

---

# Flow of Reasoning: Training LLMs for Divergent Reasoning with Minimal Examples

---

Fangxu Yu<sup>\*1</sup> Lai Jiang<sup>\*1</sup> Haoqiang Kang<sup>\*1</sup> Shibo Hao<sup>1</sup> Lianhui Qin<sup>1</sup>

## Abstract

The ability to generate diverse solutions to a given problem is a hallmark of human creativity. This divergent reasoning is also crucial for machines, enhancing their robustness and enabling them to assist humans in many applications such as scientific discovery. However, existing approaches to multi-step reasoning with large language models (LLMs) have mostly focused only on reasoning accuracy, without further discovering more diverse valid solutions. For example, supervised fine-tuning improves reasoning quality but requires vast labeled data, while reward-maximizing reinforcement learning finds top-reward solutions while neglecting the solution diversity. To fill this gap, we propose Flow of Reasoning (**FOR**), an efficient diversity-seeking LLM finetuning method aimed at improving reasoning quality and diversity with minimal data. FOR formulates multi-step LLM reasoning as a Markovian *flow* on a DAG-structured reasoning graph. This formulation allows us to incorporate and adapt principled GFlowNet approaches, for finetuning LLMs to sample divergent paths with probabilities *proportional* to the (unnormalized) reward of target problems. Extensive experiments show that, with limited training examples (e.g., 15 examples), FOR enables the discovery of diverse, creative, high-quality solutions, greatly outperforming a wide range of existing inference and training methods across six challenging reasoning tasks, including BlocksWorld (embodied reasoning), Game24 (math puzzle solving), Rubik’s Cube (spatial reasoning), 1D-ARC (abstraction reasoning), GSM8k (math reasoning), and PronToQA (logical reasoning). Code is available at <https://github.com/Yu-Fangxu/FoR>.

---

<sup>\*</sup>Equal contribution <sup>1</sup>University of California San Diego. Correspondence to: Lianhui Qin <lianhui@ucsd.edu>.

## 1. Introduction

Divergent problem solving is the ability to generate multiple diverse solutions to a given problem (Runco, 1991; Runco & Acar, 2012). As a hallmark of human intelligence, this ability drives creativity by uncovering novel ways to accomplish a task, providing more possibilities and adaptivity in different complex situations. Similarly, by encouraging machines to explore diverse solutions rather than confining to one reasoning path, we not only enhance machines’ robustness (e.g., by ranking or aggregating different solutions) (Wang et al., 2022), but also empower automated systems that assist humans in generating ideas and thinking out-of-the-box, thereby potentially facilitating task completion (Shinn et al., 2024), education (Li et al., 2023a), and scientific discovery (Jain et al., 2023a).

State-of-the-art reasoning with large language models (LLMs), however, has largely focused on improving only the reasoning *accuracy* with the topmost solution, without moving a step further to discover more *diverse* valid solutions. Specifically, *inference* methods, such as CoT (chain of thought, Wei et al., 2022), ToT (Yao et al., 2024), RAP (Hao et al., 2023), and others (Chen et al., 2024b; Besta et al., 2024), rely heavily on the underlying pretrained LLM’s capability and decoding algorithms to obtain diverse reasoning solutions. Moreover, the search-based inference (Yao et al., 2024; Hao et al., 2023) can be computationally costly when searching for multiple reasoning paths. On the other hand, the *finetuning* methods like supervised finetuning (SFT) (Yue et al., 2023; Yu et al., 2023c) often demands extensive supervision data to capture the full diversity of solutions, which can be costly to label in many applications. Alternatively, reward-maximization reinforcement learning (RL), such as proximal policy optimization (PPO, Schulman et al., 2017), trains LLMs to generate the highest-reward reasoning solution and overlooks solution diversity. As shown in the case study in Figure 1, the above-mentioned methods generate limited diverse solutions.

To overcome the limitations, we introduce Flow of Reasoning (**FOR**), a data-efficient approach that finetunes LLMs for diverse reasoning with minimal training data. FOR draws inspirations from generative flow networks (GFlowNets) for amortized diverse sampling (Bengio et al.,

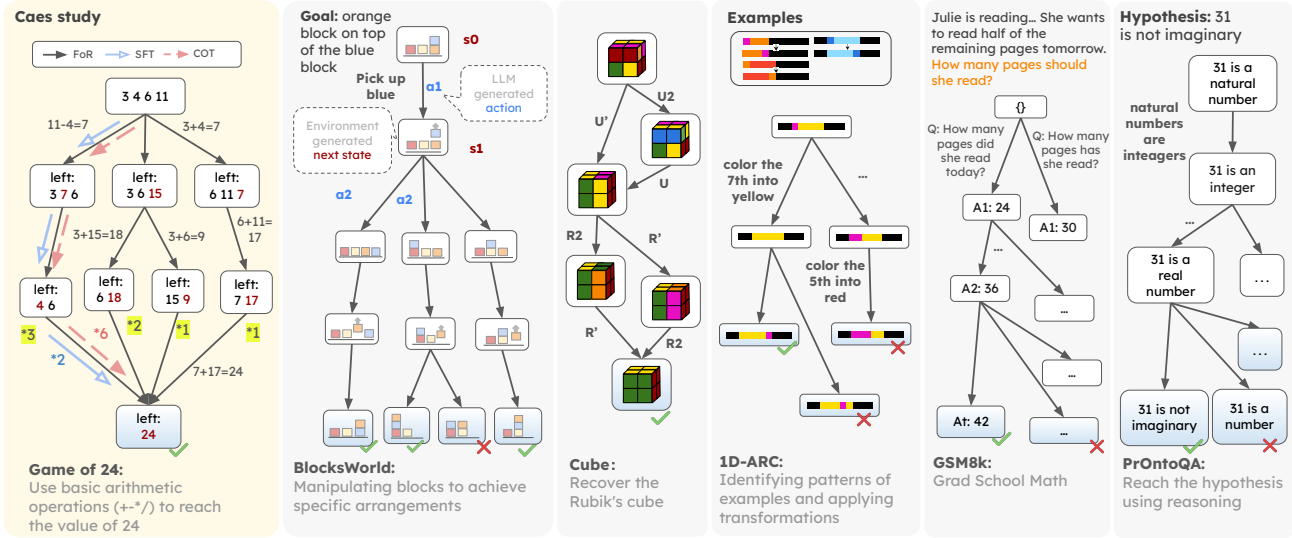


Figure 1: Multi-step LLM reasoning as a Markovian flow on six tasks, forming DAG-structured reasoning graphs. In the example of Game24 (left), we sample 20 reasoning paths from each method, respectively. Baseline methods such as SFT and CoT generate only one valid solution (leftmost path) repetitively (e.g., SFT generates this solution twice out of the 20 attempts), while our method FOR discovers three additional unique solutions.

2021) that have been studied in different domains like molecule synthesis (Koziarski et al., 2024) and operation scheduling (Zhang et al., 2023c). In particular, FOR enables diversity-seeking finetuning of multi-step LLM reasoning, to sample high-quality reasoning paths with probabilities *proportional* to the reward of target problems (as opposed to reward *maximization* in conventional RL like PPO). To this end, we formulate multi-step LLM reasoning from a Markovian flow perspective (Figure 2), where each reasoning step corresponds to an edge (action) that leads to the next node (state) in a flow graph. The reasoning process thus forms a flow that travels step-by-step from an initial state to the terminal states of the target problem. Based on this new formulation, we use the trajectory-balanced objective and adapt efficient exploration methods from the recent GFlowNet studies, enabling finetuning of LLMs to generate more accurate and diverse solutions in reasoning tasks.

FOR differs crucially from the recent GFlowNet applications on autoregressive sequence generation with or without LLMs (Hu et al., 2023a; Malkin et al., 2022a). In particular, contrary to the token-level modeling in the previous work (Hu et al., 2023a), FOR introduces higher-level modeling at the granularity of reasoning steps. This overcomes the limitations of aforementioned search-based LLM reasoning (Yao et al., 2024; Hao et al., 2023) by amortizing inference into training. Additionally, GFlowNets’ exploratory nature (Madan et al., 2024) enhances both reasoning quality and diversity, addressing the limitations of fine-tuning methods like SFT and PPO (see §4).

We evaluate the divergent reasoning capability of our ap-

proach across six challenging benchmarks: BlocksWorld (embodied reasoning; Kambhampati et al., 2024), Game24 (math puzzles; Yao et al., 2024), Rubik’s Cube (spatial reasoning; Ding et al., 2023), 1D-ARC (abstraction; Xu et al., 2023b), GSM8K (math reasoning; Cobbe et al., 2021), and ProntoQA (logical reasoning; Saparov & He, 2022; see Appendix 4.7). Empirical results show that FOR, with limited (e.g. 15) training examples, generates diverse, high-quality solutions, greatly outperforming a wide range of baselines with 20% - 85% improvements, including supervised training methods like SFT, reward-maximizing RL like PPO, diversity-seeking approaches GFN-CoT and various decoding methods, and advanced inference methods like CoT, ToT, GoT, and RAP. Ablation studies further validate the key designs in FOR that lead to robustness and effectiveness.

## 2. Related Work

**Divergent Thinking.** Divergent thinking involves two branches: lateral and vertical (Waks, 1997b). Lateral thinking explores unconventional ideas and challenges assumptions, as seen in brain teasers (Zhong et al., 2024; Summers-Stay et al., 2023), while vertical thinking focuses on structured, step-by-step reasoning. Our work focuses on vertical thinking, aiming to generate multiple different reasoning paths to solve a single problem.

**LLM Reasoning.** Recent LLMs (Achiam et al., 2023) have shown strong potential in tackling complex reasoning tasks (Hu et al., 2023c; Zhang et al., 2023d; Yu et al., 2023b). (1) **Fine-tuning LLMs**, including supervised fine-tuning (SFT) and reward-maximization reinforcement learn-

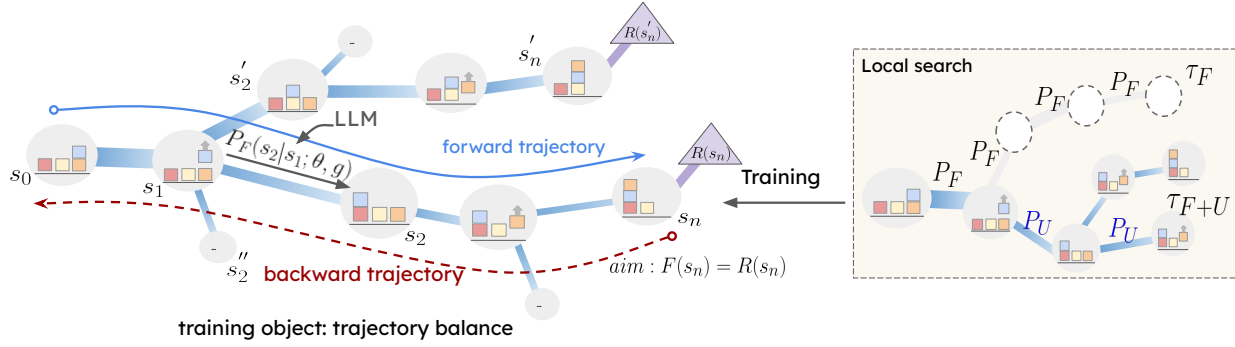


Figure 2: **Left:** The forward policy  $P_F(s_t|s_{t-1}; \theta, g)$  in the flow-based formulation is parameterized as LLM and finetuned with the trajectory balance objective (Eq.5) to achieve the desired flow  $F(s_n) = R(s_n)$  on all terminal states  $s_n$ . **Right:** FoR incorporates local search with a destroy-and-reconstruction process to augment informative trajectories in training (§3.2). This facilitates efficient exploration and improves policy learning.

ing (RL), is a key method for improving LLM reasoning abilities. **SFT**, leveraging large and high-quality datasets of reasoning chains, has proven highly effective (Yu et al., 2023c). **RL** techniques like PPO and DPO are widely used for optimizing reward-driven behavior in LLMs (Ouyang et al., 2022; Bai et al., 2022). However, both approaches tend to limit solution diversity. **(2) Prompting-based methods** engage LLMs in a step-by-step thinking process. Chain-of-Thought (CoT) (Wei et al., 2022) enhances LLMs performance by guiding them through intermediate steps to reach the final answer. Building on CoT, the methods like ToT (Yao et al., 2024) and GoT (Besta et al., 2024) model reasoning as tree and graph searches, and RAP (Hao et al., 2023) and XoT (Ding et al., 2023) use planning approaches to refine reasoning trajectories. DoT (Naik et al., 2023) enhances reasoning diversity through varied prompts.

**GFlowNets.** GFlowNets (Bengio et al., 2021) were developed to generate diverse, high-reward samples from unnormalized distributions (Shen et al., 2023b; Roy et al., 2023; Zhang et al., 2023b), making them particularly effective in domains like molecule synthesis (Kozlowski et al., 2024; Kim et al., 2024a; Lu et al., 2024), where diversity is essential. Recently, GFlowNets have been integrated with LLMs (Hu et al., 2023a; Kwon et al., 2024; Lee et al., 2024b; Song et al., 2024), but none of these have addressed multi-step reasoning problems. While Takase et al. (2024) is a concurrent work, its scope is limited to math problems and does not extend to general reasoning problems. In contrast, FOR extends GFlowNets to multi-step reasoning, modeling it as a Markovian flow through a DAG, enabling the exploration of divergent reasoning paths across six reasoning benchmarks.

### 3. FOR for Divergent Reasoning

We first formulate the divergent multi-step reasoning tasks from a Markovian flow perspective (section 3.1). Then, we

develop an efficient LLM training method using GFlowNet objectives for multi-step reasoning tasks (section 3.2).

#### 3.1. Multi-step LLM Reasoning as Markovian Flow

**Multi-Step Reasoning Problem as DAG.** Consider a multi-step reasoning problem that gives an initial state  $s_0$  and a goal  $g$ . For example, in BlocksWorld (Figure 2), an initial state is the starting configuration of the block stack, and a goal describes the desired configuration of blocks. Reasoning aims to find complete paths (or *trajectories*) that lead from the initial state to the state(s) that satisfy the goal. Given a current state  $s$ , applying an action on it leads to the transition to the next state  $s'$ , denoted as  $s \rightarrow s'$ . For example, in Figure 2, the state  $s_0$  transits to  $s_1$  after an action "pickup blue". A complete trajectory is thus a sequence of transitions  $\tau = (s_0 \rightarrow s_1 \rightarrow \dots \rightarrow s_n) \in \mathcal{T}$ , where  $s_n$  is the terminal state and  $\mathcal{T}$  is the set of all complete trajectories. Given a current state  $s_t$ , there could be multiple alternative next actions, resulting in different branches of the reasoning steps. Also, different sequences of actions can lead to the same intermediate/terminal states, as shown in Figure 1. As a result, we first formulate the multi-step reasoning problem as a directed acyclic graph (DAG).

The reasoning graph consists of diverse trajectories that lead to different terminal states. A crucial component often provided in reasoning tasks is the reward  $R(s_n) \in \mathbb{R}_{\geq 0}$ , which assigns a numerical value to any terminal state  $s_n$ . For instance, a terminal state meeting the goal  $g$  receives a high reward. As discussed in §1, to generate diverse correct reasoning trajectories for solving a task, we want to sample the trajectories with probabilities *proportional* to the reward. This significantly differs from popular reinforcement learning methods (e.g., PPO) and prompting-based planning algorithms (e.g., RAP, ToT), which focus on optimizing for only the maximum-reward trajectory.

