, calculated by domain-aware rewards  $\{r_{s,d_i,j}\}_{j=1}^m \sim \mathcal{R}_{\text{EM}}(y, \hat{y})$  of  $m$  previous training samples in the domain, which correspond to generations  $\mathcal{Y}_{s-1} : \{y_{s,d_i,j}\}_{j=1}^m \sim \pi_\theta(\cdot|q)$  and the set of ground truth  $\hat{\mathcal{Y}}_{s-1}$ . Note that  $m$  may vary across different domains and training steps due to the batch-level domain sampling strategy. Domain-aware rewards calculation over batch (Alg. 1, line 6 and 7) is computed as:  $\bar{r}_{s,d_i} = \frac{1}{m} \sum_{j=1}^m r_{s,d_i,j}$ . We employ a target reward,  $r_{\text{target}} = 1$ , to define the upper bound for the underoptimization estimation,  $v_{s,d_i}$ , of a domain. This estimation is calculated as  $\max\{0, r_{\text{target}} - \tilde{r}_{s,d_i}\}$ . The target reward quantifies the extent to which a domain's performance lags behind the desired level.

**Decaying Importance Sampling.** Furthermore, given the utilization of past rewards for domain  $d_i$ , we introduce a decaying importance-sampling strategy, which employs the exponentially weighted moving average (Holt, 2004) that considers both current and the historical estimated rewards. The historical rewards  $\{\tilde{r}_{s-1,d_i}\}_{i=1}^n$  are involved with the smoothing factor  $\alpha$  to produce normalized rewards  $\{\tilde{r}_{s,d_i}\}_{i=1}^n$ . We have  $\tilde{r}_{s,d_i} = \alpha \tilde{r}_{s-1,d_i} + (1 - \alpha) \bar{r}_{s,d_i}$ , where  $\alpha$  serves as a smoothing factor that creates a more stable estimate of the performance for a domain over time, rather than relying solely on the most recent reward  $\bar{r}_{s,d_i}$ .

**Domain Re-weighting.** Consequently, we establish a domain weight,  $w_{s,d_i}$ , which is then normalized using a standard softmax function (as detailed in Alg. 1, lines 11 to 13). In this normalization, hyperparameters  $\tau$  and  $\epsilon$  are used:  $\epsilon$  ensures a minimum sampling weight for all domains, even well-learned ones, and  $\tau$  adjusts how strongly the sampling prioritizes domains based on their rewards. We obtain the re-sampling weights  $\mathcal{W}_s := \{w_1, w_2, \dots, w_n\}$  across domains. Then, we use  $\mathcal{W}_t$  to construct a new training batch  $\mathcal{B}_s$  for the subsequent policy optimization iteration. The process will be implemented iteratively during the training and more details are described in Algorithm 1.

### 3.2. Logical Rules Data Curation

Considering the diverse and imbalanced nature of various rule-based reasoning tasks, we follow the **principles** below for the training data collection to take into account the ID and OOD performance:

---

#### Algorithm 1 Domain-aware Dynamic Sampling

---

**Input:** Policy model:  $\pi_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ ;  
 Reward model:  $\mathcal{R}_{\text{EM}}(\cdot, \cdot) : \mathcal{Y}, \hat{\mathcal{Y}} \rightarrow \{0, 1\}$ ;  
 Last weight:  $\mathcal{W}_{s-1} := \{w_1, w_2, \dots, w_n\}$ ;  
 Grouped domain:  $\mathcal{D} := \{d_1, d_2, \dots, d_n\}$ ;  
 Hyperparameters:  $\{\alpha, \epsilon, \tau\} \subset \mathbb{R}^+$ .  
**Output:** Constructed batch of samples:  $\mathcal{B}_s$ .

- 1: **procedure** TRAIN STEP  $s$  SAMPLING
- 2:   **Initialize:**  $\mathcal{B}_{s-1} \leftarrow \mathcal{W}_{s-1} \times \mathcal{D}$ ;  $r_{\text{target}} \leftarrow 1$ .
- 3:    $\mathcal{Y}_{s-1} \leftarrow \pi_\theta(\mathcal{B}_{s-1})$  ▷ **ROLLOUT**
- 4:    $\{\{r_{s,d_i,j}\}_{j=1}^m\}_{i=1}^n \leftarrow \mathcal{R}_{\text{EM}}(\mathcal{Y}_{s-1}, \hat{\mathcal{Y}}_{s-1})$
- 5:   /\* Update estimated rewards \*/
- 6:    $\{\bar{r}_{s,d_i}\}_{i=1}^n \leftarrow \{\frac{1}{m} \sum_{j=1}^m r_{s,d_i,j}\}_{i=1}^n$
- 7:    $\{\tilde{r}_{s,d_i}\}_{i=1}^n \leftarrow \{\alpha \tilde{r}_{s-1,d_i} + (1 - \alpha) \bar{r}_{s,d_i}\}_{i=1}^n$
- 8:   /\* Calculate weights by rewards \*/
- 9:   **for**  $i = 1, 2, \dots, n$  **do**
- 10:      $v_{s,d_i} \leftarrow \max\{0, r_{\text{target}} - \tilde{r}_{s,d_i}\}$
- 11:      $w_{s,d_i} \leftarrow \exp((v_{s,d_i} + \epsilon) / \tau)$
- 12:    $\mathcal{W}_s := \{w_{s,d_i}^{\text{norm}}\}_{i=1}^n$  ▷ **NORMALIZING**
- 13:      $= \{w_{s,d_i} / \sum_{j=1}^n w_{s,d_j}\}_{i=1}^n$
- 14:   /\* Re-sample w.r.t. optimized weights \*/
- 15:    $\mathcal{B}_s \leftarrow \mathcal{W}_s \times \mathcal{D}$  ▷ **SAMPLING BY  $\mathcal{W}_s$**
- 16:   **return**  $\mathcal{B}_s$

---

Table 1: Data statistics of curated tasks and data. † denotes it can be deemed as deduction reasoning since we provide rules for it explicitly. The abbr. in the table indicate Modus Ponens (MP), Universal Instantiation (UI), Hypothetical Syllogism (HS), Disjunctive Syllogism (DS), Modus Tollens (MT), respectively. “FOL”, “AR”, “CS”, and “CCR” denotes First-Order Logic, Analytical Reasoning, Constraint Satisfaction, and Categorical & Conjunctive Reasoning, respectively. “MC” represents multiple choice.

Dataset	# Train/Test	Task Format	Source	Reasoning Form	Reasoning Depth	Fiction Rule	Rule of Inference
ProofWriter (Tafjord et al., 2021)	7,997/500	Boolean	Synthetic	Deduction	0 ~ 5	✓	MP, UI
ProntoQA (Saparov and He, 2023)	8,000/500	Boolean	Synthetic	Deduction	1,3,5	✓	UI, Conjunction Simplification
Clutrr (Sinha et al., 2019)	268/67	Free Text	SemiSynthetic	Induction <sup>†</sup>	—	✗	HS
FOLIO (Han et al., 2024)	1,208/242	MC	Manual	FOL	0 ~ 7	✓	MT, DS, UI
LogicNLI (Tian et al., 2021)	8,000/500	MC	SemiSynthetic	FOL	1 ~ 5	✓	MP, MT
AR-LSAT (Zhong et al., 2022)	1,636/410	MC	RealWorld	AR	—	✓	MP, MT
Logic. Dedu. (Xu et al., 2024b)	1,200/300	MC	Manual	CS	1,3,5	✓	MP, MT
LogiQA (Liu et al., 2023b)	264/67	MC	Real-world	CCR	—	✓	MP, MT

- **Varying Reasoning Depths and Forms.** We collect data with reasoning depths varying from 0 to 7 hops to learn in a way of curriculum learning (Bengio et al., 2009) for different levels of rule composition and reasoning complexity. Besides, we cover a wide range of logical reasoning forms (e.g., deductive, inductive, and analytical reasoning), which encourages the model to handle various reasoning types and generalize its capability to diverse unseen tasks.
- **Different Rule Formats.** Our data provides rules explicitly or implicitly as inferred premises/constraints. This diverse input, including redundant or confusing rules, improves the model’s flexibility in recognizing, parsing, and applying rules across contexts.
- **Application on Multiple Rules of Inference.** Applying rules of inference is essential for deriving valid, consistent, and reliable conclusions or decisions. The diversity of rules of inference implied in the dataset facilitates learning dynamic rule employment and length generalization further.
- **Dependency on Commonsense or Context.** Our focus is on applying contextual rules adaptively, tailored to different questions. Correctly answering often requires more than simply memorizing seen rules or reasoning, or relying on common sense or existing knowledge.
- **Friendly and Robust to Reward Design and Evaluation.** Our task selection prioritizes boolean and multiple-choice questions over free text, as they are more conducive to obtaining rule-based outcome rewards and precise evaluation results.

These principles are critical for ensuring that training data captures the complexity and diversity inherent in rule-based reasoning tasks. The statistics of the training data are presented in Table 1.