**The Flow Perspective.** Sampling complex multi-step trajectories from the (often unnormalized) reward is particularly challenging (LeCun et al., 2006; Qin et al., 2022). To overcome the difficulty, we consider the above reasoning problem from a flow-based viewpoint which was initially developed in (Bengio et al., 2021) and has been studied in other machine learning settings like molecule generation (Li et al., 2023c; Lahlou et al., 2023; Li et al., 2024a; He et al., 2024). Specifically, we define a *trajectory flow* function  $F : \mathcal{T} \rightarrow \mathbb{R}_{\geq 0}$ . Analogous to the classical concept of flows in networks, the flow  $F(\tau)$  can be thought of as the volume of water traveling along this path  $\tau$ . Based on this, for any state  $s$ , we can define the *state flow*  $F(s) = \sum_{s \in \tau} F(\tau)$ , and for any edge  $s \rightarrow s'$ , the *edge flow*  $F(s \rightarrow s') = \sum_{s \rightarrow s' \in \tau} F(\tau)$ . These concepts of (unnormalized) flow are connected to the (normalized) probability distributions. Specifically, the flow trajectory determines a distribution over trajectories:

$$P(\tau) = F(\tau)/Z, \quad Z = \sum_{\tau \in \mathcal{T}} F(\tau). \quad (1)$$

With a Markov assumption, it can be shown that the distribution factorizes into step-wise distributions:

$$P(\tau) = \prod_{t=1}^n P_F(s_t | s_{t-1}), \quad (2)$$

where  $P_F(s_t | s_{t-1}) = F(s_{t-1} \rightarrow s_t) / F(s_{t-1})$ .

That is, intuitively,  $P_F(s_t | s_{t-1})$  characterizes the proportion of water at node  $s_{t-1}$  that travels toward node  $s_t$ . The distribution  $P_F$  is also called the *forward policy*, which can generate a trajectory  $\tau$  by sampling a sequence of transitions step-by-step starting from the initial state  $s_0$ . Equivalently (Bengio et al., 2023), there exists a *backward policy* that defines the distributions  $P_B(\cdot | s_t)$  over the parents of each state  $s_t$ :  $P_B(s_{t-1} | s_t) = F(s_{t-1} \rightarrow s_t) / F(s_t)$ .

Let  $\tau$  be the trajectory ending at the terminal state  $s_n$ . Recall that **our aim** in diverse LLM reasoning is to obtain a forward policy  $P_F(s_t | s_{t-1})$  such that the resulting trajectory distribution is proportional to the reward. From the flow perspective, according to Eqs.(1) and (2), this aim is equivalent to approximating a Markovian flow  $F$  such that  $F(s_n)$  equals the reward (Bengio et al., 2021):

$$F(s_n) = R(s_n), \quad \forall \text{ terminal state } s_n. \quad (3)$$

The above flow-based concepts provide a rich set of constraints that can be converted into training objectives for learning the desired forward policy. For example, the *detailed balance* constraint  $F(s_{t-1})P_F(s_t | s_{t-1}) = F(s_t)P_B(s_{t-1} | s_t)$  yields the respective objective used in molecule generation tasks (Bengio et al., 2023). In this work (§3.2), we devise the learning objective from the recent *trajectory balance* constraint shown to be more efficient (Malkin et al., 2022a). We consider the incorporation of

other more recent extensions (Jang et al., 2023; Pan et al., 2023a) like subtrajectory balance (Madan et al., 2023) as future work.

**LLM Parameterization.** We parameterize the forward policy  $P_F$  with an LLM and finetune as described in the next section. Specifically, for a reasoning task, we express its goal  $g$ , action  $a$ , and state  $s$  as natural language (see Figure 1, BlocksWorld as an example). At each reasoning step  $t$ , the LLM generates an action  $a_t \sim P_{\text{LLM}}(a | s_t; \theta, g, c)$ , where  $c$  is an appropriate prompt (e.g., instructions or in-context demonstrations). The prompts used in the experiments are detailed in Appendix C. Once an action is generated, the state transits to the next  $s_{t+1} = T(s_t, a_t)$  with a transition function  $T$ . Therefore, assuming that different actions applying to the same state  $s_t$  lead to different next states, and that action  $a_t$  leads to state  $s_{t+1}$ , we can write  $P_F(s_{t+1} | s_t; \theta, g) = P_{\text{LLM}}(a_t | s_t; \theta, g, c)$ . In the experiments, we follow previous work and define  $T$  either by an LLM with appropriate prompts and greedy decoding (e.g., BlocksWorld as in (Hao et al., 2023)) or by the environment (e.g., Rubik’s Cube as in (Ding et al., 2023)).

### 3.2. Efficient Diversity-Seeking Training of LLMs

The proposed flow-based formulation of LLM reasoning enables the adaptation from GFlowNet training methods by fine-tuning an LLM as the forward policy  $P_F$ .

**Training Objective.** Since we can sample a complete trajectory, we derive our training objective based on the trajectory balance approach (Malkin et al., 2022a), which has shown improved efficiency than other alternatives (Bengio et al., 2023; 2021). Specifically, for any complete forward trajectory  $\tau = (s_0 \rightarrow s_1 \rightarrow \dots \rightarrow s_n)$ , the trajectory balance constraint, with a task goal  $g$ , says (Figure 2):

$$Z(s_0, g) \prod_{t=1}^n P_F(s_t | s_{t-1}; g) = F(s_n) \prod_{t=1}^n P_B(s_{t-1} | s_t; g), \quad (4)$$

where we have used the fact that  $P(s_n) = F(s_n) / Z(s_0, g)$  for the terminal state  $s_n$ . Plugging in the reward  $R$ , as motivated by Eq.(3), to provide supervision signals, the constraint leads to a loss function w.r.t the parameterized forward policy  $P_F$ :

$$l(\tau; \theta, g) = \left( \log \frac{Z(s_0, g) \prod_{t=1}^n P_F(s_t | s_{t-1}; \theta, g)}{R(s_n) \prod_{t=1}^n P_B(s_{t-1} | s_t; \theta, g)} \right)^2, \quad (5)$$

$$P_B(s_{t-1} | s_t; \theta, g) := \frac{1}{|\text{Pa}(s_t)|}$$

where  $|\text{Pa}(s_t)|$  denotes the number of parents of state  $s_t$ , and (Malkin et al., 2022a) suggested a canonical choice of setting  $P_B(\cdot | s_t)$  to be uniform over the parents. Note that  $Z$  is the total flow conditioning on each goal  $g$  and initial state  $s_0$ . The term  $\log Z$  is challenging to learn accurately in an LLM setting (Hu et al., 2023a). We thus follow (Zhang



et al., 2023c) to use the log-variance approximation, which implicitly estimates  $\log Z$  given each trajectory  $\tau$ :

$$\Phi(\tau; \theta) = \log R(s_n) + \sum_{t=1}^n \log P_B(s_{t-1}|s_t; \theta, g) - \sum_{t=1}^n \log P_F(s_t|s_{t-1}; \theta, g), \quad (6)$$

where  $\Phi(\tau; \theta)$  equals to true  $\log Z$  in the optimal case. Our optimization goal then turns into minimizing the variance of  $\Phi(\tau; \theta)$  over different trajectories  $\tau$  with the loss:

$$\mathcal{L}_V(\tau; \theta) = (\Phi(\tau; \theta) - \mathbb{E}_\tau[\Phi(\tau; \theta)])^2, \quad (7)$$

where we sample trajectories  $\tau$  from a behavior policy  $\pi(\tau; \theta, g)$  for training, and  $\mathbb{E}_\tau[\Phi(\tau; \theta)]$  is estimated with a mini-batch of sampled trajectories. We have different configurations of  $\pi$ , such as on-policy, off-policy, and mixed explorations, which could impact training efficiency (see ablation studies (§4)). If  $\mathcal{L}_V(\tau; \theta)$  is globally optimized, the resulting flow satisfies Eq.(3) and  $P_F(\cdot|\cdot; \theta, g)$  samples proportionally to the reward as desired.

**Efficient Exploration.** Training with the loss function in Eq. (7) requires trajectory collection, raising the challenge of how to collect them efficiently and effectively. On-policy sampling, which generates trajectories via  $\tau \sim P_F(\tau; g)$  and optimizes  $\mathcal{L}_V(\tau; \theta)$  with a batch of trajectories, often struggles to explore the vast trajectory space due to limited exploration capabilities. This can result in inadequate estimation of the distribution needed to capture diverse reasoning paths. To address this, we want to set up a distribution  $\pi(\tau, g; \theta)$  in Eq. (7) that enhances exploration while producing various training samples for policy optimization. Inspired by recent advances in GFlowNets (Vemgal et al., 2023; Hu et al., 2023a), we combine on-policy and off-policy strategies and further enhance exploration by adapting local search (Kim et al., 2023) for LLM training.

For on-policy exploration, we use both the online policy  $P_F(s_t|s_{t-1}; \theta, g)$  and a tempered variant to generate trajectories given a goal  $g$  and initial state  $s_0$ . Off-policy exploration leverages a replay buffer prioritizing high-reward trajectories,  $\epsilon$ -sampling, and offline trajectory data (for Game24 in §4.3). To efficiently explore high-reward regions, we propose a modified local search method (Figure 2). Specifically, we select the highest-reward trajectory from each batch, truncate its latter portion, and reconstruct it using a random policy  $P_U$ . This avoids the computationally expensive forward process of LLMs, enhancing efficiency. High-reward trajectories are more likely to reconstruct correctly in early steps, with errors typically occurring later. Additional details on the full algorithm, exploration strategies, and local search are in Appendices E and F.

## 4. Experiments

### 4.1. Experimental Settings

**Baselines.** We first compare our approach with several prompting-based methods, including  $k$ -shot CoT (Wei et al., 2022), ToT (Yao et al., 2024), GoT (Besta et al., 2024), XoT (Ding et al., 2023), and RAP (Hao et al., 2023). We use the Llama-3-8B (Dubey et al., 2024) as the base model for all tasks. For fine-tuning-based methods, we evaluate SFT with diversity-enhancing decoding strategies like Temperature Sampling (Shi et al., 2024), Nucleus Sampling (Holtzman et al., 2019) ( $\eta = 0.95$ ), Typical Sampling (Meister et al., 2023) ( $\tau = 0.95$ ), and diverse beam search (Vijayakumar et al., 2016) (DBS, beam width  $k = 20$ ). Additionally, we apply fine-tuning with PPO (Schulman et al., 2017) and GFN-CoT (Hu et al., 2023a). We also compare it against OpenAI-O1, which is the latest and strongest reasoning model. All finetuning methods are trained on the same dataset as FOR. In the BlocksWorld task, we evaluated most baselines for a broad comparison, but some methods (e.g., GFN-CoT) showed suboptimal performance, which informed our decision to selectively apply high-performing baselines in the subsequent tasks.

**Evaluation.** We generate  $n$  solutions per problem and evaluate methods based on four key metrics: (1) **Accuracy (Acc)**: A problem is considered correct if at least one of the  $n$  sampled solutions is correct. (2) **Diversity**: For problems with at least one correct solution, we report the average number of semantically different correct solutions among the number of correct solutions, where higher is better (see Appendix for details). (3) **Creativity**: The proportion of unique correct solutions found by a method that are not discovered by others. (4) **Runtime**: The average time to generate one solution by using a single NVIDIA A100 80GB GPU is used as an efficiency metric. See Appendix C.1 and C.2 for details of metrics.<sup>1</sup>

### 4.2. Embodied Reasoning: BlocksWorld

BlocksWorld involves a set of blocks with unique colors that can be stacked on top of each other or moved around. The goal of this task is to enable LLMs to plan a sequence of actions to transform an initial configuration of blocks into a desired goal configuration using a series of actions. The actions are text instructions (STACK, UNSTACK, PUT, PICKUP) generated based on domain rules and block positions, and a state is the current block orientation.

**Setup.** Blocksworld examples (Valmeekam et al., 2024) are grouped by the minimum number of required actions: 30

<sup>1</sup>In our experiments, we report the average accuracy and standard deviation over three repetitions, except for search-based methods and O1-series due to time and budget constraints. Creativity is not averaged over repetitions, as the performance exhibit low variance across runs.

Table 1: Results on BlocksWorld, comparing *prompting-based* and *finetuning-based* methods on questions requiring two, four, and six steps, respectively. The standard deviations of three runs are shown in brackets.  $\alpha$  is a temperature. ♣ indicates that results from multiple runs for ToT (DFS).

Method	2-step		4-step		6-step			Runtime (s) (6-step)
	Acc. (%)	Acc. (%)	Diversity	Creativity (%)	Acc. (%)	Diversity	Creativity (%)	
Prompting-based methods								
CoT (1-shot)	48.88 (8.31)	28.57 (5.83)	1.05 (0.04)	0.00	15.82 (2.08)	1.05 (0.03)	0.00	3.57
CoT (5-shot)	68.89 (8.31)	42.86 (1.94)	1.04 (0.03)	0.00	29.63 (1.72)	1.02 (0.01)	0.00	3.68
CoT (15-shot)	64.44 (6.29)	42.06 (4.89)	1.03 (0.02)	0.00	19.53 (1.26)	1.03 (0.03)	0.00	5.32
CoT (GPT-4o, 1-shot)	93.33	54.76	1.08	0.00	67.67	1.06	0.79	3.92
ToT (DFS)	13.33	16.67	-	-	8.08	-	-	48.91
ToT (DFS) <sup>✱</sup>	40.00	42.85	1.0	0.00	31.31	1.06	1.58	44.41
ToT (BFS)	13.33	14.28	-	-	5.05	-	-	398.74
RAP	<b>100.00</b>	92.86	-	-	69.70	-	-	466.09
O1-series methods								
O1-mini (1-shot) *	100.00	100.00	1.05	0.00	93.93	1.05	2.38	10.38
O1-preview (1-shot) *	100.00	95.24	-	-	78.79	-	-	36.61
Finetuning-based methods								
SFT (α=1.0)	44.44 (3.14)	42.06 (5.44)	1.05 (0.01)	0.00	34.68 (2.52)	1.04 (0.01)	4.76	4.05
SFT (α=0.5)	42.22 (3.14)	39.68 (2.24)	1.02 (0.02)	0.00	29.63 (1.90)	1.02 (0.02)	0.79	4.07
SFT (α=0.1)	26.67 (5.44)	26.20 (3.89)	1.00 (0.00)	0.00	17.51 (1.26)	1.00 (0.00)	0.00	4.08
SFT + DBS	31.10 (3.11)	38.88 (1.12)	1.00 (0.00)	0.00	18.85 (1.25)	1.00 (0.00)	0.00	15.71
SFT + Nucleus	48.89 (3.14)	53.97 (2.97)	1.04 (0.03)	0.00	42.08 (1.71)	1.12 (0.03)	0.00	4.21
SFT + Typical	53.33 (5.44)	48.41 (2.25)	1.08 (0.02)	0.00	37.71 (2.38)	1.08 (0.02)	0.00	3.65
SFT + PPO	46.66 (5.44)	44.44 (2.24)	1.11 (0.05)	2.04	24.58 (1.72)	1.08 (0.03)	3.17	4.03
SFT + GFN-CoT	48.89 (8.81)	44.42 (2.96)	1.00 (0.00)	0.00	40.73 (1.25)	1.05 (0.03)	0.00	4.08
For (Ours)	<b>100.00</b> (0.00)	<b>98.41</b> (1.12)	<b>1.27</b> (0.02)	<b>12.24</b>	<b>78.44</b> (4.54)	<b>1.33</b> (0.03)	<b>9.52</b>	13.98

examples for 2 steps, 57 for 4 steps, and 114 for 6 steps, following Hao et al. (2023). We select the first 15 of each group as the training examples for FOR and the rest as test examples. We sampled 8, 20, and 40 times for the 2, 4, and 6-step datasets, respectively, to report diversity and creativity. All the baselines are included in §4.1.

**Reward Design.** We compose a terminal state reward and an augmented intermediate reward to evaluate trajectories. Terminal state reward is assigned to a high positive value when a trajectory reaches the goal  $g$ . The augmented intermediate reward assesses actions by using the LLM to estimate the confidence of actions in achieving their goals. A natural choice is to use the log-likelihood of actions,  $\log P_{\text{LLM}}(a_t|s_{t-1}, g)$ . However, the value of  $\log P$  is negative. To maintain monotonicity consistency and positive reward values, we use  $-1/\log P_{\text{LLM}}(a_t|s_{t-1}, g)$  instead. The total reward is defined as:  $R(s_n) = w \cdot \mathbb{I}(\text{success}) + \lambda \sum_{t=1}^n -1/\log P_{\text{LLM}}(a_t|s_{t-1}, g)$ , where  $w$  is the success weight (set to 100 and the following tasks).

**Results.** As shown in Table 1, Our method outperforms all other baselines across all metrics, showing that FOR not only achieves the best accuracy but also finds more diverse and unique solutions. Specifically, most baselines do not find any unique solutions, resulting in a creativity score of 0. It is worth noting that the O1 series improves the accuracy to a large extent, but still struggles to find diverse reasoning paths. Additionally, FOR matches the inference speed of high-efficiency baselines like  $k$ -shot CoT, ToT (DFS),

Table 2: Results on Game of 24.

Method	Acc. (%)	Diversity	Creativity (%)
<i>Prompting-based methods</i>			
CoT (5-shot)	23.00	1.04	6.60
CoT (GPT-4o, 5-shot)	<b>59.00</b>	<b>1.61</b>	<b>49.42</b>
XoT	20.00	-	-
ToT	21.00	-	-
RAP	10.00	-	-
<i>O1-series methods</i>			
OpenAI-O1-mini	94.00	-	-
<i>Finetuning-based methods</i>			
SFT ( $\alpha = 1.0$ )	29.00	1.18	11.22
FOR	48.00 (1.41)	1.30 (0.09)	29.15

and SFT-based methods, while being 30× faster than ToT (BFS) and RAP. Additional results of diversity-encouraging prompts, larger-size models, and out-of-distribution (OOD) generalization are in Appendix C.4, and training costs are detailed in Appendix C.3.

### 4.3. Math Puzzle Solving: Game of 24

Game of 24 is a mathematical reasoning task that may have multiple solutions. The objective of this task is to use 4 integers and 4 basic arithmetic operations (+, −, ×, ÷) to reach 24, where each number can only be used once. Each action  $a_t$  is defined as an equation composed of 2 numbers and an operator, and the state  $s_t$  is the left number.

**Setup.** We use the LLM-reasoner dataset (Hao et al., 2024)

Table 3: Results on Rubik’s Cube.

Method	Acc. (%)	Diversity	Creativity (%)
<i>Prompting-based methods</i>			
CoT	0.00	0.00	0.00
CoT (GPT-4)	1.09	1.00	4.35
ToT (BFS)	0.00	-	-
GoT	0.00	-	-
XoT	4.92	-	-
<i>Finetuning-based methods</i>			
SFT ( $\alpha = 1.0$ )	1.82 (0.06)	1.00 (0.00)	8.69
SFT + PPO	0.55 (0.45)	1.00 (0.00)	0.00
FOR	<b>10.87</b> (1.18)	<b>1.29</b> (0.02)	<b>82.61</b>

and randomly select 20 examples for training and 100 examples for testing. Since prior work shows that LLMs struggle to sample correct trajectories online in Game24 (Yao et al., 2024; Yu et al., 2023a), we generate offline ground-truth trajectories using Python code for fine-tuning. We sample 20 solutions for all methods.

**Reward Design.** Similar to BlocksWorld, the success reward gives a high positive reward when a trajectory  $\tau$  succeeds in reaching 24, and the augmented reward gives the product of the probability of correctness for each action  $a_t$ , given its last state  $s_{t-1}$  provided by the untrained LLM model:  $R(s_n) = w \cdot \mathbb{I}(\text{success}) + \prod_{t=1}^n P_{\text{untrained}}(a_t | s_{t-1})$ .

**Results.** As shown in Table 2, FOR demonstrates superior accuracy and diversity in solving math puzzles compared to other baselines with the same base model. Not surprisingly, O1-mini and GPT-4o achieve better performance due to the stronger intrinsic mathematical knowledge and reasoning mechanism. We also investigate the fact that GPT-4o tends to use self-verification and reflection during Game24’s inference. This may explain its superior performance in Game24.

#### 4.4. Spatial Reasoning: Rubik’s Cube

The Rubik’s Cube is a well-known puzzle requiring multi-step spatial reasoning. The model’s task is to plan a sequence of rotations to restore a shuffled cube, where each state  $s_t$  represents the block arrangement, and each action  $a_t$  is a layer rotation (e.g., 90 or 180 degrees).

**Setup.** We randomly select 15 examples from the training dataset in from (Ding et al., 2023), and evaluate different methods on a test set containing 183 examples. Each example can be solved in four steps. All the baselines are included in §4.1. For SFT, CoT and FOR, 10 solutions are sampled. See more details in Appendix C.6.

**Reward Design.** Similar to previous tasks, we have a success reward and an augmented reward that is based on the difference in minimum steps required to restore the cube from its current state. Actions that reduce steps receive higher rewards, while others receive lower rewards. Formally,  $R(s_n) = w \cdot \mathbb{I}(\text{success}) + \sum_{t=1}^n \exp(r(s_{t-1}) - r(s_t))$ , where  $r(s_t)$  represents the remaining minimum steps.

Table 4: Results on 1D-ARC.

Method	Acc. (%)	Diversity	Creativity (%)
IO	10.37 (1.21)	-	-
CoT	39.51 (1.94)	1.04 (0.01)	1.45
CoT (GPT-4o)	40.00	1.00	0.00
Program-Only	0.74	-	-
Hypo-Search	1.48	-	-
FOR	<b>50.37</b> (1.60)	<b>1.17</b> (0.02)	<b>21.74</b>

**Results.** As shown in Table 3, FOR outperforms all baselines on the Rubik’s Cube task, achieving absolute increase of 5.95% in accuracy than XoT, generating an absolute 29% increase in diverse solutions, and discovering unique and creative solutions than the baselines.

#### 4.5. Abstraction Reasoning: 1D-ARC

1D-ARC is a one-dimensional simplification of the ARC benchmark (Chollet, 2019), introduced in (Xu et al., 2023b). Each problem in the dataset contains a set of training input-output 1D grid pairs that capture a specific underlying rule and a test case that measures if the model correctly understands the rule. Following recent program search approaches (Wang et al., 2023b; Qiu et al., 2023; Butt et al., 2024), which frame the problem as a sequence of transformation functions, each action  $a_t$  is a function (e.g., horizontal mirroring), with the intermediate grid as state  $s_t$ .

**Setup.** We randomly select 5 examples from the 1d\_move\_1p, 1d\_padded\_fill, and 1d\_denoising tasks. These tasks involve moving the color bar by 1 pixel, filling in empty spaces enclosed by pixels, and removing noise-like pixels respectively. The 15 selected examples form the training set, while the remaining 45 examples from each task form the test dataset. We sample 20 solutions for each example during inference. See Appendix C.7 for details.

**Baselines.** Since there are no complex reasoning baselines (e.g., ToT) evaluated on this task, we compare against *Input-output (IO) prompting* (Mirchandani et al., 2023). *Program Only* and *Hypothesis Search* (Wang et al., 2023b) synthesize Python programs for transformation, with the latter first generating language-based transformation hypotheses before program synthesis. Fine-tuning methods are not compared due to the lack of labeled reasoning data.

**Reward Design.** In addition to a success reward, we design the augmented rewards for actions based on how much they reduce the distance to the goal. Specifically, an action receives a higher reward if it reduces the distance between the current state and the goal. The total reward is  $R(s_n) = w \cdot \mathbb{I}(\text{success}) + \sum_{t=1}^n \sum_{i=1}^K \exp(h_d(s_{t-1}^i, g^i) - h_d(s_t^i, g^i))$ , where  $h_d(\cdot, \cdot)$  is hamming distance and  $K$  is the number of training input-output pairs.

**Results.** Results in Table 4 show that FOR substantially outperforms previous methods on all metrics, especially

Table 5: Ablation results on BlocksWorld for different components in FoR with Llama-3-8b.

Method	2-step	4-step		6-step	
	Acc. (%)	Acc. (%)	Diversity	Acc. (%)	Diversity
FoR (Ours)	<b>100.00</b> (0.00)	<b>98.41</b> (1.12)	<b>1.27</b> (0.02)	<b>78.44</b> (4.54)	1.33 (0.03)
- w/o local search	100.00 (0.00)	89.68 (2.97)	1.18 (0.02)	53.90 (2.10)	1.31 (0.03)
- w/o augmented rewards	100.00 (0.00)	91.30 (1.10)	1.22 (0.02)	47.10 (1.30)	1.21 (0.01)
- w/o replay buffer	100.00 (0.00)	94.44 (2.97)	1.24 (0.04)	72.38 (1.71)	1.24 (0.01)
- w/o $\epsilon$ -sampling	100.00 (0.00)	97.61 (1.95)	1.26 (0.03)	73.39 (2.38)	1.25 (0.04)

Table 6: Results on GSM8K.

Method	Acc. (%)	Diversity-H	Diversity-G
CoT (2-shot)	45.72	1.12	1.60
CoT-SC (2-shot)	41.74	-	-
RAP	37.16	-	-
SFT ( $\alpha = 1.0$ )	52.69	1.13	1.63
<b>FoR</b>	57.39	1.26	1.72
<b>FoR + PRM</b>	<b>62.62</b>	<b>1.31</b>	<b>1.77</b>

diversity and creativity. Previous approaches (hypothesis search and program-only) generate programs at once, which easily results in errors during the intermediate process, leading to inferior performance.

#### 4.6. Mathematical Reasoning: GSM8K

GSM8K consists of grade-school math problems (Cobbe et al., 2021). Following RAP (Hao et al., 2023), we define an action  $a_t$  as an intermediate sub-question and a state as the history of all sub-questions and their corresponding answers.

**Setup.** We use the last 50 training examples in the original training dataset, and we sample 4 times for every problem at inference. We follow the baseline implementations from LLM reasoners (Hao et al., 2024). Since there is no standard metric for reasoning diversity, we manually annotate 50 test examples to evaluate the semantic difference between solutions as the diversity metric (Diversity-H), and propose an automatic metric using GPT-4o to distinguish different solutions (Diversity-G), where the prompt is provided in Appendix C.9.

**Reward Design.** We removed the success reward since the pretrained mathematical knowledge enables LLMs to perform relatively well compared to BlocksWorld. Using LLMs to verify each state is computationally expensive. The reward is defined as  $R(s_n) = w \cdot \mathbb{I}(\text{success})$ . In addition, we augment the above reward with the process reward model (PRM) using Qwen2.5-Math-PRM-7B (Zhang et al., 2025) to evaluate the intermediate reasoning steps:  $R(s_n) = w \cdot \mathbb{I}(\text{success}) + \lambda \sum_{t=1}^n \text{PRM}(a_t | s_{t-1}, g)$ .

**Results.** As shown in the table 6, in the task of GSM8K, FoR outperforms all baselines in both accuracy and diversity. These results highlight the potential of FoR for extending to more open-ended reasoning tasks.

Table 7: PrOntoQA Results. *Pred Acc* measures the accuracy of the final conclusions, while *Proof Acc* evaluates the correctness of the reasoning process (e.g., no short-cuts/hallucinations).

Method	In-Distribution		Out-of-Distribution	
	Pred Acc.(%)	Proof Acc.(%)	Pred Acc.(%)	Proof Acc.(%)
<i>Prompting-based methods</i>				
CoT	52.20 (1.23)	35.40 (1.86)	43.50 (1.48)	18.50 (1.91)
CoT (GPT-4o)	89.00	47.80	62.92	24.78
ToT (BFS)	49.80	32.20	-	-
RAP	50.70	39.50	-	-
<i>Finetuning-based methods</i>				
STaR	88.90	54.00	50.10	24.60
FoR	88.73 (1.33)	54.60 (1.50)	<b>63.07</b> (1.71)	<b>28.88</b> (2.36)
FoR + STaR	<b>90.50</b> (1.89)	<b>54.70</b> (1.41)	63.00 (2.13)	26.67 (2.80)

#### 4.7. Logical Reasoning: PrOntoQA

PrOntoQA is a logical reasoning task. Each test case includes a question (goal), a list of facts  $\mathcal{A}$  (action space), and an initial state  $s_0$ . A state  $s_t$  is the conclusion derived from the previous state  $s_{t-1}$ . Performance is evaluated using two metrics: prediction accuracy and proof accuracy. Prediction accuracy refers to the correctness of the final answer, regardless of the reasoning process. Proof accuracy, on the other hand, measures the correctness of the entire reasoning chain, ensuring that each step leading to the final answer is accurate. Both metrics are calculated using rule-based string matching. The Diversity metric is not applicable, as each question has only one valid reasoning chain.

**Setup.** We randomly select 50 examples for the training set and 120 for the test set. The evaluation is conducted on both in- and out-of-distribution (OOD) examples, with 32 samples drawn per problem during inference. In addition to the baselines described in §4.1, we adopt STaR (Zelikman et al., 2022), which applies SFT on correct examples through online sampling. We also evaluate FoR on top of the model fine-tuned by STaR. See Appendix C.8 for more experimental details.

**Reward Design.** We removed the success reward to prevent the model from arriving at correct answers through flawed reasoning paths. Instead, we apply a rule-based augmented reward that evaluates the feasibility of a fact given the previous state  $s_{t-1}$ , checking if they share the same ontology. Formally, the reward is defined as  $R(s_n) = \frac{1}{n} \sum_{t=1}^n w \cdot \mathbb{I}(s_{t-1}, s_t)$ , where  $w$  is a hyperparameter and



Table 8: Ablation results of FOR on BlocksWorld without local search.

Method	2-step	4-step		6-step	
	Acc. (%)	Acc. (%)	Diversity	Acc. (%)	Diversity
Ours - w/o local search	100.00	89.68	1.18	53.90	1.31
- w/o augmented rewards	100.00	71.43	1.04	30.30	1.17
- w/o replay buffer	100.00	78.57	1.10	34.34	1.15
- w/o $\epsilon$ -sampling	100.00	83.33	1.12	49.49	1.14

$\mathbb{I}(s_{t-1}, s_t)$  is an indicator function.  $\mathbb{I}(s_{t-1}, s_t)$  equals 1 only when the transition  $(s_{t-1}, s_t)$  is part of the ground-truth reasoning path, ensuring no shortcuts are taken.

#### 4.7.1. RESULTS

As shown in Table 7, FOR achieves superior results on both in- and out-of-distribution problems compared to all baselines. While FOR slightly outperforms the SFT-based STaR for in-distribution tasks, its advantage is far greater for out-of-distribution tasks. Moreover, combining FOR with STaR enhances in-distribution performance while preserving out-of-distribution success, revealing the complementary strengths of these methods.

#### 4.8. Additional Analysis

**Ablation Study.** To further demonstrate the effectiveness of FOR, we conduct ablation studies to analyze the impact of individual components, focusing specifically on the BlocksWorld task. Table 5 and Table 8 summarize our component-wise ablations:

1) Local search enhances exploration and high-reward trajectory collection—its removal causes a 31.3% absolute drop in 6-step accuracy; 2)  $\epsilon$ -sampling also aids exploration, though to a lesser degree; 3) Augmented intermediate rewards are critical—removing them yields a 51% decrease in 6-step accuracy and a 13% reduction in diversity, and tuning the reward weight  $\lambda$  (left plot of Fig. 3) further refines performance (overly large  $\lambda$  harms accuracy); 4) The replay buffer leverages past high-reward trajectories—dropping it incurs a 4–6% loss with local search versus an 11–19% loss without, highlighting its greater importance when search is absent. To isolate these effects, we also remove local search and reassess each component: as shown in Table 8, the replay buffer’s impact grows without search (11–19% drop vs. 4–6%), and eliminating augmented rewards still causes the largest performance degradation, underscoring their effectiveness in FOR.

**Data-Efficiency.** We test the final 20 examples from the 6-step test set and adjust the number of training examples for both FOR and SFT. As illustrated in the right plot of Figure 3, SFT’s accuracy improves with additional training data, while its diversity remains stable. However, SFT’s performance remains lower than FOR’s under any amount of training data. This is attributed to FOR’s ability to learn

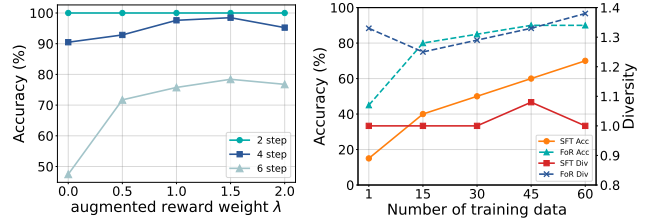


Figure 3: Additional analysis on BlocksWorld. **Left:** Accuracy of FOR across different step settings with varying intermediate reward weight ( $\lambda$ ). **Right:** Comparison of accuracy and diversity between SFT trained with varying data sizes and FOR trained on the test dataset.

from divergent reasoning trajectories, enhancing trajectory coverage and improving generalization to new cases.