## 4. Empirical Results

### 4.1. Experiment Setup

**Datasets and Benchmarks.** Besides the above mentioned training sets, we assess the models generalization on unseen tasks using subsets from BigBench Hard (Suzgun et al., 2023), BigBench Extra Hard (Kazemi et al., 2025), and ProverQA (Qi et al., 2025), as detailed in Table 2. More details regarding the training datasets are presented in Appendix §C.1.

**Compared Baselines.** We include four types of baselines: 1) **Prior rule-based reasoners (RBRs):** Hypotheses-to-Theories (Zhu et al., 2023), Logic-of-Rule (Servantez et al., 2024b), and Rule-Guided Feedback (Diallo et al., 2025); 2) **Frontier reasoners:** OpenAI-o1 (o1-2024-12-17) (Jaech et al., 2024), o3-mini (o3-mini-2025-01-31) (Zhang et al., 2025b), DeepSeek-R1 (Guo et al., 2025), and Claude-3.7-Sonnet (claude-3-7-

Table 2: OOD benchmarks statistics.

OOD Test	Examples	Levels
BBH	750	✓
BBEH	400	✓
ProverQA	1,500	✓

Task	Context (Explicit or Implicit Rules)	Question	Answer
ProofWriter	<b>RULES:</b> <i>If the bear needs the dog and the dog visits the bear then the bear likes the cat. If something is rough then it likes the dog</i> <b>FACTS:</b> <i>The bear is round. The bear visits the cat.</i>	The bear needs the cat?	True
ProntoQA	<b>RULES:</b> <i>Everything that is earthy and a wumpus is an impus. Everything that is dull and a brimpus is a numpus</i> <b>FACTS:</b> <i>Sally is dull. Sally is a brimpus.</i>	Sally is dull and a brimpus?	True
Clutrr	<b>RULES:</b> <i>If B is the son of A, and C is the grandmother of B, then C is the mother of A.</i> <b>FACTS:</b> <i>Pedro is taking his wife Dorothy out to dinner for their date tonight. Tracy loves cooking for her son. Tracy went to the store with her sister Dorothy.</i>	How is Shantel related to Pedro?	Shantel is the mother -in-law of Pedro.
LogicNLI	<b>RULES:</b> <i>All not fierce people are not brainy. If there is at least one people who is not intelligent, then Keaton is fragile and Jaime is fierce.</i> <b>FACTS:</b> <i>Jaime is fragile. Philip is not sociable. Jaime is brainy.</i>	Landon is not intelligent.	Entailment
FOLIO	<i>Rafa Nadal was born in Mallorca. Rafa Nadal is a professional tennis player. Nadal's win ratio is higher than 80%. All players in the Big 3 are professionals who have a high win ratio.</i>	Nadal was not born in Mallorca.	False
Logical Deduction	<i>On a shelf, there are five books: a blue book, a red book, a purple book, a gray book, and a white book. The white book is to the right of the gray book. The blue book is the leftmost. The red book is to the left of the gray book. The red book is the third from the left.</i>	Which of the following is true? A) The blue book is the second from the right. B) ... C) ...	D
AR-LSAT	<i>Eight new students—R, S, T, V, W, X, Y, Z—are being divided among exactly three classes—class 1, class 2, and class 3. Classes 1 and 2 will gain three new students each; class 3 will gain two new students.</i>	If T is added to class 3, which one of the following is a student who must be added to class 2?	C
LogiQA	<i>Xiao Ming forgot what day it was today, so he asked O, P, and Q. O replied I also forgot what day it is today, but you can ask P and Q both. P replied Yesterday It's the day when I lied. Q's answer is the same as P. It is known that I.O never lied;</i>	What day is today? A) Monday B) Tuesday C) Thursday D) Sunday	C

Figure 3: Demonstration of the curated training data examples.

sonnet-20250219 with extended thinking mode) (Anthropic, 2025) with standard step-by-step zero-shot CoT prompting (Wei et al., 2022a); **3) Behavioral cloning** (Pomerleau, 1988)<sup>1</sup>: SFT without CoT (Wei et al., 2022b), SFT with short CoT (Yeo et al., 2025), and SFT with distilled long CoT (Yeo et al., 2025); **4) Advanced RLVRs:** we compare the recent RLVR approaches to show our effectiveness, including GRPO (Shao et al., 2024), Dr. GRPO (Liu et al., 2025), and DAPO (Yu et al., 2025). Due to RULEREASONER only adding negligible wall-clock time during batch sampling, we use DAPO without its dynamic sampling to exclude additional computational overhead for fair comparison purposes.

**Evaluation Metrics.** All tasks in the work are evaluated using the algebraic mean of hard exact match, which is also equivalent to pass@1 accuracy under strict extraction and comparison.

**Implementation Details. Training Setup:** We use Qwen3 base models (4B and 8B)<sup>2</sup> (Yang et al., 2024a;b) as our base models and employ veRL (Sheng et al., 2024) for RL and supervised post-training. In RL, we set train and mini batch sizes to 64 for strict on-policy updates, and a rollout size of 64 per question. For hyperparameters in DADS, we use a  $\tau$  of 0.5 for moderately frequent domain weight updates and an  $\epsilon$  of 0.1 for minimum sampling probability per domain, with a smoothing factor  $\alpha$  of 0.5. **Inference Setup:** We employ random sampling (temperature  $\tau = 0.6$  and  $top-p = 0.95$  with a maximum output length of 2,048. For behavioral cloning and RLVR baselines, we use Qwen3-8B-Base for full-parameter SFT or online RL. We perform five runs per test set and calculate the mean and standard deviation of the performance. We also report one-sided p-values for statistical significance.

## 4.2. RULEREASONER Improves RLVR Performance and Efficiency

**In-Distribution Performance.** As shown in Table 3, we first report the in-domain tasks performance to depict the effectiveness of RULEREASONER. Compared with the cutting-edge LRMs, RULEREASONER-8B surprisingly outperforms with a large performance gap. Notably, on eight ID tasks, OpenAI-o1 lags behind RULEREASONER-8B with 4.1% point, whereas Claude-3.7-Sonnet underperforms with 4.5% point. Also, RULEREASONER-8B outperforms prior strong RBRs such as HtT and Chain-of-Logic, which are built directly on top of OpenAI o3-mini for most of tasks, except for the AR-LSAT and Logical Deduction. This implies that **RULEREASONER benefits**

<sup>1</sup>Following RL literature nomenclature, we refer to models trained with the negative log-likelihood loss as behavioral cloning and perform task-focused supervised training to maximize baseline performance.

<sup>2</sup><https://github.com/QwenLM/Qwen3>

Table 3: Comparison with all baselines on eight ID benchmarks. RULEREASONER significantly outperforms most of other methods. Average is the macro mean across all samples of domains.

	Induction	Deduction		FOL		Others			Avg. Results
	Clutrr	ProntoQA	ProofWriter	FOLIO	LogicNLI	AR-LSAT	Logic. Dedu.	LogiQA	
<b>Prior RBRs</b>									
HtT (Zhu et al., 2023)	40.3	92.0	88.0	71.0	54.0	97.0	100.0	79.1	77.7
RGFB (Diallo et al., 2025)	31.3	94.0	88.0	74.0	55.0	95.0	100.0	79.1	77.1
Chain-of-Logic (Servantez et al., 2024b)	44.8	91.0	92.0	80.0	54.0	97.0	100.0	80.6	80.0
<b>Frontier Reasoners</b>									
OpenAI o1 (Jaech et al., 2024)	52.2	91.0	91.0	77.0	60.0	98.0	88.0	82.1	79.9
OpenAI o3-mini (Zhang et al., 2025b)	40.3	94.0	93.0	74.0	55.0	96.3	100.0	77.6	78.8
Claude-3.7-Sonnet (Anthropic, 2025)	65.7	92.8	90.0	74.7	58.0	76.2	97.0	81.5	79.5
DeepSeek-R1 (Guo et al., 2025)	71.6	40.0	27.0	72.7	49.0	89.7	98.3	85.0	66.7
<b>Behavioral Cloning</b>									
SFT w/o CoT (Wei et al., 2022b)	37.5	96.0	88.8	73.4	74.8	37.5	85.9	76.1	71.2
SFT w/ Short CoT (Yeo et al., 2025)	77.6	92.6	87.0	82.9	73.8	54.8	87.6	88.0	80.9
SFT w/ Long CoT (Yeo et al., 2025)	83.5	95.6	89.2	83.4	76.6	68.6	79.6	79.1	81.9
<b>Advanced RLVRs</b>									
GRPO (Shao et al., 2024)	73.1	95.4	96.4	72.3	66.6	36.3	90.3	70.1	75.0
Dr. GRPO (Liu et al., 2025)	68.6	96.0	95.6	73.9	75.4	32.1	84.3	65.6	73.9
DAPO (Yu et al., 2025)	86.5	96.0	94.8	80.9	65.8	40.0	95.3	74.6	79.2
<b>RULEREASONER (Ours)</b>									
RULEREASONER-4B	82.0 <sub>0.4</sub>	95.0 <sub>0.6</sub>	96.3 <sub>0.3</sub>	78.9 <sub>0.8</sub>	66.6 <sub>0.4</sub>	38.6 <sub>0.5</sub>	96.3 <sub>0.2</sub>	80.5 <sub>0.7</sub>	79.2 <sub>0.6</sub>
RULEREASONER-8B	<b>95.5<sub>0.3</sub></b>	<b>96.4<sub>0.4</sub></b>	<b>97.0<sub>0.2</sub></b>	<b>84.7<sub>0.6</sub></b>	<b>70.4<sub>0.1</sub></b>	<b>46.8<sub>0.3</sub></b>	<b>98.3<sub>0.4</sub></b>	<b>83.5<sub>0.3</sub></b>	<b>84.0<sub>0.5</sub></b>

from RLVRs to obtain higher improvement in rule understanding and utilization. In addition, RULEREASONER-8B also outperforms recent RLVR methods which are trained with lower intra-task performance variance on *all eight tasks*, for instance, higher average performance of 84.0% (+4.8%) yet with a lower variance of 3.1% (-0.5%), comparing to DAPO (79.2% with a variance of 3.6%). This demonstrates that **RULEREASONER not only develops impressive individual task performance, but also maintains the domain performance balance.**