**Discussion.** FOR achieves higher reasoning accuracy and diversity by leveraging its exploratory nature (Madan et al., 2024) to sample multiple valid trajectories, covering a broader solution space without getting stuck on incorrect paths like SFT. This increases the likelihood of reaching a correct solution. A detailed case study that compares FOR with SFT is provided in Appendix G. Furthermore, FOR demonstrates robust performance across varying training set sizes, different LLM families, and larger model scales (see Appendix D).

## 5. Conclusion

We introduce FOR, which efficiently trains LLMs to generate diverse, high-quality reasoning paths proportional to unnormalized rewards. By leveraging a flow-based multi-step reasoning formulation, FOR adapts and improves GFlowNets training strategies. Across six reasoning tasks, FOR outperforms finetuning- and prompting-based baselines in both accuracy and diversity. Limitations are discussed in Appendix I.

## Impact Statement

This study introduces FOR, a framework that enhances LLMs’ ability to solve multi-step reasoning problems. This approach lays the foundation for scalable, principled LLM reasoning, with potential applications in automated theorem proving, robotics, and scientific discovery, contributing to more robust and interpretable AI systems.

## References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.
- Bansal, H., Hosseini, A., Agarwal, R., Tran, V. Q., and Kazemi, M. Smaller, weaker, yet better: Training llm reasoners via compute-optimal sampling. *arXiv preprint arXiv:2408.16737*, 2024.
- Barke, S., Gonzalez, E. A., Kasibatla, S. R., Berg-Kirkpatrick, T., and Polikarpova, N. Hysynth: Context-free llm approximation for guiding program synthesis. *arXiv preprint arXiv:2405.15880*, 2024.
- Bengio, E., Jain, M., Korablyov, M., Precup, D., and Bengio, Y. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394, 2021.
- Bengio, Y., Lahlou, S., Deleu, T., Hu, E. J., Tiwari, M., and Bengio, E. Gflownet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.
- Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17682–17690, 2024.
- Butt, N., Manczak, B., Wiggers, A., Rainone, C., Zhang, D., Defferrard, M., and Cohen, T. Codeit: Self-improving language models with prioritized hindsight replay. *arXiv preprint arXiv:2402.04858*, 2024.
- Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., Chu, P., Dong, X., Duan, H., Fan, Q., Fei, Z., Gao, Y., Ge, J., Gu, C., Gu, Y., Gui, T., Guo, A., Guo, Q., He, C., Hu, Y., Huang, T., Jiang, T., Jiao, P., Jin, Z., Lei, Z., Li, J., Li, J., Li, L., Li, S., Li, W., Li, Y., Liu, H., Liu, J., Hong, J., Liu, K., Liu, K., Liu, X., Lv, C., Lv, H., Lv, K., Ma, L., Ma, R., Ma, Z., Ning, W., Ouyang, L., Qiu, J., Qu, Y., Shang, F., Shao, Y., Song, D., Song, Z., Sui, Z., Sun, P., Sun, Y., Tang, H., Wang, B., Wang, G., Wang, J., Wang, J., Wang, R., Wang, Y., Wang, Z., Wei, X., Weng, Q., Wu, F., Xiong, Y., Xu, C., Xu, R., Yan, H., Yan, Y., Yang, X., Ye, H., Ying, H., Yu, J., Yu, J., Zang, Y., Zhang, C., Zhang, L., Zhang, P., Zhang, P., Zhang, R., Zhang, S., Zhang, S., Zhang, W., Zhang, W., Zhang, X., Zhang, X., Zhao, H., Zhao, Q., Zhao, X., Zhou, F., Zhou, Z., Zhuo, J., Zou, Y., Qiu, X., Qiao, Y., and Lin, D. Internlm2 technical report, 2024.
- Chen, Q., Zhang, B., Wang, G., and Wu, Q. Weak-eval-strong: Evaluating and eliciting lateral thinking of llms with situation puzzles. *arXiv preprint arXiv:2410.06733*, 2024a.
- Chen, W., Ma, X., Wang, X., and Cohen, W. W. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022.
- Chen, Y. and Mauch, L. Order-preserving gflownets. *arXiv preprint arXiv:2310.00386*, 2023.
- Chen, Z., White, M., Mooney, R., Payani, A., Su, Y., and Sun, H. When is tree search useful for llm planning? it depends on the discriminator. *arXiv preprint arXiv:2402.10890*, 2024b.
- Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- de Miquel, A. Arc\_kaggle - abstraction and reasoning challenge kaggle competition code. [https://github.com/alejandrodemiquel/ARC\\_Kaggle](https://github.com/alejandrodemiquel/ARC_Kaggle), 2021. Accessed: 2024-09-29.
- Deleu, T. and Bengio, Y. Generative flow networks: a markov chain perspective. *arXiv preprint arXiv:2307.01422*, 2023.
- Deleu, T., Góis, A., Emezue, C., Rankawat, M., Lacoste-Julien, S., Bauer, S., and Bengio, Y. Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence*, pp. 518–528. PMLR, 2022.
- Deleu, T., Nishikawa-Toomey, M., Subramanian, J., Malkin, N., Charlin, L., and Bengio, Y. Joint bayesian inference of graphical structure and parameters with a single generative flow network. *Advances in Neural Information Processing Systems*, 36, 2024.

- Ding, R., Zhang, C., Wang, L., Xu, Y., Ma, M., Zhang, W., Qin, S., Rajmohan, S., Lin, Q., and Zhang, D. Everything of thoughts: Defying the law of penrose triangle for thought generation. *arXiv preprint arXiv:2311.04254*, 2023.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Feng, X., Wan, Z., Wen, M., Wen, Y., Zhang, W., and Wang, J. Alphazero-like tree-search can guide large language model decoding and training. *arXiv preprint arXiv:2309.17179*, 2023.
- Ghari, P. M., Tseng, A., Eraslan, G., Lopez, R., Biancalani, T., Scalia, G., and Hajiramezanali, E. Generative flow networks assisted biological sequence editing. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.
- Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., and Hu, Z. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.
- Hao, S., Gu, Y., Luo, H., Liu, T., Shao, X., Wang, X., Xie, S., Ma, H., Samavedhi, A., Gao, Q., et al. Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. *arXiv preprint arXiv:2404.05221*, 2024.
- Havrilla, A., Du, Y., Raparthy, S. C., Nalmpantis, C., Dwivedi-Yu, J., Zhuravinskyi, M., Hambro, E., Sukhbaatar, S., and Raileanu, R. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*, 2024.
- He, H., Chang, C., Xu, H., and Pan, L. Looking backward: Retrospective backward synthesis for goal-conditioned gflownets. *arXiv preprint arXiv:2406.01150*, 2024.
- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- Hernandez-Garcia, A., Saxena, N., Jain, M., Liu, C.-H., and Bengio, Y. Multi-fidelity active learning with gflownets. *arXiv preprint arXiv:2306.11715*, 2023.
- Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.
- Hosseini, A., Yuan, X., Malkin, N., Courville, A., Sordoni, A., and Agarwal, R. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*, 2024.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- Hu, E. J., Jain, M., Elmoznino, E., Kaddar, Y., Lajoie, G., Bengio, Y., and Malkin, N. Amortizing intractable inference in large language models. *arXiv preprint arXiv:2310.04363*, 2023a.
- Hu, E. J., Malkin, N., Jain, M., Everett, K. E., Graikos, A., and Bengio, Y. Gflownet-em for learning compositional latent variable models. In *International Conference on Machine Learning*, pp. 13528–13549. PMLR, 2023b.
- Hu, Y., Yang, H., Lin, Z., and Zhang, M. Code prompting: a neural symbolic method for complex reasoning in large language models. *arXiv preprint arXiv:2305.18507*, 2023c.
- Huang, S., Ma, S., Li, Y., Huang, M., Zou, W., Zhang, W., and Zheng, H.-T. Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles. *arXiv preprint arXiv:2308.10855*, 2023.
- Ismayilzada, M., Paul, D., Bosselut, A., and van der Plas, L. Creativity in ai: Progresses and challenges. *arXiv preprint arXiv:2410.17218*, 2024.
- Jain, M., Bengio, E., Hernandez-Garcia, A., Rector-Brooks, J., Dossou, B. F., Ekbote, C. A., Fu, J., Zhang, T., Kilgour, M., Zhang, D., et al. Biological sequence design with gflownets. In *International Conference on Machine Learning*, pp. 9786–9801. PMLR, 2022.
- Jain, M., Deleu, T., Hartford, J., Liu, C.-H., Hernandez-Garcia, A., and Bengio, Y. Gflownets for ai-driven scientific discovery. *Digital Discovery*, 2(3):557–577, 2023a.
- Jain, M., Raparthy, S. C., Hernández-García, A., Rector-Brooks, J., Bengio, Y., Miret, S., and Bengio, E. Multi-objective gflownets. In *International conference on machine learning*, pp. 14631–14653. PMLR, 2023b.
- Jang, H., Kim, M., and Ahn, S. Learning energy decompositions for partial inference of gflownets. *arXiv preprint arXiv:2310.03301*, 2023.
- Jiang, Y., Ilievski, F., Ma, K., and Sourati, Z. BRAIN-TEASER: Lateral thinking puzzles for large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14317–14332, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.885. URL <https://aclanthology.org/2023.emnlp-main.885>.

- Jin, C., Wu, Y., Cao, J., Xiang, J., Kuo, Y.-L., Hu, Z., Ullman, T., Torralba, A., Tenenbaum, J. B., and Shu, T. Mmtom-qa: Multimodal theory of mind question answering. *arXiv preprint arXiv:2401.08743*, 2024.
- Jung, J., Qin, L., Welleck, S., Brahman, F., Bhagavatula, C., Bras, R. L., and Choi, Y. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv preprint arXiv:2205.11822*, 2022.
- Kambhampati, S., Valmeekam, K., Guan, L., Stechly, K., Verma, M., Bhambri, S., Saldyt, L., and Murthy, A. Llm’s can’t plan, but can help planning in llm-modulo frameworks. *arXiv preprint arXiv:2402.01817*, 2024.
- Kang, H., Sachdeva, E., Gupta, P., Bae, S., and Lee, K. Gflowvln: Enhancing multi-step reasoning in vision-language models with generative flow networks. *arXiv preprint arXiv:2503.06514*, 2025.
- Kim, H., Kim, M., Choi, S., and Park, J. Genetic-guided gflownets: Advancing in practical molecular optimization benchmark. *arXiv preprint arXiv:2402.05961*, 2024a.
- Kim, M., Yun, T., Bengio, E., Zhang, D., Bengio, Y., Ahn, S., and Park, J. Local search gflownets. *arXiv preprint arXiv:2310.02710*, 2023.
- Kim, M., Choi, S., Son, J., Kim, H., Park, J., and Bengio, Y. Ant colony sampling with gflownets for combinatorial optimization. *arXiv preprint arXiv:2403.07041*, 2024b.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.
- Koziarski, M., Rekesh, A., Shevchuk, D., van der Sloot, A., Gaiński, P., Bengio, Y., Liu, C.-H., Tyers, M., and Batey, R. A. Rgfn: Synthesizable molecular generation using gflownets. *arXiv preprint arXiv:2406.08506*, 2024.
- Kraaijveld, K., Jiang, Y., Ma, K., and Ilievski, F. Columbus: Evaluating cognitive lateral understanding through multiple-choice rebuses. *arXiv preprint arXiv:2409.04053*, 2024.
- Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J. D., Singh, A., Baumli, K., Iqbal, S., Bishop, C., Roelofs, R., et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.
- Kwon, O. J., Matsunaga, D. E., and Kim, K.-E. Gdpo: Learning to directly align language models with diversity using gflownets. *arXiv preprint arXiv:2410.15096*, 2024.
- Lahlou, S., Deleu, T., Lemos, P., Zhang, D., Volokhova, A., Hernández-García, A., Ezzine, L. N., Bengio, Y., and Malkin, N. A theory of continuous generative flow networks. In *International Conference on Machine Learning*, pp. 18269–18300. PMLR, 2023.
- LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.
- Lee, H., Kim, S., Lee, S., Hwang, S., Lee, J., Lee, B.-J., and Kim, S. Arcle: The abstraction and reasoning corpus learning environment for reinforcement learning. *arXiv preprint arXiv:2407.20806*, 2024a.
- Lee, S., Kim, M., Cherif, L., Dobre, D., Lee, J., Hwang, S. J., Kawaguchi, K., Gidel, G., Bengio, Y., Malkin, N., et al. Learning diverse attacks on large language models for robust red-teaming and safety tuning. *arXiv preprint arXiv:2405.18540*, 2024b.
- Li, C., Liu, C.-H., Liu, D., Cai, Q., and Pan, L. Bifurcated generative flow networks. *arXiv preprint arXiv:2406.01901*, 2024a.
- Li, Q., Fu, L., Zhang, W., Chen, X., Yu, J., Xia, W., Zhang, W., Tang, R., and Yu, Y. Adapting large language models for education: Foundational capabilities, potentials, and challenges. *arXiv preprint arXiv:2401.08664*, 2023a.
- Li, W., Li, Y., Li, Z., Hao, J., and Pang, Y. Dag matters! gflownets enhanced explainer for graph neural networks. *arXiv preprint arXiv:2303.02448*, 2023b.
- Li, Y., Luo, S., Wang, H., and Hao, J. Cflownets: Continuous control with generative flow networks. *arXiv preprint arXiv:2303.02430*, 2023c.
- Li, Z., Chen, G., Shao, R., Jiang, D., and Nie, L. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*, 2024b.
- Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let’s verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.
- Lin, B. Y., Wu, Z., Yang, Y., Lee, D.-H., and Ren, X. RiddleSense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1504–1515, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.131. URL <https://aclanthology.org/2021.findings-acl.131>.



- Liu, D., Jain, M., Dossou, B. F., Shen, Q., Lahlou, S., Goyal, A., Malkin, N., Emezue, C. C., Zhang, D., Hassen, N., et al. Gflowout: Dropout with generative flow networks. In *International Conference on Machine Learning*, pp. 21715–21729. PMLR, 2023a.
- Liu, S., Cai, Q., He, Z., Sun, B., McAuley, J., Zheng, D., Jiang, P., and Gai, K. Generative flow network for list-wise recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1524–1534, 2023b.
- Lu, S. Z., Lu, Z., Hajiramezanali, E., Biancalani, T., Bengio, Y., Scalia, G., and Koziarski, M. Cell morphology-guided small molecule generation with gflownets. In *ICML 2024 Workshop on Structured Probabilistic Inference* {\&} *Generative Modeling*, 2024.
- Luong, T. Q., Zhang, X., Jie, Z., Sun, P., Jin, X., and Li, H. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 2024.
- Ma, G., Bengio, E., Bengio, Y., and Zhang, D. Baking symmetry into gflownets. *arXiv preprint arXiv:2406.05426*, 2024.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.
- Madan, K., Rector-Brooks, J., Korablyov, M., Bengio, E., Jain, M., Nica, A. C., Bosc, T., Bengio, Y., and Malkin, N. Learning gflownets from partial episodes for improved convergence and stability. In *International Conference on Machine Learning*, pp. 23467–23483. PMLR, 2023.
- Madan, K., Zhan, A., Lamb, A., Bengio, E., Pan, L., Berseth, G., and Bengio, Y. Goal2flownet: Learning diverse policy covers using GFlownets for goal-conditioned RL, 2024. URL <https://openreview.net/forum?id=VCscggkg2t>.
- Malkin, N., Jain, M., Bengio, E., Sun, C., and Bengio, Y. Trajectory balance: Improved credit assignment in gflownets. *Advances in Neural Information Processing Systems*, 35:5955–5967, 2022a.
- Malkin, N., Lahlou, S., Deleu, T., Ji, X., Hu, E., Everett, K., Zhang, D., and Bengio, Y. Gflownets and variational inference. *arXiv preprint arXiv:2210.00580*, 2022b.
- Meister, C., Pimentel, T., Wiher, G., and Cotterell, R. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023.
- Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., and Scialom, T. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.
- Michail, A., Clematide, S., and Opitz, J. Paraphrasus: A comprehensive benchmark for evaluating paraphrase detection models. *arXiv preprint arXiv:2409.12060*, 2024.
- Min, D., Hu, N., Jin, R., Lin, N., Chen, J., Chen, Y., Li, Y., Qi, G., Li, Y., Li, N., et al. Exploring the impact of table-to-text methods on augmenting llm-based question answering with domain hybrid data. *arXiv preprint arXiv:2402.12869*, 2024a.
- Min, D., Xu, Z., Qi, G., Huang, L., and You, C. Unihgkr: Unified instruction-aware heterogeneous knowledge retrievers. *arXiv preprint arXiv:2410.20163*, 2024b.
- Mirchandani, S., Xia, F., Florence, P., Ichter, B., Driess, D., Arenas, M. G., Rao, K., Sadigh, D., and Zeng, A. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*, 2023.
- Mishra, S., Finlayson, M., Lu, P., Tang, L., Welleck, S., Baral, C., Rajpurohit, T., Taffjord, O., Sabharwal, A., Clark, P., et al. Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*, 2022.
- Naik, R., Chandrasekaran, V., Yuksekgonul, M., Palangi, H., and Nushi, B. Diversity of thought improves reasoning abilities of large language models. *arXiv preprint arXiv:2310.07088*, 2023.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Pan, L., Zhang, D., Courville, A., Huang, L., and Bengio, Y. Generative augmented flow networks. *arXiv preprint arXiv:2210.03308*, 2022.
- Pan, L., Malkin, N., Zhang, D., and Bengio, Y. Better training of gflownets with local credit and incomplete trajectories. In *International Conference on Machine Learning*, pp. 26878–26890. PMLR, 2023a.
- Pan, L., Zhang, D., Jain, M., Huang, L., and Bengio, Y. Stochastic generative flow networks. In *Uncertainty in Artificial Intelligence*, pp. 1628–1638. PMLR, 2023b.
- Qin, L., Welleck, S., Khashabi, D., and Choi, Y. Cold decoding: Energy-based constrained text generation with langevin dynamics. *Advances in Neural Information Processing Systems*, 35:9538–9551, 2022.

- Qiu, L., Jiang, L., Lu, X., Sclar, M., Pyatkin, V., Bhagavatula, C., Wang, B., Kim, Y., Choi, Y., Dziri, N., et al. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559*, 2023.
- Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.
- Roy, J., Bacon, P.-L., Pal, C., and Bengio, E. Goal-conditioned gflownets for controllable multi-objective molecular design. *arXiv preprint arXiv:2306.04620*, 2023.
- Runco, M. A. *Divergent thinking*. Ablex Publishing Corporation Norwood, NJ, 1991.
- Runco, M. A. and Acar, S. Divergent thinking as an indicator of creative potential. *Creativity research journal*, 24 (1):66–75, 2012.
- Saparov, A. and He, H. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *arXiv preprint arXiv:2210.01240*, 2022.
- Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., and Leike, J. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Sendera, M., Kim, M., Mittal, S., Lemos, P., Scimeca, L., Rector-Brooks, J., Adam, A., Bengio, Y., and Malkin, N. On diffusion models for amortized inference: Benchmarking and improving stochastic control and sampling. *arXiv preprint arXiv:2402.05098*, 2024.
- Shen, M. W., Bengio, E., Hajiramezanali, E., Loukas, A., Cho, K., and Biancalani, T. Towards understanding and improving gflownet training. In *International Conference on Machine Learning*, pp. 30956–30975. PMLR, 2023a.
- Shen, T., Pandey, M., and Ester, M. Tacogfn: Target conditioned gflownet for structure-based drug design. *arXiv preprint arXiv:2310.03223*, 2023b.
- Shi, C., Yang, H., Cai, D., Zhang, Z., Wang, Y., Yang, Y., and Lam, W. A thorough examination of decoding methods in the era of llms. *arXiv preprint arXiv:2402.06925*, 2024.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- Shridhar, M., Yuan, X., Côté, M.-A., Bisk, Y., Trischler, A., and Hausknecht, M. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.
- Singh, A., Co-Reyes, J. D., Agarwal, R., Anand, A., Patil, P., Liu, P. J., Harrison, J., Lee, J., Xu, K., Parisi, A., et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.
- Song, Z., Yang, C., Wang, C., An, B., and Li, S. Latent logic tree extraction for event sequence explanation from llms. *arXiv preprint arXiv:2406.01124*, 2024.
- Summers-Stay, D., Voss, C. R., and Lukin, S. M. Brainstorm, then select: a generative language model improves its creativity score. In *The AAAI-23 Workshop on Creative AI Across Modalities*, 2023. URL <https://openreview.net/forum?id=8HwKaJlwvl>.
- Takase, R., Tsunokake, M., Tsuchiya, Y., and Inuzuka, S. Gflownet fine-tuning for diverse correct solutions in mathematical reasoning tasks. *arXiv preprint arXiv:2410.20147*, 2024.
- Tan, J. C. M. and Motani, M. Large language model (llm) as a system of multiple expert agents: An approach to solve the abstraction and reasoning corpus (arc) challenge. *arXiv preprint arXiv:2310.05146*, 2023.
- Tian, Y., Peng, B., Song, L., Jin, L., Yu, D., Mi, H., and Yu, D. Toward self-improvement of llms via imagination, searching, and criticizing. *arXiv preprint arXiv:2404.12253*, 2024.
- Todd, G., Merino, T., Earle, S., and Togelius, J. Missed connections: Lateral thinking puzzles for large language models. *arXiv preprint arXiv:2404.11730*, 2024.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.
- Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., and Higgins, I. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.
- Valmeekam, K., Marquez, M., Sreedharan, S., and Kambhampati, S. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36, 2024.

- Vemgal, N., Lau, E., and Precup, D. An empirical study of the effectiveness of using a replay buffer on mode discovery in gflowtnets. *arXiv preprint arXiv:2307.07674*, 2023.
- Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.
- Waks, S. Lateral thinking and technology education. *Journal of Science Education and Technology*, 6:245–255, 1997a. URL <https://api.semanticscholar.org/CorpusID:108865906>.
- Waks, S. Lateral thinking and technology education. *Journal of Science Education and Technology*, 6(4):245–255, 1997b. doi: 10.1023/A:1022534310151. URL <https://doi.org/10.1023/A:1022534310151>.
- Wang, P., Li, L., Shao, Z., Xu, R., Dai, D., Li, Y., Chen, D., Wu, Y., and Sui, Z. Math-shepherd: A label-free step-by-step verifier for llms in mathematical reasoning. *arXiv preprint arXiv:2312.08935*, 2023a.
- Wang, R., Zelikman, E., Poesia, G., Pu, Y., Haber, N., and Goodman, N. D. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*, 2023b.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Wu, Y., Sun, Z., Li, S., Welleck, S., and Yang, Y. An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*, 2024.
- Xie, J., Zhang, K., Chen, J., Zhu, T., Lou, R., Tian, Y., Xiao, Y., and Su, Y. Travelplanner: A benchmark for real-world planning with language agents. *arXiv preprint arXiv:2402.01622*, 2024a.
- Xie, Y., Kawaguchi, K., Zhao, Y., Zhao, J. X., Kan, M.-Y., He, J., and Xie, M. Self-evaluation guided beam search for reasoning. *Advances in Neural Information Processing Systems*, 36, 2024b.
- Xu, Y., Khalil, E. B., and Sanner, S. Graphs, constraints, and search for the abstraction and reasoning corpus. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4115–4122, 2023a.
- Xu, Y., Li, W., Vaezipoor, P., Sanner, S., and Khalil, E. B. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *arXiv preprint arXiv:2305.18354*, 2023b.
- Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: synergizing reasoning and acting in language models (2022). *arXiv preprint arXiv:2210.03629*, 2023.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- Yu, F., Gao, A., and Wang, B. Outcome-supervised verifiers for planning in mathematical reasoning. *arXiv preprint arXiv:2311.09724*, 2023a.
- Yu, F., Jiang, L., Huang, S., Wu, Z., and Dai, X. Persuasivetom: A benchmark for evaluating machine theory of mind in persuasive dialogues. *arXiv preprint arXiv:2502.21017*, 2025.
- Yu, J., He, R., and Ying, R. Thought propagation: An analogical approach to complex reasoning with large language models. *arXiv preprint arXiv:2310.03965*, 2023b.
- Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023c.
- Yuan, L., Cui, G., Wang, H., Ding, N., Wang, X., Deng, J., Shan, B., Chen, H., Xie, R., Lin, Y., et al. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*, 2024a.
- Yuan, W., Pang, R. Y., Cho, K., Sukhbaatar, S., Xu, J., and Weston, J. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024b.
- Yuan, Z., Yuan, H., Li, C., Dong, G., Tan, C., and Zhou, C. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.
- Yue, X., Qu, X., Zhang, G., Fu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

- Zelikman, E., Wu, Y., Mu, J., and Goodman, N. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.
- Zhang, D., Malkin, N., Liu, Z., Volokhova, A., Courville, A., and Bengio, Y. Generative flow networks for discrete probabilistic modeling. In *International Conference on Machine Learning*, pp. 26412–26428. PMLR, 2022.
- Zhang, D., Dai, H., Malkin, N., Courville, A., Bengio, Y., and Pan, L. Let the flows tell: Solving graph combinatorial optimization problems with gflownets. *arXiv preprint arXiv:2305.17010*, 2023a.
- Zhang, D., Pan, L., Chen, R. T., Courville, A., and Bengio, Y. Distributional gflownets with quantile flows. *arXiv preprint arXiv:2302.05793*, 2023b.
- Zhang, D. W., Rainone, C., Peschl, M., and Bondesan, R. Robust scheduling with gflownets. *arXiv preprint arXiv:2302.05446*, 2023c.
- Zhang, Z., Yao, Y., Zhang, A., Tang, X., Ma, X., He, Z., Wang, Y., Gerstein, M., Wang, R., Liu, G., et al. Igniting language intelligence: The hitchhiker’s guide from chain-of-thought reasoning to language agents. *arXiv preprint arXiv:2311.11797*, 2023d.
- Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023e.
- Zhang, Z., Zheng, C., Wu, Y., Zhang, B., Lin, R., Yu, B., Liu, D., Zhou, J., and Lin, J. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025.
- Zhao, Z., Lee, W. S., and Hsu, D. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36, 2024.
- Zhong, S., Huang, Z., Gao, S., Wen, W., Lin, L., Zitnik, M., and Zhou, P. Let’s think outside the box: Exploring leap-of-thought in large language models with creative humor generation, 2024. URL <https://arxiv.org/abs/2312.02439>.
- Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.
- Zhu, D., Li, Y., Shao, Y., Hao, J., Wu, F., Kuang, K., Xiao, J., and Wu, C. Generalized universal domain adaptation with generative flow networks. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8304–8315, 2023.
- Zhu, X., Wang, J., Zhang, L., Zhang, Y., Huang, Y., Gan, R., Zhang, J., and Yang, Y. Solving math word problems via cooperative reasoning induced language models. *arXiv preprint arXiv:2210.16257*, 2022.
- Zimmermann, H., Lindsten, F., van de Meent, J.-W., and Naesseth, C. A. A variational perspective on generative flow networks. *arXiv preprint arXiv:2210.07992*, 2022.

## A. Additional Related Work

**Reasoning with LLM.** Recent LLMs (Achiam et al., 2023; Touvron et al., 2023; Bai et al., 2022; Chowdhery et al., 2023) have demonstrated great potentials in tackling complex reasoning tasks (Cobbe et al., 2021; Mishra et al., 2022; Hendrycks et al., 2021; Rein et al., 2023; Mialon et al., 2023). (1) **Fine-tuning LLMs** is a primary way to enhance their reasoning abilities, including SFT and reward-maximization reinforcement learning (RL) approaches. **SFT** with large-scale and high-quality datasets of reasoning chains has proven very effective (Yu et al., 2023c; Yue et al., 2023; Yuan et al., 2024a). Various methods for constructing training samples have been proposed when ground truth reasoning chains are not available. For example, STaR (Zelikman et al., 2022) uses online sampling with self-correction to find positive samples. ReST<sup>EM</sup> (Singh et al., 2023) and V-STaR (Hosseini et al., 2024) filter samples with external verifiers. **RL** techniques, particularly reward-maximizing policy optimization methods like PPO, are widely employed in LLMs (Ouyang et al., 2022; Bai et al., 2022; Havrilla et al., 2024; Luong et al., 2024). However, both maximum likelihood training (i.e. SFT) and reward-maximizing policy optimization (e.g., PPO) do not encourage models to generate diverse solutions. (2) **prompting-based reasoning algorithms** aim to better elicit the knowledge inside LLMs without tuning their parameters. Techniques such as CoT (Wei et al., 2022) and its variants (Chen et al., 2022; Li et al., 2024b; Zhang et al., 2023e; Zhou et al., 2022; Kojima et al., 2022) have improved LLM performance by enabling them to generate intermediate steps before arriving at a final answer. To provide reasoning more guidance, self-evaluation (Xie et al., 2024b; Shinn et al., 2024; Madaan et al., 2024) and reward models are introduced to enhance reasoning process (Uesato et al., 2022; Lightman et al., 2023). Besides, a more relevant series of works combine LLM reasoning capabilities with planning and search algorithms such as MCTS (Hao et al., 2023; Feng et al., 2023; Zhao et al., 2024), tree and graph search (Jung et al., 2022; Zhu et al., 2022; Yao et al., 2024; Besta et al., 2024; Yao et al., 2023). Moreover, recent studies turn to amortizing computation for reasoning paths (Yuan et al., 2023; Wu et al., 2024; Bansal et al., 2024), such as self-correct (Kumar et al., 2024; Saunders et al., 2022), self-improvement (Tian et al., 2024; Yuan et al., 2024b).



**GFlowNets.** GFlowNets (Bengio et al., 2021) were originally proposed to learn policies for sampling from unnormalized distributions, with a primary motivation from scientific discovery (Jain et al., 2023a), which requires generating diverse high-reward samples (Shen et al., 2023b; Roy et al., 2023; Zhang et al., 2023b; Ma et al., 2024; Pan et al., 2023b), such as molecular generation (Koziarski et al., 2024; Kim et al., 2024a; Lu et al., 2024) and biological sequence generation (Ghari et al., 2023; Jain et al., 2022). Beyond the science domain, GFlowNets have also been applied in various downstream applications such as recommendation systems (Liu et al., 2023b), domain adaptation (Zhu et al., 2023), combinatorial optimization (Zhang et al., 2023a; Kim et al., 2024b) and explainability of deep neural networks (Li et al., 2023b). Additionally, GFlowNets have proven to be suitable for sampling from posterior distributions (Hu et al., 2023b; Deleu et al., 2022; 2024; Zhang et al., 2022). As a reinforcement learning method, prior works have incorporated intermediate feedback with GFlowNets to address sparse reward issues (Pan et al., 2023a; Jang et al., 2023; Pan et al., 2022) and multi-objective rewards (Jain et al., 2023b; Hernandez-Garcia et al., 2023; Chen & Mauch, 2023). There are also theoretical analyses treating GFlowNets as recurrent MCMC (Deleu & Bengio, 2023) and variational inference (Malkin et al., 2022b; Zimmermann et al., 2022) that are used to model the distribution over trajectories.

**Diverse Thinking.** Vertical and lateral thinking (Waks, 1997a; Ismayilzada et al., 2024) are two distinct approaches that differ significantly in their focus and methodology. Vertical thinking emphasizes logical, structured, and sequential reasoning, often following a step-by-step approach to solve problems. Our work aligns with this paradigm to generate multiple correct, structured reasoning processes to achieve specific goals. In contrast, lateral thinking prioritizes creativity and innovation, encouraging the exploration of unconventional perspectives and challenging established assumptions. There are multiple works that evaluate (Huang et al., 2023; Chen et al., 2024a; Kraaijveld et al., 2024; Todd et al., 2024) and promote (Zhong et al., 2024; Summers-Stay et al., 2023; Jiang et al., 2023; Lin et al., 2021) the lateral thinking ability of LLMs.