**Out-of-Distribution Performance.** As illustrated in Figure 1, RULEREASONER-8B surpasses frontier LRMs across *three OOD benchmarks*. Specifically, it shows a remarkable 10.4% improvement compared to OpenAI-o1. As depicted in Table 4, RULEREASONER-8B consistently increases performance across the three OOD benchmarks, achieving the highest average performance gains of  $\Delta 56.0\%$ , including  $\Delta 71.4\%$  on BBH,  $\Delta 48.4\%$  on ProverQA, and  $\Delta 48.2\%$  on BBEH. These findings highlight the effectiveness of RULEREASONER in enhancing the general rule-based reasoning capabilities of SLMs. As shown in Table 4, the SFT baseline lags behind RULEREASONER in both ID and OOD evaluations. Notably, while SFT improves ID performance to nearly match RULEREASONER (81.9% versus 84.0%), its OOD performance remains significantly lower (34.4% versus 54.5%). We conclude that, in contrast to RLVR, SFT does not effectively generalize to unseen rules or OOD scenarios, which is also aligned to Chu et al. (2025). Moreover, a relative OOD demonstration in Table 5 reveals that **RULEREASONER elicits SLMs to extrapolate, allowing them to generalize to new tasks by applying unseen rules, through a concise and logically rigorous reasoning trajectory.**

### 4.3. Analyses and Discussion

**Advantages of DADS.** As depicted in Figure 4, we find that RULEREASONER enhances task performance across all domains without tradeoff. Specifically, RULEREASONER takes care of the underperformed task such as AR-LSAT and increases its domain weights consistently. Even for the low-portion domains (e.g., ProofWriter) and the continuously decreased domains such as Clutrr, RULEREASONER

Table 4: Comparison of average improvement. % denotes ID performance and % denotes OOD performance, respectively. Unlike the task-focused settings in §4.1, † indicates full-set mix supervised training to obtain stronger OOD performance for SFT.

Model	Pass@1	Avg. $\Delta$
Qwen3-8B	27.4 / 34.2	—
+ SFT <sup>†</sup>	81.9 / 66.6	54.5 / 34.4
+ GRPO	75.0 / 75.8	47.6 / 41.6
+ RULEREASONER	<b>84.0 / 81.7</b>	<b>56.6 / 47.5</b>



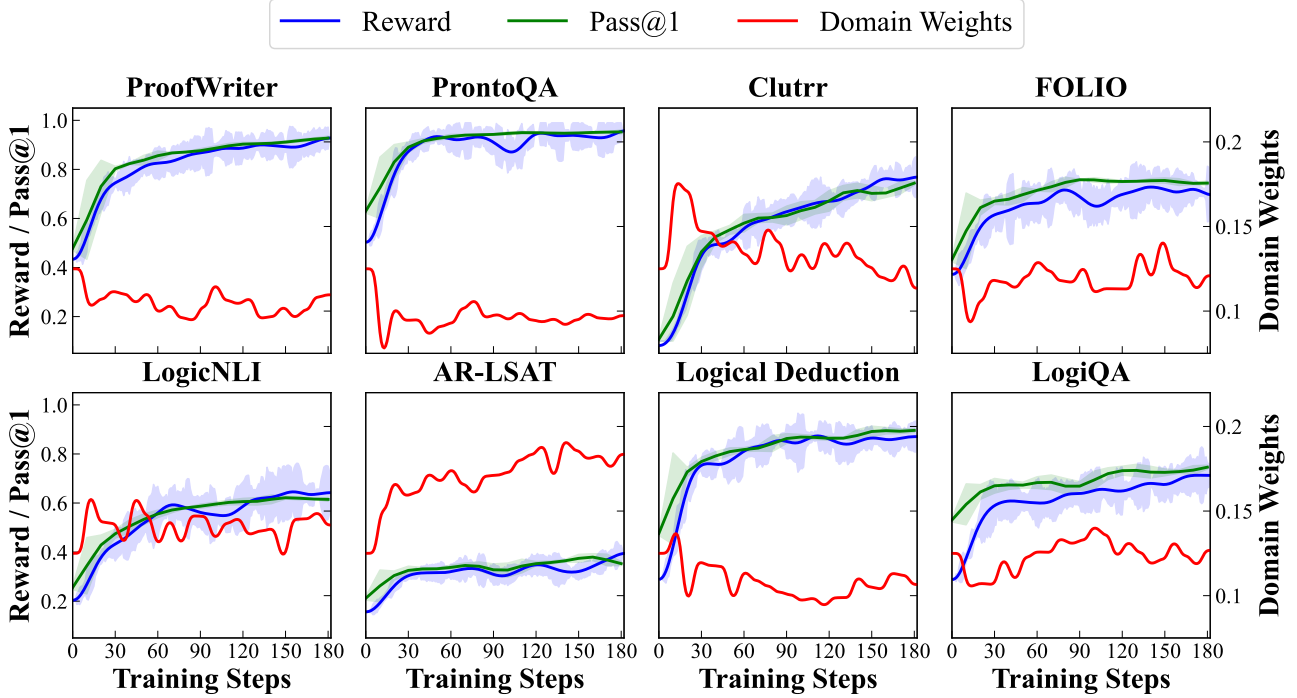


Figure 4: Learning dynamics by domains. “Reward” represents the training-time reward obtained from each task and “Pass@1” denotes the validation pass@1 performance. We employ exponential moving average smoothing for clearly displaying the curves “Reward”, “Pass@1”, and “Domain Weights”.

still steadily improves their training rewards and validation accuracy without reaching an obvious plateau. Interesting, analogous to the phenomenon described by Zucchet et al. (2025), the knowledge acquisition period during pre-training is accelerated fast on transition, but led to overfitting by the imbalanced data distributions. Hence, we conclude that DADS serves as an online data scheduling strategy, stabilizing the dynamics of on-policy RL training and mitigates over-optimization.

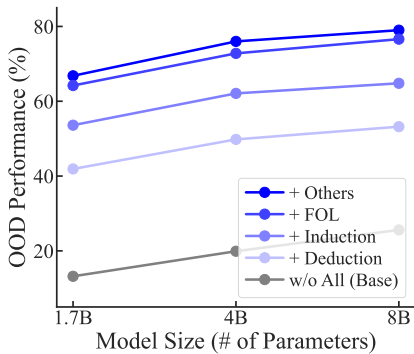


Figure 5: Impact on incremental task mixing recipes.

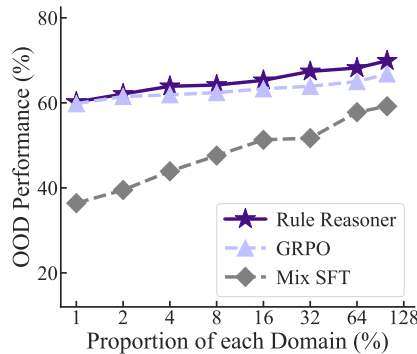


Figure 6: Impact on training sample efficiency.

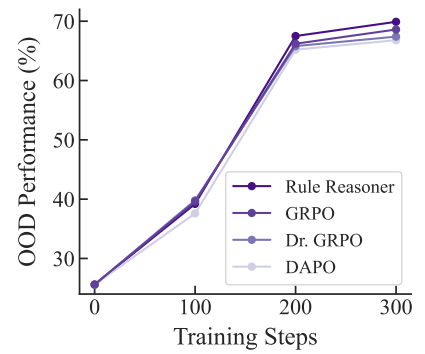


Figure 7: Comparisons on different RLVR methods.

**Impact of Task Mixing Recipe.** Figure 5 illustrates the impact of mixing recipes of incremental tasks on the average OOD performance across models of different sizes. The incremental mixing strategies show a consistent upward trend as the model size increases indicating that a larger model capacity generally contributes to better generalization ability for OOD performance. The mixed collection of various reasoning types enhances the OOD performance nearly to 80% while the baseline “w/o

All” at around 25%. It further validates the benefit of our training data curation principles and the effectiveness of task mixing for learning generalization.

**Investigation on Training Sample Efficiency.** Figure 6 assesses the OOD performance with scaling training sample sizes among the selected three methods at training step 500. We sample training data with different proportion of domain by log scale. It is evident that for all three methods, as the proportion of each domain increases, the OOD performance shows a continuous upward trend. RULEREASONER consistently outperforms GRPO and Mix SFT at every proportion level. For the same proportion of data, RULEREASONER obviously achieves higher performance than Mix SFT as well as a slight advantage over GPRO. For example, when the domain proportion is 1%, RULEREASONER achieves an OOD performance of about 60% point, while the other two is around 59% point and 42% point. Similarly, our method is more effective with high training sample efficiency to achieve the same OOD performance, which implies the effectiveness of the online data sampling policy in Section 3.

As shown in Figure 7, we also prove that RULEREASONER improves sample efficiency with fewer training steps than the existing RLVR methods. It achieves comparable performance with  $\sim 66.8\%$  on average, roughly fewer 100 steps compared to the GRPO. Similarly, DAPO requires more steps to match RULEREASONER’s performance, which indicates that the dynamic batch-level data curricula in RULEREASONER benefit RLVR and accelerate it to achieve higher performance with fewer costs.

**Case Studies: RULEREASONER Improves the Quality of the Rule-based Application for Reasoning.** We show examples from RULEREASONER-8B in various test sets in Table 5 to demonstrate how our training recipe improves the performance of the model. Before training, the model fails to solve the problems because of various reasons, such as poor reasoning logic, reasoning distracted by the noisy rules, etc. After training, the model is able to derive a valid reasoning chain and answer rigidly. Additionally, we provide more successful and failed cases with their trajectories in Table 7 and 8.

Table 5: Reasoning chains before and after training by RULEREASONER for comparison.

Case Type	Input	Before Training	After Training
Explicit Rules (ProofWriter)	<b>Rules:</b> If something is smart then it is furry. If something is cold it is green. Cold, smart things are green. All furry things are white. If something is green, it is smart. All furry things are green. All cold, green things are rough. Cold things are green. <b>Facts:</b> Dave is blue. Dave is cold. Dave is green. <b>Question:</b> Dave is furry?	Step 1: Analyze the given rules and facts. Rule 1: If something is smart, then it is furry. ... Fact 1: Dave is blue. Fact 2: Dave is cold. ... Step 2: Determine if Dave is furry based on the given rules and facts. From Fact 2 and Rule 2, we know that Dave is green. From Fact 2 and Rule 8, we know that Dave is green. <b>However, we also have Rule 4 stating that all furry things are white, and Dave is blue (Fact 1). This creates a contradiction, as Dave cannot be both furry and blue.</b> Since we have a contradiction, we cannot determine if Dave is furry. The answer is Unknown. <b>Answer: Unknown. ✗</b>	Step-by-Step Reasoning: 1. Dave is cold. From Rule 8: If something is cold, then it is green. Since Dave is cold, Dave is green. This is consistent with the given fact that Dave is green. <b>From Rule 5: If something is green, then it is smart. Since Dave is green, Dave is smart. From Rule 1: If something is smart, then it is furry. Since Dave is smart, Dave is furry.</b> From Rule 6: All furry things are green. This is consistent with the given fact that Dave is green. 3. Conclusion: Since Dave is smart (from Rule 5), and from Rule 1, if something is smart, then it is furry, we can conclude Dave is furry. <b>Answer: True. ✓</b>
Implicit Rules (BBH)	<b>Context:</b> The following paragraphs each describe a set of three objects arranged in a fixed order. The statements are logically consistent within each paragraph. On a branch, there are three birds: a robin, an owl, and a hawk. The robin is to the right of the hawk. The owl is the leftmost. <b>Options:</b> (A) The robin is the rightmost (B) The owl is the rightmost (C) The hawk is the rightmost. <b>Question:</b> Which of the following options is true?	To answer this question, let’s analyze the information provided in the context step by step: 1. There are three birds on a branch: ... Now, let’s evaluate the options: (A) The robin is the rightmost: <b>This option is not true because the robin is to the right of the hawk, but not necessarily the rightmost bird.</b> (B) The owl is the rightmost: This option is not true because the owl is the leftmost bird, not the rightmost. <b>Based on the information provided, none of the options (A, B, or C) are true.</b> However, since we must choose one of the given options, we can conclude that the correct answer is: (B) Please note that this answer is not entirely accurate, as the owl is actually the leftmost bird. The question might have a mistake or missing information. <b>Answer: B. ✗</b>	To determine the correct order of the three birds on the branch, let’s analyze the information provided: 1. The owl is the leftmost. This means the owl is in the first position on the branch. 2. The robin is to the right of the hawk. This means the robin comes after the hawk in the order. Now, let’s determine the positions of the hawk and the robin: Since the owl is the leftmost, the owl must be in the first position. The robin is to the right of the hawk, so the hawk must be in the second position, and the robin must be in the third position. <b>So, the order from left to right is: 1. Owl 2. Hawk 3. Robin</b> Now, let’s evaluate the options: (A) The robin is the rightmost. So, this statement is true. (B) The owl is the rightmost. ... So, this statement is false. <b>Answer: A. ✓</b>

## 5. Conclusion

We introduce RULEREASONER, a novel training framework tailored to enhance both effectiveness and efficiency of multi-domain training for RLVR. RULEREASONER effectively harmonizes diverse rule-based reasoning capabilities across various reasoning tasks, resulting in demonstrably higher in-distribution and out-of-distribution performance efficiently compared to prior rule-based reasoners, state-of-the-art LRMs, strong supervised training approaches, and RLVR methods. We aim to explore further research into data-centric approaches to facilitate reasoning model efficiency in future work.

**Limitations & Future Work.** In this study, due to the scarcity and imbalance nature of rule-based reasoning data, the current methods may not encompass all possible rule formats and complexities encountered in real-world applications with limited rule coverage, which hinders the performance from tasks generalization. Besides, the performance of our method is still constrained by the quality of rule filtering, particularly easy to be distracted when dealing with noisy or redundant rules that can negatively impact reasoning performance. Furthermore, while the method demonstrates effectiveness with smaller models (4B and 8B parameters), its scalability to larger-scale modeling remains unverified that could be more effective in complex scenarios. Finally, it remains to be explored that whether the approach shows better extrapolating to long logical reasoning chains requiring extensive rule compositions or deep logical deductions. These limitations highlight areas for future improvement in expanding rule diversity and robustness as well as model capacity scaling to meet the needs of supporting longer reasoning trajectories.

**Acknowledgments.** We would like to thank Zixia Jia, Hengli Li, and Xubo Qin from BIGAI for their valuable contributions in discussing the project, and Tong Wu from BIGAI for his initial trials, helpful discussions on RLVR training, and the infra-system operation and maintenance.

## References

- Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.720. URL <https://aclanthology.org/2024.acl-long.720/>.
- Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. Can LLMs reason with rules? logic scaffolding for stress-testing and improving LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7523–7543, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.406. URL <https://aclanthology.org/2024.acl-long.406/>.
- Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. Chain of logic: Rule-based reasoning with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2721–2733, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.159. URL <https://aclanthology.org/2024.findings-acl.159/>.
- Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. *Advances in Neural Information Processing Systems*, 37:73572–73604, 2024.
- Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. Symbolic working memory enhances language models for complex rule application. *arXiv preprint arXiv:2408.13654*, 2024b.
- Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36:52430–52452, 2023a.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, 2024.
- Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. Testing the general deductive reasoning capacity of large language models using ood examples. *Advances in Neural Information Processing Systems*, 36:3083–3105, 2023.

- Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. Large language models are in-context semantic reasoners rather than symbolic reasoners. *arXiv preprint arXiv: 2305.14825*, 2023a.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022a. URL [https://openreview.net/forum?id=\\_VjQlMeSB\\_J](https://openreview.net/forum?id=_VjQlMeSB_J).
- Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*, 2025.
- Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model, 2025. URL <https://arxiv.org/abs/2503.24290>, 2025.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf).
- Gaurav Srivastava, Shuxiang Cao, and Xuan Wang. Towards reasoning ability of small language models. *arXiv preprint arXiv:2502.11569*, 2025.
- Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. *arXiv preprint arXiv:2502.12143*, 2025a.
- Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025.
- David Silver and Richard S Sutton. Welcome to the era of experience. *Preprint of a chapter to appear in Designing an Intelligence, edited by George Konidaris, MIT Press (forthcoming)*, 2025.
- Xuefeng Li, Haoyang Zou, and Pengfei Liu. Limr: Less is more for rl scaling. *arXiv preprint arXiv:2502.11886*, 2025b.
- Yuxin Zuo, Kaiyan Zhang, Shang Qu, Li Sheng, Xuekai Zhu, Biqing Qi, Youbang Sun, Ganqu Cui, Ning Ding, and Bowen Zhou. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025.
- Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. Large language models can learn rules. *arXiv preprint arXiv:2310.07064*, 2023.
- Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. Chain of logic: Rule-based reasoning with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2721–2733, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.159. URL <https://aclanthology.org/2024.findings-acl.159/>.
- Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. Large language models are in-context semantic reasoners rather than symbolic reasoners. *arXiv preprint arXiv:2305.14825*, 2023b.
- L Luo, J Ju, B Xiong, YF Li, G Haffari, and S Pan. Chatrule: Mining logical rules with large language models for knowledge graph reasoning. *arxiv 2024. arXiv preprint arXiv:2309.01538*, 2024.
- Zixia Jia, Junpeng Li, Shichuan Zhang, Anji Liu, and Zilong Zheng. Combining supervised learning and reinforcement learning for multi-label classification tasks with partial labels. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13553–13569, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.731. URL <https://aclanthology.org/2024.acl-long.731/>.
- Jiayi Gui, Yiming Liu, Jiale Cheng, Xiaotao Gu, Xiao Liu, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. Logicgame: Benchmarking rule-based reasoning abilities of large language models. *arXiv preprint arXiv:2408.15778*, 2024.