## B. Preliminaries and Background

GFlowNets (Bengio et al., 2021; 2023; Liu et al., 2023a) are a class of models that amortize the cost of sampling from an intractable target distribution over terminal states  $\mathcal{X}$  by learning a neural network-facilitated approximation of the target distribution using its unnormalized density or reward function. The task of sampling from this distribution resorts to a decision-making process. Below, we introduce GFlowNets with more details.

**Settings.** We are given a pointed directed acyclic graph

(DAG)  $\mathcal{G} = (\mathcal{S}, \mathcal{A})$ , where  $\mathcal{S}$  is a finite set of vertices (states), and  $\mathcal{A} \subseteq \mathcal{S} \times \mathcal{S}$  is a set of directed edges (actions). If  $s \rightarrow s'$  is an action, we say  $s$  is a parent of  $s'$  and  $s'$  is a child of  $s$ . There is exactly one state that has no incoming edge, called the initial state  $s_0 \in \mathcal{S}$ . States that have no outgoing edges are called *terminal states*. We denote by  $\mathcal{X}$  the set of terminal states. A complete trajectory is a sequence  $\tau = (s_0 \rightarrow \dots \rightarrow s_n)$  such that each  $s_i \rightarrow s_{i+1}$  is an action and  $s_n = x \in \mathcal{X}$ . We denote by  $\mathcal{T}$  the set of complete trajectories and the terminal state as  $\tau_x$ .

Here we define the reward  $R : \mathcal{X} \rightarrow \mathbb{R}^+$ , and define a forward transition probability function, or a forward policy,  $P_F(\cdot|s)$ , which is a distribution over the children of every state  $s \in \mathcal{S}$ . The forward policy is typically parametrized by a neural network that takes a representation of  $s$  as input and produces the logits of a distribution over its children. Any forward policy  $P_F$  induces a distribution over complete trajectories  $\tau \in \mathcal{T}$  (denoted by  $P_F$  as well), which in turn defines a marginal distribution over terminal states  $x \in \mathcal{X}$ :

$$P_F(\tau) = P_F(s_0 \rightarrow \dots \rightarrow s_n) = \prod_{i=0}^{n-1} P_F(s_{i+1}|s_i) \quad \forall \tau \in \mathcal{T} \quad (8)$$

Given a forward policy  $P_F$ , terminal states  $x \in \mathcal{X}$  can be sampled from  $P_F$  by sampling trajectories  $\tau$  from  $P_F(\tau)$  and taking their final states  $s_n$ . GFlowNets aim to find a forward policy  $P_F$  such that the induced distribution  $P_F^\top(x)$  is proportional to the reward function:

$$P_F^\top(x) \propto R(x) \quad (9)$$

**Training.** Training GFlowNets considers achieving a consistent flow (Bengio et al., 2023; Malkin et al., 2022a), which means the flow for the forward direction should equal to the flow for the backward direction. Below we introduce relevant objectives.

**Detailed Balance (DB).** The DB objective (Bengio et al., 2023) requires learning two objectives in addition to parametric forward policy  $P_F(\cdot|s)$ : 1. A *Backward policy*, which is distribution  $P_B(s'|s; \theta)$  over the parents of any non-initial state. 2. A *State flow function*:  $F(\cdot; \theta) : \mathcal{S} \rightarrow \mathbb{R}_{>0}$ . Then DB loss for a single transition  $s \rightarrow s'$  is defined as:

$$\mathcal{L}_{DB} = \left( \log \frac{F(s; \theta) P_F(s'|s; \theta)}{F(s'; \theta) P_B(s|s'; \theta)} \right)^2 \quad (10)$$

if  $\mathcal{L}_{DB}$  is optimized to 0 for each transition, then the forward policy  $P_F$  satisfies 9.

**Trajectory Balance (TB).** Trajectory balance (Malkin et al., 2022a) introduces a backward policy  $P_B$ , which is a learned distribution  $P_B(\cdot|s')$  over the parents of every state  $s \in \mathcal{S}$ . We have the following backward probability:

$$P_B(\tau) = P_B(s_0 \rightarrow \dots \rightarrow s_n | s_n) = \prod_{i=0}^{n-1} P_B(s_i | s_{i+1}) \quad \forall \tau \in \mathcal{T} \quad (11)$$

Then we give the trajectory balance constraint for any complete trajectory  $\tau = (s_0 \rightarrow \dots \rightarrow s_n)$ :

$$F(s_0) \prod_{i=1}^n P_F(s_i | s_{i-1}) = F(s_n) \prod_{i=1}^n P_B(s_{i-1} | s_i) \quad (12)$$

where we have  $P(s_n) = F(s_n)/Z$ . To enforce this constraint, we convert equation 12 into an objective for optimization. Suppose that a model with parameters  $\theta$  outputs estimated forward policy  $P_F(\cdot | s; \theta)$  and backward policy  $P_B(\cdot | s; \theta)$  for states  $s$ , as well as a scalar  $Z_\theta$  estimating  $F(s_0)$ . GFlowNets aim to approximate a Markovian Flow  $F_\theta$  that  $F(s_n) = R(s_n)$ . For a complete trajectory  $\tau = (s_0 \rightarrow \dots \rightarrow s_n)$ , define the trajectory balance loss as follows:

$$\mathcal{L}_{TB}(\tau; \theta) = \left( \log \frac{Z_\theta \prod_{t=0}^{n-1} P_F(s_{t+1} | s_t; \theta)}{R(s_n) \prod_{t=0}^{n-1} P_B(s_t | s_{t+1}; \theta)} \right)^2 \quad (13)$$

If  $\mathcal{L}_{TB}$  is made equal to 0 for every complete trajectory  $\tau$ , then equation 9 satisfies for all  $x \in X$  and  $Z$  is the inverse constant of proportionality:  $Z = \sum_{x \in \mathcal{X}} R(x)$ .

**Conditional GFlowNets.** In a GFlowNet, both the policy and reward function can be conditioned on additional information. For instance, in the tasks we focus on, a GFlowNet policy generates actions sequentially for an embodied reasoning problem, starting from an initial state  $s_0$  and a goal  $g$ . Furthermore, the allowable actions vary depending on the specific  $s_0$  in each case. The conditional GFlowNets we train achieve amortization by sharing the policy model across different  $s_0$  and  $g$ , enabling the model to generalize to initial states and targets that were not seen during training.

## C. Experimental details

### C.1. Diversity Metric

We define the following metric to measure the diversity of reasoning paths found by different approaches. Under the same number of samplings at inference time, we count the number of different successful trajectories a policy finds for the successful example on average.

$$\text{Diversity} = \frac{\sum_{i=1}^n S_i \cdot \mathbb{I}(S_i \geq 1)_i}{\sum_{i=1}^n \mathbb{I}(S_i \geq 1)_i} \geq 1 \quad (14)$$

where  $n$  is the total number of problems,  $S_i$  is the number of successful trajectories found for the  $i$ -th question, and

Table 9: Training time shown is seconds when training on the BlocksWorld.

Method	Runtime (s)
SFT	196.37
SFT+PPO	1740.96
FOR	6833.37

$\mathbb{I}(S_i \geq 1)$  is an indicator function that is 1 if there is at least one successful trajectory found for the  $i$ -th question and 0 otherwise. Thus, the denominator is the number of examples in which a model finds at least one trajectory, and the nominator is the sum of all successful trajectories a model finds across all examples. The smallest diversity is 1 when a method can only find at most one successful trajectory on average, and diversity = 1.5 indicates a method is able to find 1.5 different successful trajectories on average.

### C.2. Creativity Metric

We define the following metric to quantify the creativity of a reasoning method. Given the same number of samples during inference, we calculate the ratio of unique successful trajectories that a method identifies in the test dataset  $D_{test}$ , which are not found by any other methods. Let  $\mathcal{M} = \{m_1, m_2, \dots, m_{|\mathcal{M}|}\}$  represent the set of reasoning methods. For the  $i$ -th problem, the  $l$ -th method has a solution set  $S_i^l$ , where  $1 \leq l \leq |\mathcal{M}|$ . The complete set of solutions across all methods is defined as:

$$S = \bigcup_{i=1}^n \bigcup_{l=1}^{|\mathcal{M}|} S_i^l \quad (15)$$

Then we can define the creativity metric of method  $m_l$  as:

$$\text{Creativity}(m_l) = \frac{1}{|S|} \sum_{i=1}^{|D_{test}|} \sum_{s \in S_i^l} \mathbb{I}(s, i, l), \quad (16)$$

where for the  $i$ -th problem, if the solution  $s \in S_i^l$  is found only by method  $m_l$  and not by any other method  $m_k$  (where  $k \neq l$ ), then  $\mathbb{I}(s, i, l) = 1$ . Otherwise,  $\mathbb{I}(s, i, l) = 0$ . The indicator function  $\mathbb{I}(s, i, l)$  is defined as:

$$\mathbb{I}(s, i, l) = \begin{cases} 1, & \text{if } s \notin \bigcup_{k \neq l} S_i^k \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

### C.3. Efficiency Analysis

All experiments were conducted using a server with a single NVIDIA A100 GPU. Below we report the average of 3 times training for 6-step training cost on BlocksWorld dataset for 10 epochs. We compare with SFT, PPO and table 9 shows the results.

PPO and FoR need much more training costs because they need exploration and interaction with environments to collect trajectories for training, and SFT only trains on ground-truth trajectories which take less time.

#### C.4. BlocksWorld.

**FoR Setup.** During the training, we finetune the LLM with LoRA (Hu et al., 2021) with  $r = 32$ ,  $\alpha = 64$ , and dropout=0.1. We set  $\epsilon$  from 0.3 and decrease it to 0.01,  $\beta$  from 1 to 2, and the probability  $\delta$  using replay buffer increases from 0.3 to 0.5 throughout the iterations linearly. The learning rate is set to  $1e-4$  with a cosine annealing schedule, and the number of training iterations is set to 10. Reward weight  $\lambda$  is set to 1.5. In our ablation study when setting  $\lambda = 0$ , we add a small number  $b = 0.5$  to avoid log 0. Table 7 shows the template we use for the forward policy in the 6-step setting, and its difference between 2-step and 4-step is only replacing the 6-step demonstration to 2-step and 4-step. During testing, we sample 8, 20, and 40 trajectories for 2, 4, and 6 steps respectively. As long as one trajectory reaches the goal, we label this instance as solved, all the baselines conform to the same rule.

**Additional details for baselines.** We compare FoR the following baselines:

(1) *Chain-of-Thoughts prompting (CoT)* (Wei et al., 2022): It concatenates  $k$  problems with ground truth solutions and the test problem, and prompts the LLM to generate a solution. We test the setting where  $k = 1, 5, 15$ , and pass the test cases to LLMs at the same times as FoR, and the test case is regarded as solved if at least one plan is correct.

(2) *Tree-of-Thoughts prompting (ToT)* (Yao et al., 2024): This approach constructs a tree of actions and searches for the solution with the highest reward. For each action, the reward includes (a) the likelihood of the LLM predicting the action and (b) self-evaluation, where the LLM is prompted with the question, "Is this action good?" and the answer is mapped to a reward value. We implement ToT with both breadth-first search (BFS) and depth-first search (DFS), terminating after generating 10 solutions.

(3) *Reasoning-via-Planning (RAP)* (Hao et al., 2023): This method also conducts a tree search for the optimal solution. Different from ToT, it alternatively predicts the next action and predicts the resulting block arrangement. Besides the rewards used in ToT, if the predicted block arrangement matches the goal, a high reward will be assigned.

(4) *Supervised Fine-Tuning (SFT)*: We use problems in the training set and their corresponding ground truth solutions to finetune the LLM. Note that this is an easier setting than FoR which does not have access to ground truth solutions. We train LLM with the same iterations as FoR.

(5) *Proximal Policy Optimization (PPO)* (Schulman et al., 2017): This is a widely-used reward-maximization reinforcement learning method for LLM training. We design the objective to encourage the LLM to generate solutions that satisfy the goal. Following the common practice of previous work (Ouyang et al., 2022; Wang et al., 2023a), we penalize the policy if it deviates too much from the reference policy. Formally, the objective is  $\max_{\pi_{\theta}} \mathbb{E}_{\tau \sim \pi_{\theta}} [R(x, y)] - \beta \mathbb{D}_{\text{KL}} [\pi_{\theta}(y | x) \| \pi_{\text{ref}}(y | x)]$ .

(6) *GfN-CoT* (Hu et al., 2023a): This approach adapts the GFlowNets training paradigm, which is a diversity-seeking RL method, to enable posterior sampling of the intermediate reasoning process from LLMs.

**Performance on OOD settings.** We further assess performance on out-of-distribution (OOD) settings. Specifically, we train the model using FoR and SFT on a 2-step training set and evaluate it on a 4-step test set, and then train the model on the 4-step training set and evaluate it on the 6-step test set. This allows us to analyze their generalization on OOD problems. For prompting-based baselines, we use 2-step and 4-step examples as demonstrations, respectively.

According to the result in table 10, FoR maintains the highest accuracy (71.43%) on OOD tasks compared to other methods like CoT and SFT, which range from 9.52% to 14.28%. FoR also achieves greater diversity (by an absolute improvement of 0.2 over SFT), highlighting its superior generalization and solution exploration capabilities.

**Additional results on search-based approaches.** To assess whether search-based methods can generate diverse solutions, we execute ToT (DFS) multiple times (equal to the number of runs performed for our method) and compare their performance. The results shown in Table 13 demonstrate that though search-based methods improve the diversity, they still underperform FoR.

**Additional baseline results with diversity-encouraging instruction.** To further stimulate the diverse reasoning ability in the baseline approaches, we add a diversity-encouraging prompt as instruction:

*Please carefully understand the goals and initial states, then come up with diverse solutions and think outside the box.*

We evaluate multiple baseline methods using Llama-3-8B as the base model, following the exact same settings as in Section 4.2. The results for BlocksWorld are reported in Table 11. The numbers in parentheses indicate the performance difference compared to the original prompt without the diversity-encouraging instruction.

We observe that diversity-encouraging prompts for the CoT and SFT baselines lead to improvements in both diversity and accuracy, with average absolute gains of 0.03 and 5.11%,

Table 10: OOD results on BlocksWorld.

Method	2-step to 4-step			4-step to 6-step		
	Acc. (%)	Diversity	Creativity (%)	Acc. (%)	Diversity	Creativity (%)
CoT (1-shot)	9.52	1.00	3.12	2.02	1.00	0.00
CoT (5-shot)	14.28	1.00	3.12	12.12	1.08	3.45
CoT (15-shot)	11.90	1.00	3.12	8.08	1.00	0.00
ToT (BFS)	9.52	-	-	8.08	-	-
ToT (DFS)	4.76	-	-	6.06	-	-
RAP	<b>80.95</b>	-	-	34.34	-	-
SFT ( $\alpha=1.0$ )	11.92	1.00	9.37	28.28	1.03	1.15
FOR (Ours)	71.43	<b>1.20</b>	<b>59.38</b>	<b>65.65</b>	<b>1.25</b>	<b>60.92</b>

Table 11: Baseline results with diversity-encouraging instruction prompt. "+" denotes the performance improvement compared to the absence of the prompt.

Method	4-step			6-step		
	Acc. (%)	Diversity	Creativity (%)	Acc. (%)	Diversity	Creativity (%)
CoT (1-shot)	16.67 (-10.90)	1.00 (-0.05)	0.0 (0.00)	11.11 (-4.71)	1.09 (+0.04)	0.0 (0.00)
CoT (5-shot)	59.52 (+16.66)	1.12 (+0.08)	2.04 (+2.04)	33.33 (+3.70)	1.03 (+0.00)	0.79 (+0.79)
CoT (15-shot)	52.38 (+12.32)	1.09 (+0.06)	0.0 (0.00)	13.13 (-6.40)	1.07 (+0.04)	0.0 (0.00)
SFT ( $\alpha=1.0$ )	59.52 (+17.46)	1.10 (+0.05)	0.0 (0.00)	47.47 (+12.79)	1.10 (+0.06)	0.0 (0.00)
FOR (Ours)	<b>98.41</b>	<b>1.27</b>	<b>12.24</b>	<b>78.44</b>	<b>1.33</b>	<b>9.52</b>

respectively. However, FOR still outperforms them, achieving average absolute improvements of 0.19 in diversity, 9.46% in creativity, and 34.93% in accuracy compared to the best baseline for each metric.