- Qiyuan He, Jianfei Yu, and Wenya Wang. Large language model-enhanced symbolic reasoning for knowledge base completion. *arXiv preprint arXiv:2501.01246*, 2025.
- Ruiwen Zhou, Wenyue Hua, Liangming Pan, Sitao Cheng, Xiaobao Wu, En Yu, and William Yang Wang. Rulearena: A benchmark for rule-guided reasoning with llms in real-world scenarios. *arXiv preprint arXiv:2412.08972*, 2024.
- Kaiyu He, Mian Zhang, Shuo Yan, Peilin Wu, and Zhiyu Zoey Chen. Idea: Enhancing the rule learning ability of large language model agent through induction, deduction, and abduction. *arXiv preprint arXiv:2408.10455*, 2024.
- Xiaojuan Tang, Jiaqi Li, Yitao Liang, Muhan Zhang, and Zilong Zheng. Mars: Situated inductive reasoning in an open-world environment. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks*, 2024.
- Aïssatou Diallo, Antonis Bikakis, Luke Dickens, Anthony Hunter, and Rob Miller. Rule-guided feedback: Enhancing reasoning by enforcing rule adherence in large language models. *arXiv preprint arXiv:2503.11336*, 2025.
- Xiangyu Peng, Congying Xia, Xinyi Yang, Caiming Xiong, Chien-Sheng Wu, and Chen Xing. Regenesi: Llms can grow into reasoning generalists via self-improvement. *arXiv preprint arXiv:2410.02108*, 2024.
- Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. Can LLMs reason with rules? logic scaffolding for stress-testing and improving LLMs. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7523–7543, Bangkok, Thailand, August 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.406. URL <https://aclanthology.org/2024.acl-long.406/>.
- Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. Symbolic working memory enhances language models for complex rule application. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17583–17604, Miami, Florida, USA, November 2024d. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.974. URL <https://aclanthology.org/2024.emnlp-main.974/>.
- Yifan Zhang, Yifeng Liu, Huizhuo Yuan, Yang Yuan, Quanquan Gu, and Andrew C Yao. On the Design of KL-Regularized Policy Gradient Algorithms for LLM Reasoning. *arXiv e-prints*, art. arXiv:2505.17508, May 2025.
- Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *arXiv preprint arXiv:2505.16400*, 2025.
- Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, and Sida I Wang. Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution. *arXiv preprint arXiv:2502.18449*, 2025.
- Jiaqi Li, Xinyi Dong, Yang Liu, Zhizhuo Yang, Quansen Wang, Xiaobo Wang, SongChun Zhu, Zixia Jia, and Zilong Zheng. Reflectevo: Improving meta introspection of small llms by learning self-reflection. *arXiv preprint arXiv:2505.16475*, 2025c.
- Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025.
- Hengli Li, Chenxi Li, Tong Wu, Xuekai Zhu, Yuxuan Wang, Zhaoxin Yu, Eric Hanchen Jiang, Song-Chun Zhu, Zixia Jia, Ying Nian Wu, et al. Seek in the dark: Reasoning via test-time instance-level policy gradient in latent space. *arXiv preprint arXiv:2505.13308*, 2025d.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.
- Xiaojiang Zhang, Jinghui Wang, Zifei Cheng, Wenhao Zhuang, Zheng Lin, Minglei Zhang, Shaojie Wang, Yinghan Cui, Chao Wang, Junyi Peng, et al. Srpo: A cross-domain implementation of large-scale reinforcement learning on llm. *arXiv preprint arXiv:2504.14286*, 2025a.
- Taiwei Shi, Yiyang Wu, Linxin Song, Tianyi Zhou, and Jieyu Zhao. Efficient reinforcement finetuning via adaptive curriculum learning. *arXiv preprint arXiv:2504.05520*, 2025.

- Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=XfHWcNTShp>. Survey Certification.
- Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=1XuByUeHhd>.
- Charles C Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting*, 20(1):5–10, 2004.
- Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL <https://aclanthology.org/2021.findings-acl.317/>.
- Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=qFVBzXxR2V>.
- Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1458. URL <https://aclanthology.org/D19-1458/>.
- 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, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1229. URL <https://aclanthology.org/2024.emnlp-main.1229/>.
- Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. Diagnosing the first-order logical reasoning ability through LogicNLI. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.303. URL <https://aclanthology.org/2021.emnlp-main.303/>.
- Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Yining Chen, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. Analytical reasoning of text. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2306–2319, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.177. URL <https://aclanthology.org/2022.findings-naacl.177/>.
- Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13326–13365, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.720. URL <https://aclanthology.org/2024.acl-long.720/>.
- Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962, 2023b. doi: 10.1109/TASLP.2023.3293046.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL <https://aclanthology.org/2023.findings-acl.824/>.
- Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, et al. Big-bench extra hard. *arXiv preprint arXiv:2502.19187*, 2025.

- Chengwen Qi, Ren Ma, Bowen Li, He Du, Binyuan Hui, Jinwang Wu, Yuanjun Laili, and Conghui He. Large language models meet symbolic provers for logical reasoning evaluation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=C25SgeXWjE>.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- Brian Zhang, Eric Mitchell, Hongyu Ren, Kevin Lu, Max Schwarzer, Michelle Pokrass, Shengjia Zhao, Ted Sanders, Adam Kalai, Alexandre Passos, Benjamin Sokolowsky, Elaine Ya Le, Erik Ritter, Hao Sheng, Hanson Wang, Ilya Kostrikov, James Lee, Johannes Ferstad, Michael Lampe, Prashanth Radhakrishnan, Sean Fitzgerald, Sébastien Bubeck, Yann Dubois, Yu Bai, Andy Applebaum, Elizabeth Proehl, Evan Mays, Joel Parish, Kevin Liu, Leon Maksin, Leyton Ho, Miles Wang, Michele Wang, Olivia Watkins, Patrick Chao, Samuel Miserendino, Tejal A. Patwardhan, Antonia Woodford, Beth Hoover, Jake Brill, Kelly Stirman, Neel Ajjarapu, Nick Turley, Nikunj Handa, Olivier Godement, Akshay Nathan, Alyssa Huang, Andy Wang, Ankit Gohel, Ben Eggers, Brian Yu, Bryan Ashley, Chengdu Huang, Davin Bogan, Emily Sokolova, Eric Horacek, Felipe Petroski Such, Jonah Cohen, Joshua Gross, Justin Becker, Kan Wu, Larry Lv, Lee Byron, Manoli Liodakis, Max Johnson, Mike Trpcic, Murat Yesildal, Rasmus Rygaard, R. J. Marsan, Rohit Ram-chandani, Rohan Kshirsagar, Sara Conlon, Tony Xia, Siyuan Fu, Srinivas Narayanan, Sulman Choudhry, Tomer Kaftan, Trevor Creech, Andrea Vallone, Andrew Duberstein, Enis Sert, Eric Wallace, Grace Zhao, Irina Kofman, Jieqi Yu, Joaquin Quiñero Candela, Madeleine Boyd, Mehmet Ali Yatbaz, Mike McClay, Mingxuan Wang, Sandhini Agarwal, Saachi Jain, Sam Toizer, Santiago Hernández, Steve Mostovoy, Tao Li, Young Cha, Yunyun Wang, Lama Ahmad, Troy Peterson, Carpus Chang, Kristen Ying, Aidan Clark, Dane Stuckey, Jerry Tworek, Jakub W. Pachocki, Johannes Heidecke, Kevin Weil, Liam Fedus, Mark Chen, Sam Altman, and Wojciech Zaremba. Openai o3-mini system card. In *OpenAI Blog Post*, 2025b. URL <https://cdn.openai.com/o3-mini-system-card-feb10.pdf>.
- Anthropic. Claude 3.7 sonnet. <https://www.anthropic.com/news/claude-3-7-sonnet>, 2025. Accessed: 2025-04-11.
- Dean A. Pomerleau. Alvin: An autonomous land vehicle in a neural network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988. URL [https://proceedings.neurips.cc/paper\\_files/paper/1988/file/812b4ba287f5ee0bc9d43bbf5bbe87fb-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/1988/file/812b4ba287f5ee0bc9d43bbf5bbe87fb-Paper.pdf).
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022b. URL <https://openreview.net/forum?id=gEzrGCozdqR>.
- Edward Yeo, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning in LLMs. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025. URL <https://openreview.net/forum?id=AgtQlhMQOV>.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding rl-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024b.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.
- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *The Second Conference on Parsimony and Learning (Recent Spotlight Track)*, 2025. URL <https://openreview.net/forum?id=d3E3LWmTar>.
- Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations. *arXiv preprint arXiv:2503.21676*, 2025.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

- Mehdi Fatemi, Banafsheh Rafiee, Mingjie Tang, and Kartik Talamadupula. Concise reasoning via reinforcement learning. *arXiv preprint arXiv:2504.05185*, 2025.
- Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*, 2025b.
- Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>.
- Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12): 3848–3860, August 2023. ISSN 2150-8097. doi: 10.14778/3611540.3611569. URL <https://doi.org/10.14778/3611540.3611569>.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf).
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.



## A. Ethical Considerations

We adhere to ethical principles to ensure the responsible development and application of our proposed techniques. The research conducted in the paper conforms in every respect, with the NeurIPS code of ethics guidelines. Our work focuses on enhancing the rule-based reasoning abilities of SLMs without directly involving human subjects or sensitive information. The study also acknowledges several ethical implications such as the transparency of rule-based systems, though advantageous for interpretability, raises concerns about accountability if rules are misapplied in high-stakes domains. We advocate for rigorous validation of rules against diverse datasets in avoid of human biases on manually crafted rules. We also acknowledge the potential broader impacts of our research, recognize the environmental and computational costs associated with LLM training, and strive to optimize our methods for efficiency.

## B. Prompts

In this work, we use the same prompt template for each dataset for model training and evaluation.

### B.1. Prompts for Dataset with Explicit Rules.

**Instruction:** Please answer the question based on the given rules and facts using either of [A/B/C/D] (or [True/False/Unknown]). Fill in the answer between <answer> and </answer>. Provide your step by step reasoning process between <think> and </think>.

**Input:**  
 Rules: {Rules}  
 Facts: {Facts}  
**Question:** {Question}  
**Options:** {Options} (OPTIONAL)

### B.2. Prompts for Dataset with Implicit Rules.

**Instruction:** Please answer the question based on the given contexts using either of [A/B/C/D] (or [True/False/Unknown]). Fill in the answer between <answer> and </answer>. Provide your step by step reasoning process between <think> and </think>.

**Input:**  
 Context: {Context}  
**Question:** {Question}  
**Options:** {Options} (OPTIONAL)

## C. Data Details

### C.1. Data Sources

We list the training and evaluation data sources associated with the urls used in the paper as below. The followings are the training and validation data sources:

- ProofWriter (2021): <https://allenai.org/data/proofwriter>
- ProntoQA (2023): <https://github.com/asaparov/prontoqa>
- Clutrr (2019): <https://github.com/SiyuanWangw/RuleApplication/blob/master/Data/clutrr>
- AR-LSAT (2022): <https://github.com/SiyuanWangw/RuleApplication/blob/master/Data>
- FOLIO (2024): <https://github.com/Yale-LILY/FOLIO/blob/main/data/v0.0>
- LogicNLI (2021): <https://github.com/omnilabNLP/LogicNLI/blob/main/dataset>

- LogicalDeduction (2024): <https://github.com/Aiden0526/SymbCoT/tree/main/data>
- LogiQA (2023): <https://github.com/csitfun/LogiQA2.0/blob/main/logiqa/DATA/LOGIQA>

The followings are the OOD test data sources:

- BigBench-Hard (2023): <https://huggingface.co/datasets/lukaemon/bbh>
- ProverQA (2025): <https://huggingface.co/datasets/opendatalab/ProverQA>
- BigBench-Extra-Hard (2025): <https://github.com/google-deepmind/bbeh>

## C.2. Dataset Curation Details

For ProntoQA, we randomly negate some of the proof questions to avoid learning the shortcut of answer “True”. For ProofWriter, we randomly sample ten percent of the original source data considering the imbalance nature of the whole training data. Then we use DeepSeek-R1 to generate the reasoning process including short CoT and long CoT sequences for each QA sample. For LogiQA, we use data with the reasoning type both categorical reasoning and conjunctive reasoning that leverages the implicit rule application and reasoning. For BigBench-Hard, we use the subset of “logical\_deduction” with three, five, and seven objects with varying levels of difficulties and select BoardgameQA and ZebraPuzzles from BigBench-Extra-Hard to keep consistent with our task definition for OOD evaluation.

## D. Additional Evaluation Results

**Challenges of Different Rule Settings.** We investigate the task performance on Clutrr, with three-level rule settings in Figure 8 in the followings: 1) Ordered Rules: rules are arranged in their application order. 2) Shuffled Rules: rules are provided in a random order. 3) Noisy Rules: rules are shuffled and include irrelevant ones. aligns with real-world scenarios, rules may contain distractors. To our expectation, the task with ordered rules achieves the best performance among them likely due to the logical sequence aiding in task execution. Shuffled Rules, while still contain only the relevant rules but in a random order, show a moderate performance drop. Noisy rules result in the most significant performance reduction with the added complexity of redundant rules as distractors, highlighting the negative effect on task performance.

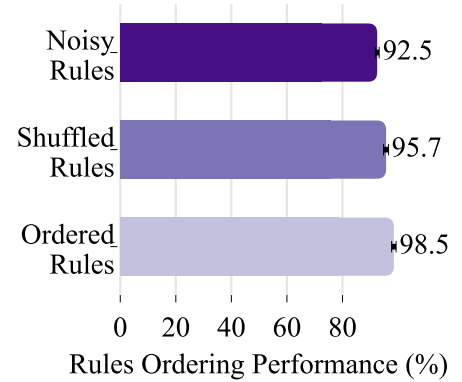


Figure 8: Comparison of performance on challenging rule settings.

**RULEREASONER can Adapt to Varying Rule Complexity.** As depicted in Figure 9, we present the extended OOD evaluation results, with test sets separated by rule complexity (*i.e.*, task difficulty). The BBH, ProverQA, and BBEH benchmarks consist of questions requiring reasoning up to various difficulties that hinge on the diverse factors of query complexity.

Specifically, for BBH, we divide the original test set into three difficulty levels based on the multi-hop number of the query. For ProverQA, we adopt the original difficulty levels from its source, which is separated by the number of reasoning steps. For BBEH, the test set was categorized into three levels according to the query length (in tokens): Easy [0, 1068), Medium [1068, 2175), and Hard [2175, 2741). Thus, we test the generalization capabilities on the three subsets of each OOD benchmark and report their performance on the higher difficulty questions. Not surprisingly, performance in easy subset exhibit

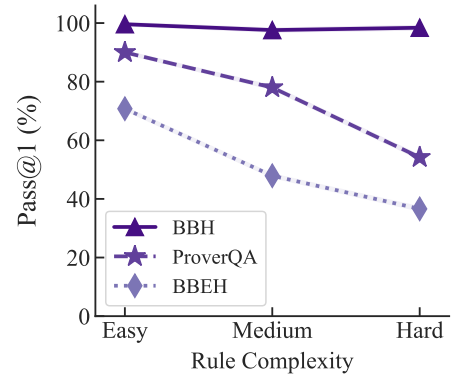


Figure 9: Comparison of performance on varying task complexity.

substantially stronger than corresponding medium- and hard-level subsets, with an average pass@1 of 86.7% compared to the 63.0% (-23.7%) of hard subset across three benchmarks. Interestingly, we notice that RULEREASONER-4B drops significantly along with subsets in different difficulties, while maintain the still performance in BBH. One possible explanation is that our base model, Qwen3-8B-Base, might have encountered test set leakage, given that BBH was published in late 2022 but Qwen3 models were released in 2025 (Yang et al., 2025a). Therefore, we suggest assessing our models to more challenging benchmarks to achieve more robust and reasonable results. We leave this direction for future work.