**Additional case study.** In Table 12, we show examples generated by FOR. We observe that after training, FOR can sample the terminal state with probability approximately proportional to the rewards, leading to an approximate sampling of different plans with the same probability. This empirically verifies the efficacy of the training objective.

### C.5. Game of 24.

**FOR Setup.** See Figure 8 for the prompt template used in the experiment of the Game of 24. We use LoRA to train the model with  $r = 8$ ,  $\alpha = 32$ , dropout=0.1. We load the LLM in fp16, and set the hyperparameters as follows: batch size = 4, learning rate = 1e-5, number of epochs = 5, and the reward weight  $w = 100$ .

### C.6. Rubik’s Cube

**FOR Setup.** The training hyperparameters are identical to BlocksWorld. During testing, we sample 10 trajectories. See Figure 9 for the prompt template of the Rubik’s Cube task.

**Additional details for baselines.** Apart from the baselines in Blocksworld, we further compare them with GoT and XoT.

*Graph-of-Thought (GoT)* (Besta et al., 2024): GoT builds upon the ToT method by introducing the ability to create graph-like thought structures, achieved through the aggrega-

tion and refinement of thoughts during intermediate search stages. While this approach allows for more adaptable thought structures, it still requires several LLM inference calls for evaluation, leading to substantial computational expenses.

*Everything-of-Thought (XoT)* (Ding et al., 2023): XOT is a collaborative framework combining LLMs with MCTS to optimize the thought generation process, aiding LLMs in solving complex problems. It first trains a small network to explore the space fast while LLMs refine and correct the thoughts generated by MCTS.

### C.7. 1D-ARC

**FOR Setup.** Except that we train the model for 1 iteration, other training hyperparameters are identical to BlocksWorld. We use the hand-crafted transformation functions in ARC Challenge 2nd-place (de Miquel, 2021) on Kaggle 2020. See Figure 11 for the prompt template of the 1D-ARC task. Part of the prompt is adapted from (Tan & Motani, 2023). For CoT and FoR, we sampled 20 times. IO methods directly predict the output grids without an explicit reasoning process, while program-only and Hypothesis Search approaches generate a large number of candidate programs and choose the best candidates, which is time-consuming. As a result, we do not report diversity and creativity metrics for these methods.

**Additional details for baselines.** In addition to IO and CoT, we also compare our approach with Hypothesis Search which belongs to discrete program search methods (Barke et al., 2024; Xu et al., 2023a; Lee et al., 2024a).



Table 12: An example of the probability of two trajectories to be sampled by FoR.




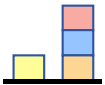
Initial State	goal	Trajectory	Terminal State	$P_F(\tau)$
	the red block is on top of the blue block	1. unstack the orange block from on top of the blue block 2. put down the orange block 3. unstack the blue block from on top of the red block 4. put down the blue block 5. pick up the red block 6. stack the red block on top of the blue block		0.47
	the red block is on top of the blue block	1. unstack the orange block from on top of the blue block 2. put down the orange block 3. unstack the blue block from on top of the red block 4. stack the blue block on top of the orange block 5. pick up the red block 6. stack the red block on top of the blue block		0.46

Table 13: Performance of Search-based methods with multiple runs. ♣ denotes results from a single run.

Method	2-step	4-step		6-step	
	Acc. (%)	Acc. (%)	Diversity	Acc. (%)	Diversity
ToT (DFS)♣	13.33	16.67	-	8.08	-
ToT (DFS)	40.00	42.85	1.0	31.31	1.06
FoR	100.00	98.41	1.27	78.44	1.33

**Hypothesis Search** (Wang et al., 2023b): The method first generates multiple hypotheses describing the underlying transformation rules in natural language, and then selects a subset of potentially correct hypotheses. Based on these selected hypotheses, numerous Python programs are synthesized, which are subsequently tested on the training input-output pairs to verify whether they pass all the cases. If a program successfully passes all the training input-output pairs, it is considered to have accurately captured the underlying transformation rules.

**OOD data creation.** We separate the in-distribution and OOD data by topics and ontology. We use the animal-related problems as in-distribution examples and the number-related problems as OOD examples.

### C.8. Logical Reasoning

**Setup.** We use LoRA to train the model with  $r = 8$ ,  $\alpha = 32$ , dropout=0.1. We load the LLM in fp16, and set the hyperparameters as follows: batch size = 4, learning rate =  $5e-6$ , number of epochs = 40, and the reward weight  $w = 100$ . See Table 10 for the prompt template of the logical reasoning task.

**Additional details for Baselines.** Apart from CoT, ToT, and

RAP, we compare FoR with STaR (Zelikman et al., 2022), which uses online sampling to filter our positive examples consistent with ground truth trajectories to finetune the LLM. Note that this is an easier setting than FoR, which doesn’t have access to ground truth solutions. It also indicates an upper bound of SFT methods that do not rely on ground truth solutions, like. All baselines use Llama3 8B as the base model.

### C.9. Mathematical Reasoning

**Automatic Diversity Metric.** To distinguish different solutions, we adapt a paraphrase-based method (Michail et al., 2024) to automatically evaluate the diversity with GPT-4o to measure the number of different solutions for each problem. The prompt is shown in Figure 13 **Additional results.** To further investigate the relationship between accuracy and the number of samples, we set the sampling counts to 1, 5, 10, and 20, and compared the performance under these configurations. As shown in Figure 4, FoR consistently outperforms SFT by a significant margin, ranging from 3.5% to 6.6%. The higher accuracy demonstrates that aligning with the reward distribution of solutions enables FoR to better explore the reasoning space, increasing the possibility of reaching the correct solution.

## D. Additional Analysis Results

**Performance with various LLMs.** We conduct experiments to evaluate FoR’s scalability and diversity gains using larger Llama-3-70B model and different family models, Qwen2.5 (7B & 72B) (Yang et al., 2024) and Intern2.5-

Table 14: Performance with larger models and different families of models.

Method	4-step		6-step	
	Acc. (%)	Diversity	Acc. (%)	Diversity
CoT 5-shot (Llama3-8B)	28.57	1.05	15.82	1.05
CoT 5-shot (Llama3-70B)	45.23	1.05	46.46	1.11
FoR (Llama3-8B)	98.41	1.27	78.44	1.33
FoR (Llama3-70B)	100.00	1.38	87.65	1.40
FoR (Qwen2.5-7B)	100.00	1.24	86.86	1.36
FoR (Qwen2.5-72B)	100.00	1.41	90.13	1.46
FoR (InternLM2.5-7B-Chat)	100.00	1.26	83.83	1.31

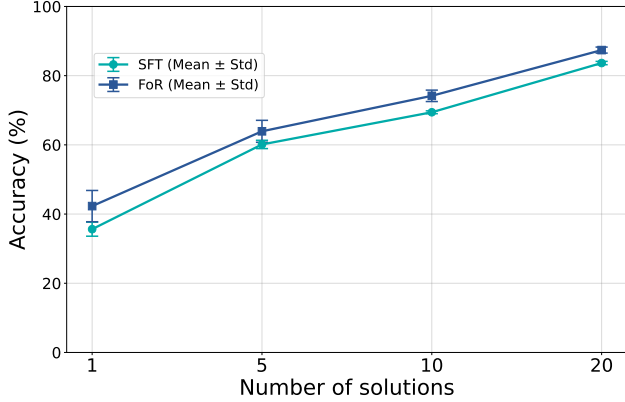


Figure 4: Accuracy of FOR and SFT across the different number of sampled solutions.

7B-Chat (Cai et al., 2024) on the BlocksWorld task. Results shown in Table 14 indicate that FOR consistently improves accuracy and diversity with larger models. Compared to CoT baselines, FOR exhibits greater diversity gains as model size increases. This suggests that the gains of FOR become more pronounced as model capacity increases. Even with minimal data (15 examples), FOR yields clear improvements, highlighting its potential for robust performance across different base models.

## E. Exploration and Training

FoR employs the following techniques to explore during the training phase:

1. Online training: (1) we employ the online policy  $P_F(a_t|s_{t-1}, \alpha)$ , and its tempered version (2) Similar to  $\epsilon$ -greedy, we sample action at step  $t$  by  $P_F$  with probability  $\epsilon$ , and sample with uniform distribution over action space  $P_U(a_t|s_{t-1})$  with  $(1 - \epsilon)$  probability. (3) To further explore the high-reward region, we modified the local search (Kim et al., 2023; Zhang et al., 2022). More specifically, we select the trajectory with the highest reward in a batch and conduct a destroy and reconstruction process for augmenting the trajectories to enable a higher probability of sampling successful trajectories, referring to Appendix F for more details.

2. Offline training: (1) Experience replay represents a significant advancement in reinforcement learning, offering enhancements in both learning efficiency and stability, as evidenced by recent empirical studies in GFlowNets (Vemgal et al., 2023; Shen et al., 2023a). To optimize the utility of the trajectories collected, we set up a prioritized replay buffer (PRB). This buffer facilitates the sampling of trajectories in proportion to their reward value,  $R(\tau)$ , or its logarithmic value, thereby prioritizing potentially more informative experiences. (2) For tasks (e.g. Game of 24) that have a large space, online sampling diverse trajectories with LLMs is computationally expensive. Therefore, we integrate the offline trajectories to have a larger coverage of space and improve the efficiency, which means  $\delta = 0$ .

Algorithm 1 describes the training framework.

## F. Modified Local Search

Local search is a simple data augmentation technique for GFlowNets (Kim et al., 2023; Zhang et al., 2022; Sendera et al., 2024), which is designed to enhance training efficiency. Different from the original local search which is conducted on each sampled trajectory, we select the trajectory in a batch with the highest reward to conduct a local search. Here we denote the trajectory reward  $R(\tau)$  as the reward of the terminal state of the trajectory  $R(\tau = (s_0 \rightarrow \dots \rightarrow s_n)) = R(s_n)$ . More specifically, we illustrate our modified local search for one instance as follows:

- **Sampling:** Sample a set of complete trajectories  $\{\tau_1, \dots, \tau_M\}$  using forward policy  $P_F$  and select the  $\tau_m$  with the largest reward  $R(\tau_m)$
- **Searching:** We destroy  $\tau_m$  by backtracking  $K$ -step into a partial trajectory and reconstruct the complete trajectory from the partial trajectory:

$$\begin{aligned} \tau_{destroy} &= (s_0 \rightarrow \dots \rightarrow s'_{n-K}), \\ \tau_{recon} &= (s'_{n-K} \rightarrow \dots \rightarrow s'_n) \end{aligned} \quad (18)$$

**Algorithm 1** FoR Training

---

```

1: Input:  $I$ : number of iterations,  $P_F$ : initial LLM policy,  $\mathcal{D}$ : Prioritized Replay Buffer,  $M$ : Batch-size,  $\delta$ :
   online-offline ratio,  $\mathcal{E}$ : Training Dataset,  $\mathcal{O}$ : offline Data
2: Output: Trained policy  $P_F$ 
3: for  $i = 1$  to  $I$  do
4:   Sample from training dataset  $\mathcal{E}$  with initial state  $s_0$  and goal  $g$ 
5:   Sample  $u \sim [0, 1]$ 
6:   if  $u < \delta$  then
7:     // Exploration
8:     Sample  $M$  online trajectories  $\{\tau_1, \dots, \tau_M\}$  with forward policy  $P_F$ 
9:     Select trajectory  $\tau_m \in \{\tau_1, \dots, \tau_M\}$  with the largest  $R(\tau_m)$ 
10:     $\{\tau'_1, \dots, \tau'_{N'}\} \leftarrow \text{Local Search}(\tau_m)$ 
11:    Update  $\mathcal{D} \leftarrow \mathcal{D} \cup \{\tau_1, \dots, \tau_M\} \cup \{\tau'_1, \dots, \tau'_{N'}\}$ 
12:   else
13:     // Exploitation
14:     if is Game24 then
15:       Sample  $M$  offline trajectories from Offline Data  $\mathcal{O}$ 
16:     else
17:       Sample  $M$  offline trajectories from  $\mathcal{D}$ 
18:   Exploit  $M$  (with  $N'$ ) trajectories to compute the objective function in Eq. 7
19:   Update the parameter in  $P_F$  with respect to Eq. 7
20: return  $P_F$ 

```

---

We obtain the local searched trajectory  $\tau'$ :

$$\tau' = (s_0 \rightarrow \dots \rightarrow s'_{n-K} \rightarrow \dots \rightarrow s'_n) \quad (19)$$

Where the  $\tau_{recon}$  is completed by the random policy  $P_U$  which randomly selects a feasible action for efficiency. We can obtain a set of reconstructed trajectories  $\{\tau'_1, \dots, \tau'_{N'}\}$

- **Filtering:** We now need to evaluate the collected reconstructed trajectories  $\{\tau'_1, \dots, \tau'_{N'}\}$  and determine whether to accept or reject  $\tau' \in \{\tau'_1, \dots, \tau'_{N'}\}$ . Specifically, we accept  $\tau'$  as follows:

$$A(\tau, \tau') = 1_{R(\tau') > R(\tau)} \quad (20)$$

This means we greedily filter out the candidates  $\{\tau'_1, \dots, \tau'_{N'}\} \subset \{\tau'_1, \dots, \tau'_N\}$  that have a higher reward than  $\tau_m$ , which has a higher possibility of reaching the goal. Then we return these trajectories and add them into the replay buffer  $\mathcal{D}$ .

## G. Case Study

**Balance between diversity and accuracy.** According to Figure 5, we use the problem (3,4,6,11) to show how FoR achieves such high performance while focusing on diversity. As illustrated in the figure, we compare trajectories sampled 20 times by both SFT and FoR. While both methods produce diverse trajectories initially, FoR demonstrates better

capability in reaching successful final steps from various middle steps. For example, FoR successfully transitions from intermediate steps (3,6,15) to target 24, whereas SFT fails to do so. This highlights the effectiveness of FoR's design in simultaneously promoting diversity and ensuring accuracy.

**Better robustness due to exploratory nature.** According to Figure 6, we use the problem (7,9,9,13) to demonstrate the robustness of FoR. As shown in the figure, SFT repeatedly fails by getting stuck in a single second state of (9,3,16) 20 times, while FoR successfully discovers multiple diverse trajectories leading to the correct solution. This robustness can be attributed to the exploratory nature of FoR's training objective, which encourages the model to sample diverse successful trajectories. By expanding the search space through high-reward exploration, FoR increases the chance of finding successful outcomes. This capability not only improves the robustness of the model but also enhances its generalization to new scenarios, showcasing the effectiveness of FoR in addressing complex reasoning tasks.

**Other examples** Figure 14 shows generated samples for the BlocksWorld, Figure 15 for Game24, and Table 15 for ProOntoQA, respectively.

## H. Prompts.

Below, we detail all the prompts we used in all settings from Figure 7, 8, 9, 10, 11, 12.