**Test-time Scalability.** In the cutting-edge discussion on the essence and usefulness of a longer thinking process, (Fatemi et al., 2025) find that the extra generated tokens do not help improve the final prediction accuracy, while (Yeo et al., 2025; Yang et al., 2025b) hold the opposite positions which claim that accurate results are not necessary with the long reasoning process. As depicted in Figure 10, we perform repeated sampling to investigate the upper limit of performance for each question in the way of Brown et al. (2024), illustrated in brown curves. Similarly, we take the majority vote and normalized weighted sum methods in the way of Wang et al. (2023), respectively. To investigate

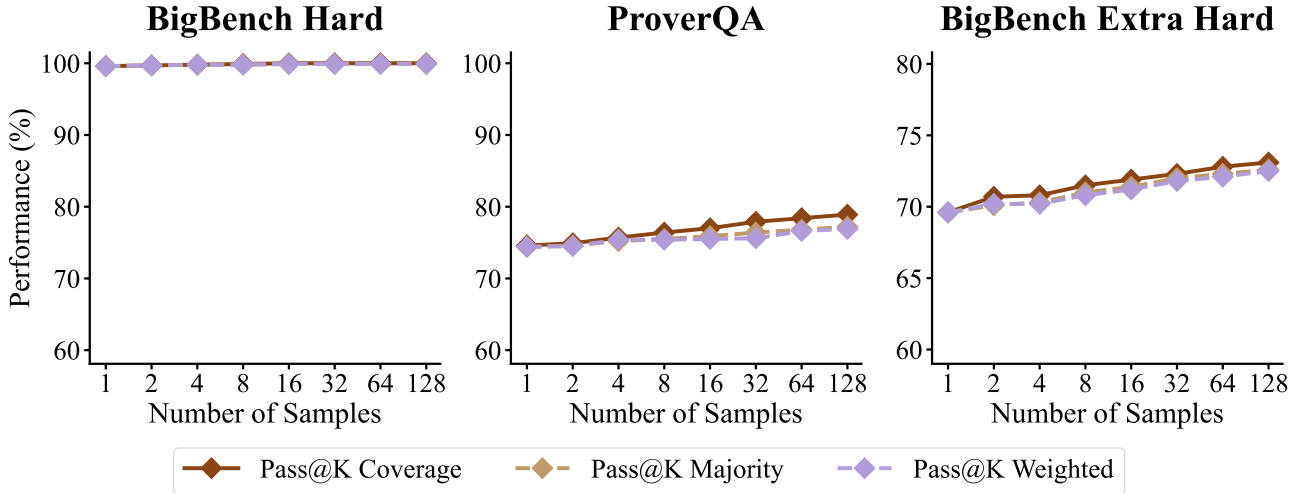


Figure 10: Comparison of OOD performance in parallel test-time scaling methods.

the interesting question, we study two experimental context scaling strategies on RULEREASONER, including 1) training-time iterative scaling and 2) test-time scaling. For test-time scaling, we report sequential and parallel strategies, respectively. As show in Figure 10, test-time scaling demonstrates varied effectiveness across benchmarks. For BigBench Hard, all Pass@K methods achieve near-perfect performance ( $\sim 100\%$ ) with minimal scaling, indicating the dataset’s limited complexity for distinguishing scaling benefits. On ProverQA, Pass@K Coverage consistently outperforms both Majority and Weighted approaches, with performance gaps widening as sample size increases ( $\sim 1.7\%$  at 128 samples). BigBench Extra Hard reveals the most substantial scaling benefits, where Coverage method achieves 73.1% pass@k performance compared to  $\sim 72\%$  for alternative approaches at 128 samples. The consistent superiority of Coverage sampling across challenging benchmarks (ProverQA and BigBench Extra Hard) suggests that diverse solution exploration outweighs consensus-based aggregation for complex reasoning tasks. These findings support the position that extended reasoning processes, when properly sampled, do enhance prediction accuracy for sufficiently difficult problems in rule-based reasoning.

## E. Computational Infrastructure & Hyperparameters

We list the details of the computational infrastructure and hyperparameters of training and inference used in this work in Table 6.

Table 6: Hyper-parameters of RULEREASONER-4B and RULEREASONER-8B on-policy training and inference.

Computational Infrastructure 4 × A100-SXM4-80GB GPU			
Hyperparameter	Assignment	Hyperparameter	Assignment
Base model	RULEREASONER-4B (Qwen3-4B-Base)	Base model	RULEREASONER-8B (Qwen3-8B-Base)
Training strategy	FSDP (Zhao et al., 2023)	Training strategy	FSDP (Zhao et al., 2023)
Training epochs	3	Training epochs	3
Training precision	bfloat16	Training precision	bfloat16
Mini batch size	65,536 tokens	Mini batch size	65,536 tokens
Micro batch size	16,384 tokens	Micro batch size	16,384 tokens
Max input length	1,024	Max input length	1,024
Max output length	2,048	Max output length	2,048
Max learning rate	1e − 6	Max learning rate	1e − 6
Optimizer	AdamW	Optimizer	AdamW
Adam beta weights	0.9, 0.95	Adam beta weights	0.9, 0.95
KL loss coefficient	0	KL loss coefficient	0
Rollout per sample	64	Rollout per sample	64
Clip ratio	0.2	Clip ratio	0.2
Gradient clipping	1.0	Gradient clipping	1.0
Evaluation steps	10	Evaluation steps	10
Save steps	10	Save steps	10
Training engine	PyTorch (Paszke et al., 2019)	Training engine	PyTorch (Paszke et al., 2019)
Decay factor $\alpha$	0.5	Decay factor $\alpha$	0.5
Base reward $\epsilon$	0.1	Base reward $\epsilon$	0.1
Scaling factor $\tau$	0.8	Scaling factor $\tau$	0.5
Rollout engine	vLLM (Kwon et al., 2023)	Rollout engine	vLLM (Kwon et al., 2023)
Rollout temperature	0.8	Rollout temperature	0.8
Rollout top-p	1	Rollout top-p	1
Random seeds	{21, 42, 84}	Random seeds	{21, 42, 84}

## F. Example Trajectories

We provide additional positive and negative examples (*i.e.*, failed cases) with fine-grained reasoning trajectories from the test sets of various tasks to demonstrate clear step-by-step logical reasoning patterns and well-organized thought process of RULEREASONER after training in Table 7 and Table 8.



Table 7: Positive examples of RULEREASONER on various datasets.

Dataset	Input	RULEREASONER
Clutrr	<p><b>Context:</b> James took his son Francisco out on a hunting trip. Natasha, Francisco's daughter, wanted to come, but they thought she was too young. Instead, James's wife Frances stayed home with her.</p> <p><b>Question:</b> How is Natasha related to Frances?</p>	<p>To determine how Natasha is related to Frances, let's analyze the information provided and trace the family relationships step by step. Given Information: 1. James took his son Francisco out on a hunting trip. This means James is Francisco's father ... Step-by-Step Reasoning: 1. James is Francisco's father, Francisco is James's son. 2. Since Frances is James's wife, she is also Francisco's mother. 3. Since Francisco is James's son and Frances's son, Natasha is Francisco's daughter, making Natasha James's granddaughter and Frances's granddaughter. 4. Conclusion: Natasha is the granddaughter of Frances.</p> <p>Answer: granddaughter ✓</p>
Folio	<p><b>Context:</b> Diamond Mine is a professional wrestling stable, formed in WWE. Roderick Strong leads Diamond Mine. Diamond Mine includes the Creed Brothers, and Ivy Nile. Imperium has a feud with Diamond Mine.</p> <p><b>Question:</b> Roderick strong leads a professional wrestling stable.</p>	<p>To determine whether the statement "Roderick Strong leads a professional wrestling stable" is true, false, or unknown based on the given context, let's analyze the information provided: 1. The context states: ... 3. From the context, we know that Roderick Strong leads Diamond Mine, and Diamond Mine is a professional wrestling stable. 4. Therefore, since Roderick Strong leads Diamond Mine, and Diamond Mine is a professional wrestling stable, it follows that Roderick Strong leads a professional wrestling stable.</p> <p>Answer: true ✓</p>
Logical Deduction	<p><b>Context:</b> A fruit stand sells five fruits: apples, loquats, cantaloupes, watermelons, and kiwis. The watermelons are more expensive than the loquats. The watermelons are less expensive than the kiwis. The loquats are more expensive than the apples. The cantaloupes are the cheapest.</p> <p><b>Options:</b> A) The apples are the third-most expensive. B) The loquats are the third-most expensive. C) The cantaloupes are the third-most expensive. D) The watermelons are the third-most expensive. E) The kiwis are the third-most expensive.</p> <p><b>Question:</b> Which of the following is true?</p>	<p>To determine the correct order of the fruits from most expensive to least expensive, let's analyze the information provided: 1. The cantaloupes are the cheapest. This means cantaloupes are the least expensive. 2. The loquats are more expensive than the apples. Loquats &gt; Apples 3. The watermelons are more expensive than the loquats. Watermelons &gt; Loquats 4. The watermelons are less expensive than the kiwis. Kiwis &gt; Watermelons Now, let's piece this information together to establish the order from most expensive to least expensive: the order from most expensive to least expensive is: Kiwis &gt; Watermelons &gt; Loquats &gt; Apples &gt; Cantaloupes Now, let's evaluate the options: A) The apples are the third-most expensive. Apples are fourth in the order, so this is false. B) The loquats are the third-most expensive. Loquats are third in the order, so this is true. C) The cantaloupes are the third-most expensive. Cantaloupes are fifth in the order, so this is false. D) The watermelons are the third-most expensive. Watermelons are second in the order, so this is false. E) The kiwis are the third-most expensive. Kiwis are first in the order, so this is false</p> <p>Answer: B ✓</p>
ProverQA	<p><b>Context:</b> Every protozoa is either parasitic or free-living, but not both. Frederick lives in a host. Any protozoa that lives in a host or feeds on nutrients is parasitic. Queenie lives in a host. Queenie does not feed on nutrients.</p> <p><b>Options:</b> A) True B) False C) Uncertain</p> <p><b>Question:</b> Queenie is free-living.</p>	<p>To determine whether the statement "Queenie is free-living", let's analyze the information provided and the options step by step. Given Information: 1. Every protozoa is either parasitic or free-living, but not both. This means if a protozoa is parasitic, it cannot be free-living, and vice versa ... Step-by-Step Reasoning: 1. Queenie lives in a host. From the given information, any protozoa that lives in a host or feeds on nutrients is parasitic. Since Queenie lives in a host, she falls into the category of protozoa that lives in a host. 2. Any protozoa that lives in a host or feeds on nutrients is parasitic. Since Queenie lives in a host, she is parasitic. 3. Every protozoa is either parasitic or free-living, but not both. If Queenie is parasitic, she cannot be free-living. 4. Conclusion: Since Queenie is parasitic, she is not free-living.</p> <p>Answer: B ✓</p>
LogiQA	<p><b>Context:</b> Customs is a state organ that supervises, checks and collects customs duties on all goods and articles entering and leaving the country.</p> <p><b>Options:</b> A) Check other people's luggage B) Check the body of others C) Take portraits of others D) Detain others for one day</p> <p><b>Question:</b> According to the above definition, when on duty, customs officer can?</p>	<p>To determine the correct answer, we need to analyze the role of a customs officer based on the given context and see which option aligns with their duties. Reasoning: 1. The primary role of a customs officer is to supervise, check, and collect customs duties on goods and articles entering and leaving the country. 2. To ensure compliance with customs regulations, customs officers may need to inspect luggage or packages to verify the contents and ensure they are properly declared. 3. Checking the body of others (Option B) is not a standard duty of a customs officer, as it would involve invasive procedures that are not typically part of customs inspection. 4. Taking portraits of others (Option C) is unrelated to the duties of a customs officer. 5. Detaining others for one day (Option D) is a more severe action and would require specific legal grounds, which are not mentioned in the context of customs duties. Conclusion: The most appropriate action for a customs officer, based on their role, is to check other people's luggage (Option A) to ensure compliance with customs regulations.</p> <p>Answer: A ✓</p>
ProntoQA	<p><b>Rules:</b> Every vumpus is a rompus and a sterpus. Every rompus is a brimpus and a lempus. Zumpuses are wumpuses and dumpuses. Each grimpus is a shumpus and a tumpus. Each lorpus is muffled and an impus. Every numpus is a gorpus. Gorpuses are zumpuses and jompuses. Every jompus is a vumpus and a grimpus.</p> <p><b>Facts:</b> Alex is a brimpus and a lempus. Alex is a zumpus and a gorpus</p> <p><b>Question:</b> Alex is not a vumpus?</p>	<p>We need to analyze the information provided and trace Alex's properties through the relationships defined. Given Rules: 1. Every vumpus is a rompus and a sterpus. 2. Every rompus is a brimpus and a lempus ... Given Facts: Alex is a brimpus and a lempus. Alex is a zumpus and a gorpus. Step-by-Step Reasoning: 1. Alex is a zumpus and a gorpus. From rule 7: Every gorpus is a zumpus and a jompus. Since Alex is a gorpus, Alex is a zumpus and a jompus. From rule 3: Zumpuses are wumpuses and dumpuses. Since Alex is a zumpus, Alex is a wumpus and a dumpus. From rule 8: Every jompus is a vumpus and a grimpus. Since Alex is a jompus, Alex is a vumpus and a grimpus. 2. Alex is a vumpus. From the above reasoning, since Alex is a jompus, and every jompus is a vumpus, Alex must be a vumpus. 3. Conclusion: The statement "Alex is not a vumpus" would be false because we have determined that Alex is indeed a vumpus.</p> <p>Answer: false ✓</p>

Table 8: Negative examples of RULEREASONER on various datasets.

Dataset	Input	RULEREASONER
LogiQA	<p><b>Context:</b> As far as the role of individual and nation is concerned, cultural quality mainly infiltrates people's mind and improves people's realm by forming a specific cultural atmosphere; In other words, the role of cultural quality is indirect and abstract</p> <p><b>Options:</b> A)Cultural quality is more important than humanistic spirit B)Humanistic spirit is more important than cultural quality C)Cultural quality and humanistic spirit are one thing D)When vigorously improving the quality of national culture, we should attach great importance to the construction of humanistic spirit</p> <p><b>Question:</b> The function of humanistic spirit is more direct and specific</p>	<p>Context Analysis: The role of cultural quality is described as indirect and abstract. It mainly infiltrates people's minds and improves their realm by forming a specific cultural atmosphere. The function of humanistic spirit is described as more direct and specific. Question Analysis: The question asks us to compare the importance of cultural quality and humanistic spirit based on their functions. It suggests that since the humanistic spirit is more direct and specific, we should consider its importance in relation to cultural quality. Reasoning: The context tells us that cultural quality is indirect and abstract, while the humanistic spirit is direct and specific. Since the humanistic spirit is more direct and specific, it likely has a more immediate and tangible impact compared to cultural quality, which is more indirect and abstract. Therefore, it would be reasonable to conclude that the humanistic spirit is more important than cultural quality in terms of its direct and specific function. Conclusion: Given that the humanistic spirit is more direct and specific, it is more important than cultural quality.</p> <p><b>Answer:</b> B ✗</p>
AR-LSAT	<p><b>Context:</b> While on vacation, Sukanya receives several e-mail messages from work, each message from one of three associates: Hilary, Jerome, and Lula. Sukanya receives at least one and no more than two messages from each of them. Sukanya receives each message on the day it is sent. No more than one message is sent each day. The messages are received in a manner consistent with the following: The first message is not from Lula. Both the first and last messages are from the same person. Exactly once Sukanya receives a message from Jerome on the day after receiving one from Hilary. Of the first three messages, exactly one is from Jerome.</p> <p><b>Options:</b> A) Exactly one of the messages is from Lula. B) Exactly two of the messages are from Jerome. C) The second message is from Lula. D) The third message is from Hilary. E) The fourth message is from Jerome.</p> <p><b>Question:</b> If Sukanya receives exactly four e-mail messages, then which one of the following must be true?</p>	<p>Step-by-Step Reasoning: 1. The first message is not from Lula. Both the first and last messages are from the same person Therefore, the last message must also be from the same person as the first message. So, the first and fourth messages are from either Hilary or Jerome. 2. Messages from Jerome: Since there is only one Jerome message in the first three messages, the Jerome message must be in one of the first three messages. The Jerome message on the day after Hilary must be the fourth message (since there are only four messages). 3. Messages from Lula: Sukanya receives at least one and no more than two messages from each associate. Since the first message is not from Lula, Lula must be in the second, third, or fourth message. If the first and fourth messages are from Hilary, then the second and third messages must include Lula and Jerome. If the first and fourth messages are from Jerome, then the second and third messages must include Hilary and Lula. 4. Analyzing the Options: Option A: Since Sukanya receives at least one message from each associate and there are four messages, it's possible that Lula sends two messages. This option is not necessarily true. Option B: The first and fourth messages are from the same person (either Hilary or Jerome). If the first and fourth messages are from Jerome, then Jerome sends two messages. If the first and fourth messages are from Hilary, then Jerome must send one message in the first three messages and one in the fourth message. Therefore, Jerome could send either one or two messages. This option is not necessarily true. Option C: The second message could be from Lula, but it's not necessarily true based on the constraints. This option is not necessarily true. Option D: The third message could be from Hilary, but it's not necessarily true based on the constraints. Option E: Since exactly once Sukanya receives a message from Jerome on the day after receiving one from Hilary, and there are only four messages, the Jerome message on the day after Hilary must be the fourth message. Therefore, the fourth message must be from Jerome.</p> <p><b>Answer:</b> E ✗</p>
LogicNLI	<p><b>Rules:</b> If there is at least one people who is popular, then Curt is not unable. If there is someone who is loyal, then Brock is timid and Lucille is not popular. If there is someone who is timid, then Marcus is unable and Lucille is loyal. Garth being not shiny is equivalent to Jed being not unable and Marcus being light. If someone who is not timid is also loyal, then he is not unable. If there is at least one people who is light or not shiny, then Silas is timid. If Lucille is light, then Brock is not popular. If Curt is not light, then Marcus is popular. Silas is light if and only if Silas is loyal and Brock is not unable. If someone is not timid, then he is loyal. If there is someone who is both unable and not timid, then Jed is shiny. Someone who is both not shiny and unable is always light.</p> <p><b>Facts:</b> Silas is light. Marcus is not timid. Marcus is popular. Marcus is light. Lucille is not unable. Curt is loyal. Brock is unable. Jed is loyal. Curt is unable. Garth is light. Brock is not shiny. Jed is popular.</p> <p><b>Question:</b> Brock is unable.</p>	<p>Given Facts: 1. Silas is light. 2. Marcus is not timid. ... Given Rules: 1. If there is at least one person who is popular, then Curt is not unable. 2. If there is someone who is loyal, then Brock is timid and Lucille is not popular. ... Analysis: From the given facts, we can see that **Fact 7** directly states: Brock is unable. This means that the statement "Brock is unable" is **entailed** by the facts provided. Conclusion: Since the fact "Brock is unable" is explicitly given in the facts, the statement Brock is unable is **entailed** by the information.</p> <p><b>Answer:</b> entailment ✗</p>