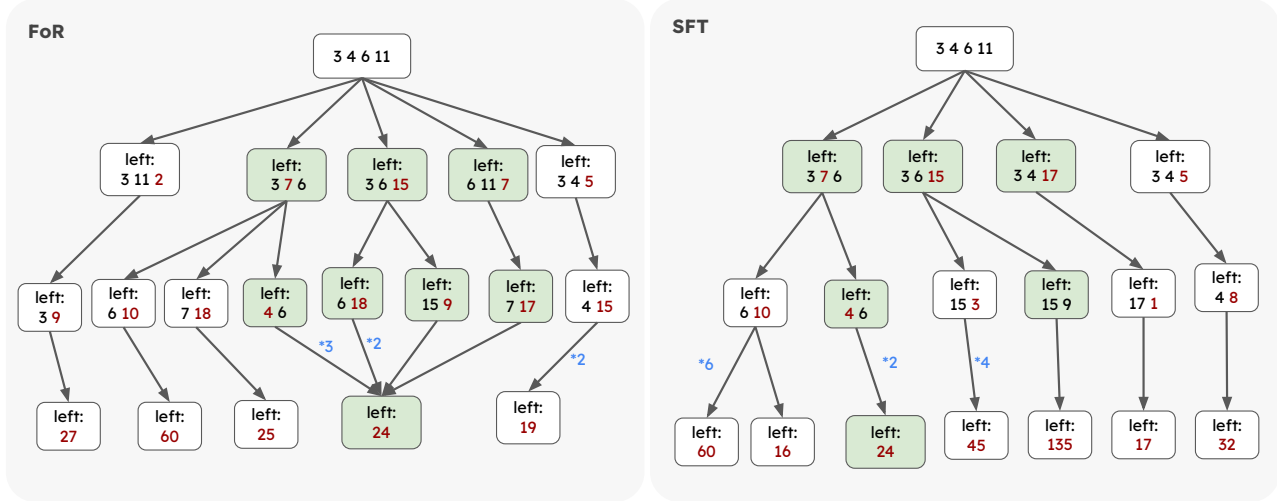


Figure 5: Problem (3,4,6,11). Green blocks represent the states that can achieve 24. Blue numbers represent the sample times of trajectories bigger than 1. This shows that FoR can achieve 24 in multiple different states while SFT usually fails to do so.

Table 15: Examples for PrOntoQA.

Query: True or false: 31 is not imaginary. (OOD)	
State	Action
31 is a natural number.	Natural numbers are integers.
31 is an integer.	Integers are real numbers.
31 is a real number.	Real numbers are not imaginary.
31 is not imaginary.	Finish.
Query: True or false: Wren is not bony. (In-distribution)	
State	Action
Wren is a painted lady.	Each painted lady is a butterfly.
Wren is a butterfly.	Each butterfly is a lepidopteran.
Wren is a lepidopteran.	Each lepidopteran is an insect.
Wren is an insect.	Each insect is an arthropod.
Wren is an arthropod.	Each arthropod is not bony.
Wren is not bony.	Finish.

Future work can extend FoR to broader reasoning tasks, including reasoning on structured data (Min et al., 2024a;b), social reasoning (Jin et al., 2024; Yu et al., 2025), and multimodal reasoning (Zhang et al., 2023e; Kang et al., 2025).

## I. Limitations and Future Work

Future work should further address the two limitations of FoR.

The first is **acquisition of a large number of trajectories efficiently**. Online sampling with LLMs is computationally expensive, leading to more efficient and effective strategies to explore more complex settings, such as real-world settings, AlfWorld (Shridhar et al., 2020), and TravelPlanner (Xie et al., 2024a) to be further studied.

The second is **facilitating FoR long-range steps reasoning**. LLMs fall short in long-range planning and reasoning, thus methods like MCTS (Feng et al., 2023) or an automatic reasoning system (Trinh et al., 2024) can be combined with LLMs for long-horizon divergent reasoning.



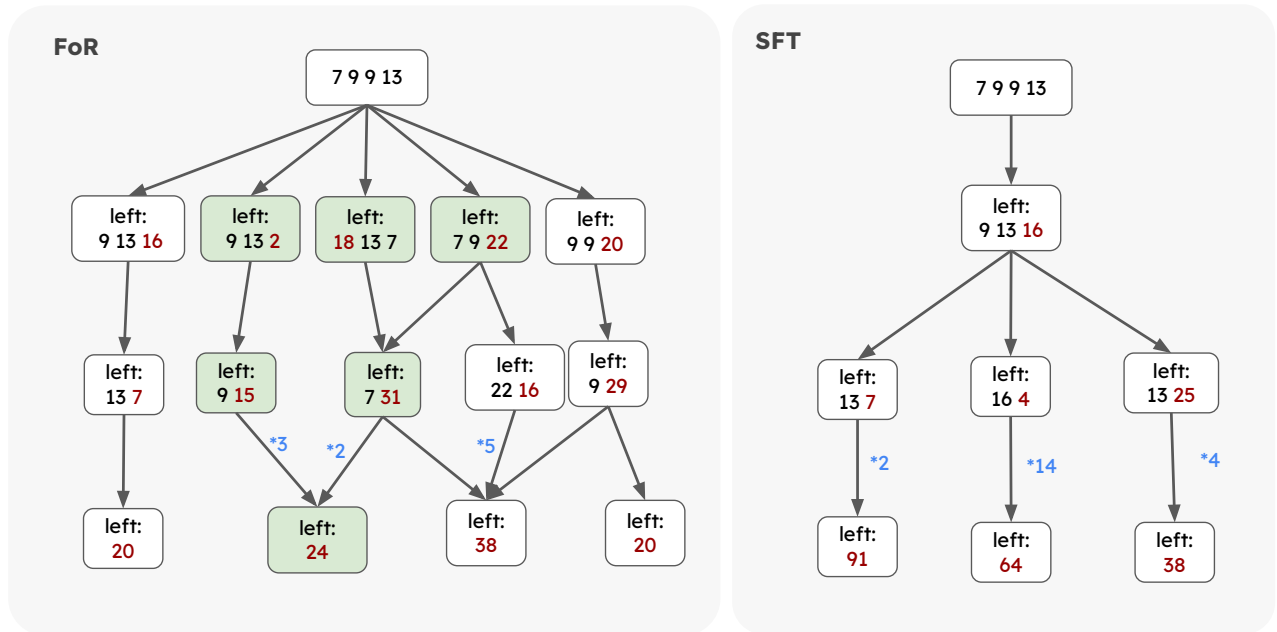


Figure 6: Problem (7,9,9,13). Green blocks represent the states that can achieve 24. Blue numbers represent the sample times of trajectories bigger than 1. This shows that FoR can achieve 24 in multiple different states while SFT usually fails to do so.

Prompt for BlocksWorld

I am playing with a set of blocks where I need to arrange the blocks into stacks.  
Here are the actions I can do

Pick up a block  
Unstack a block from on top of another block  
Put down a block  
Stack a block on top of another block

I have the following restrictions on my actions:  
I can only pick up or unstack one block at a time.  
I can only pick up or unstack a block if my hand is empty.  
I can only pick up a block if the block is on the table and the block is clear.  
A block is clear if the block has no other blocks on top of it and if the block is not picked up.  
I can only unstack a block from on top of another block if the block  
I am unstacking was really on top of the other block.  
I can only unstack a block from on top of another block if the block I am unstacking is clear.  
Once I pick up or unstack a block, I am holding the block.  
I can only put down a block that I am holding.  
I can only stack a block on top of another block if I am holding the block being stacked.  
I can only stack a block on top of another block if the block onto which I am stacking the block is clear.  
Once I put down or stack a block, my hand becomes empty.

[STATEMENT]

As initial conditions I have that, the orange block is clear, the hand is empty, the red block is on top of the blue block, the orange block is on top of the red block and the blue block is on the table.

My goal is to have that the blue block is on top of the orange block.

My plan is as follows:

[PLAN]

unstack the orange block from on top of the red block  
put down the orange block  
unstack the red block from on top of the blue block  
put down the red block  
pick up the blue block  
stack the blue block on top of the orange block  
[PLAN END]

[STATEMENT]

As initial conditions I have that, <current state>

My goal is to My goal is to have that <goals>

My plan is as follows:

[PLAN]

<action>

Figure 7: Prompt template for the embodied reasoning task (6-step).

Prompt for Game of 24

Use numbers and basic arithmetic operations (+ - \* /) to obtain 24.  
 For each step, you are only allowed to choose two of the remaining numbers to obtain a new number.  
 Input: 4 4 6 8  
 Steps:  
 $4 + 8 = 12$  (left: 4 6 12)  
 $6 - 4 = 2$  (left: 2 12)  
 $2 * 12 = 24$  (left: 24)  
 Input: 2 9 10 12  
 Steps:  
 $12 * 2 = 24$  (left: 9 10 24)  
 $10 - 9 = 1$  (left: 1 24)  
 $24 * 1 = 24$  (left: 24)  
 Input: 4 9 10 13  
 Steps:  
 $13 - 10 = 3$  (left: 3 4 9)  
 $9 - 3 = 6$  (left: 4 6)  
 $4 * 6 = 24$  (left: 24)  
 Input: 1 4 8 8  
 Steps:  
 $8 / 4 = 2$  (left: 1 2 8)  
 $1 + 2 = 3$  (left: 3 8)  
 $3 * 8 = 24$  (left: 24)  
 Input: 5 5 5 9  
 Steps:  
 $5 + 5 = 10$  (left: 5 9 10)  
 $10 + 5 = 15$  (left: 9 15)  
 $15 + 9 = 24$  (left: 24)  
 Input: <input>  
 Steps:  
 <action>

Figure 8: Prompt template for the mathematical puzzle task.

Prompt for Rubik's Cube

You are a virtual expert in solving a 2x2 Pocket Cube. Your task is to restore a scrambled 2x2 Rubik's Cube to its original state. All the given problems can be solved in 1 to 4 moves. You cannot exceed more than 11 moves. Provide the sequence of moves required for the restoration. Please follow the instructions and rules below to complete the solving:

1. A 2x2 Pocket Cube has six faces, namely: [Upper, Front, Bottom, Left, Right, Back] Each consisting of a 2x2 grid of squares, with each square having its own color.
2. Colors in the Cube are represented in numbers: [0, 1, 2, 3, 4, 5]
3. You must make a move to the Cube to achieve a Restored State. Note that we just need each face to have the same numbers, no matter which face has which color.
4. A restoration of a Pocket Cube is to move squares in each face to have the same numbers.
5. You are only allowed to use the following moves [U, U', U2, R, R', R2, F, F', F2].

Now strictly follow the above process to form Restoration Moves.

[STATEMENT]  
 As initial state of the cube, I have that  
 [Initial Cube State]:  
 <current state>  
 [Process]:  
 [Step 1]  
 [Move] <action>

Figure 9: Prompt template for the spatial Reasoning task.

Prompt for PrOntoQA

Given a list of facts, and a current claim, output one possible fact as the next step ONLY BASED ON THE LAST CLAIM without using your knowledge. Be sure to copy the EXACT sentence in the facts. Do NOT change any wording. Do NOT create your own words. Give me the next step ONLY.

Facts 1: Each lepidopteran is an insect. Each arthropod is a protostome. Every animal is multicellular. Protostomes are invertebrates. Each whale is bony. Each painted lady is a butterfly. Invertebrates are animals. Butterflies are lepidopterans. Each insect is six-legged. Every insect is an arthropod. Arthropods are not bony.

Query 1: True or false: Sally is not bony.

Claim 1.1: Sally is an insect.

Next 1.1: Every insect is an arthropod.

Claim 1.2: Sally is an arthropod.

Next 1.2: Arthropods are not bony.

Claim 1.3: Sally is not bony.

Next 1.3: Finish.

Facts 2: Lepidopterans are insects. Every animal is multicellular. Each insect is an arthropod. Each invertebrate is an animal. Insects are six-legged. Arthropods are small. Arthropods are invertebrates. Each butterfly is a lepidopteran. Whales are not small.

Query 2: True or false: Polly is not small.

Claim 2.1: Polly is a lepidopteran.

Next 2.1: Lepidopterans are insects.

Claim 2.2: Polly is an insect.

Next 2.2: Each insect is an arthropod.

Claim 2.3: Polly is an arthropod.

Next 2.3: Arthropods are small.

Claim 2.4: Polly is small.

Next 2.4: Finish.

Facts 3: <facts>

Query 3: <query>

Claim 3.1: <initial state>

Next 3.1: <action>

Figure 10: Prompt template for logical reasoning task.

Prompt for 1D-ARC

You are provided with a series of input-output pairs, where each value from 'a' to 'j' represents a different color, and '.' denotes a blank cell. For example, [['.', 'a', '.'], ['.', '.', 'b']] represents a grid with 2 rows and 3 columns, where color 'a' is at position (1,0) and color 'b' is at position (2,1).

Coordinates are expressed in 2D positions (row, col), with 'row' indicating the row number and 'col' indicating the column number, both using zero-based indexing. The input-output pairs may not cover all possibilities, so you should infer the simplest possible relationship between them.

Your task is to reason through a sequence of Python functions that can transform the input grid into the output grid. Please strictly follow this process to form the appropriate Python function.

[STATEMENT]

You have the following input-output pairs:

[Initial Grid State]:

<init\_state>

Based on the provided list of Python functions, select the appropriate function to achieve the transformation from the input to the output:

<python\_function>

Now, please choose one function from the above list:

<action>

Figure 11: Prompt template for abstraction reasoning task.



Prompt for GSM8K

Given a question, please decompose it into sub-questions. For each sub-question, please answer it in a complete sentence, ending with "The answer is". When the original question is answerable, please start the subquestion with "Now we can answer the question: ".

Question 1: Four years ago, Kody was only half as old as Mohamed. If Mohamed is currently twice as 30 years old, how old is Kody?

Question 1.1: How old is Mohamed?

Answer 1.1: He is currently  $30 * 2 = 60$  years old. The answer is 60.

Question 1.2: How old was Mohamed four years ago?

Answer 1.2: Four years ago, he must have been  $60 - 4 = 56$  years old. The answer is 56.

Question 1.3: How old was Kody four years ago?

Answer 1.3: Kody was half as old as Mohamed four years ago. Thus, Kody was  $56 / 2 = 28$  years old. The answer is 28.

Question 1.4: Now we can answer the question: How old is Kody?

Answer 1.4: She is currently  $28 + 4 = 32$  years old. The answer is 32.

Question 2: On a moonless night, three fireflies danced in the evening breeze. They were joined by four less than a dozen more fireflies before two of the fireflies flew away. How many fireflies remained?

Question 2.1: How many fireflies joined?

Answer 2.1: The fireflies were joined by four less than a dozen more fireflies, which are  $12 - 4 = 8$  fireflies. The answer is 8.

Question 2.2: Now we can answer the question: How many fireflies remained?

Answer 2.2: Three fireflies were dancing originally. They were joined by 8 fireflies before two of them flew away. So there were  $3 + 8 - 2 = 9$  remaining. The answer is 9.

Question 3: <question>

Figure 12: Prompt template for the mathematical Reasoning task.

Prompt for counting different solutions

Analyze the given problem-solving approaches and identify the number of distinct methods used to arrive at the solution. Consider variations in reasoning, calculation steps, and logical structuring. Directly return the total count of unique solution approaches without additional explanations.

Solution 1: "Question 3: Claire makes a 3 egg omelet every morning for breakfast. How many dozens of eggs will she eat in 4 weeks?

Question 3.1: How many eggs does she eat per day?

Answer 3.1: She makes a 3 egg omelet every morning for breakfast. So she eats 3 eggs per day. The answer is 3.

Question 3.2: How many days are in a week?

Answer 3.2: There are 7 days in a week. The answer is 7."

Solution 2: "Question 3: Claire makes a 3 egg omelet every morning for breakfast. How many dozens of eggs will she eat in 4 weeks?

Question 3.1: How many eggs does she eat in a day?

Answer 3.1: She eats 3 eggs in a day. The answer is 3.

Question 3.2: How many days are there in 4 weeks?

Answer 3.2: There are  $4 * 7 = 28$  days in 4 weeks. The answer is 28.

Question 3.3: How many eggs does she eat in 4 weeks?

Answer 3.3: She eats 3 eggs in a day. So she eats  $3 * 28 = 84$  eggs in 4 weeks. The answer is 84.

Question 3.4: Now we can answer the question: How many dozens of eggs will she eat in 4 weeks?

Answer 3.4: She eats 84 eggs in 4 weeks. 84 eggs is  $84 / 12 = 7$  dozens. The answer is 7."

Number: 2

Solution 1: Three of the 16 eggs go to Janet breakfast. Four of the remaining 13 eggs go to her muffins. That leaves 9 eggs to sell. The eggs sell for 2 each, so she makes  $9 * 2 = 18$ . The answer is 18.

Solution 2: Janet lays 16 eggs a day. She eats three eggs for breakfast and bakes muffins with four this, so she has  $16 - 3 - 4 = 9$  eggs left. She sells these eggs at 2 each, so she makes  $9 * 2 = 18$  at the market. The answer is 18.

Number: 1

Solution 1: <solution1>

Solution 2: <solution2>

Number:

Figure 13: Prompt template measures the number of different solutions.

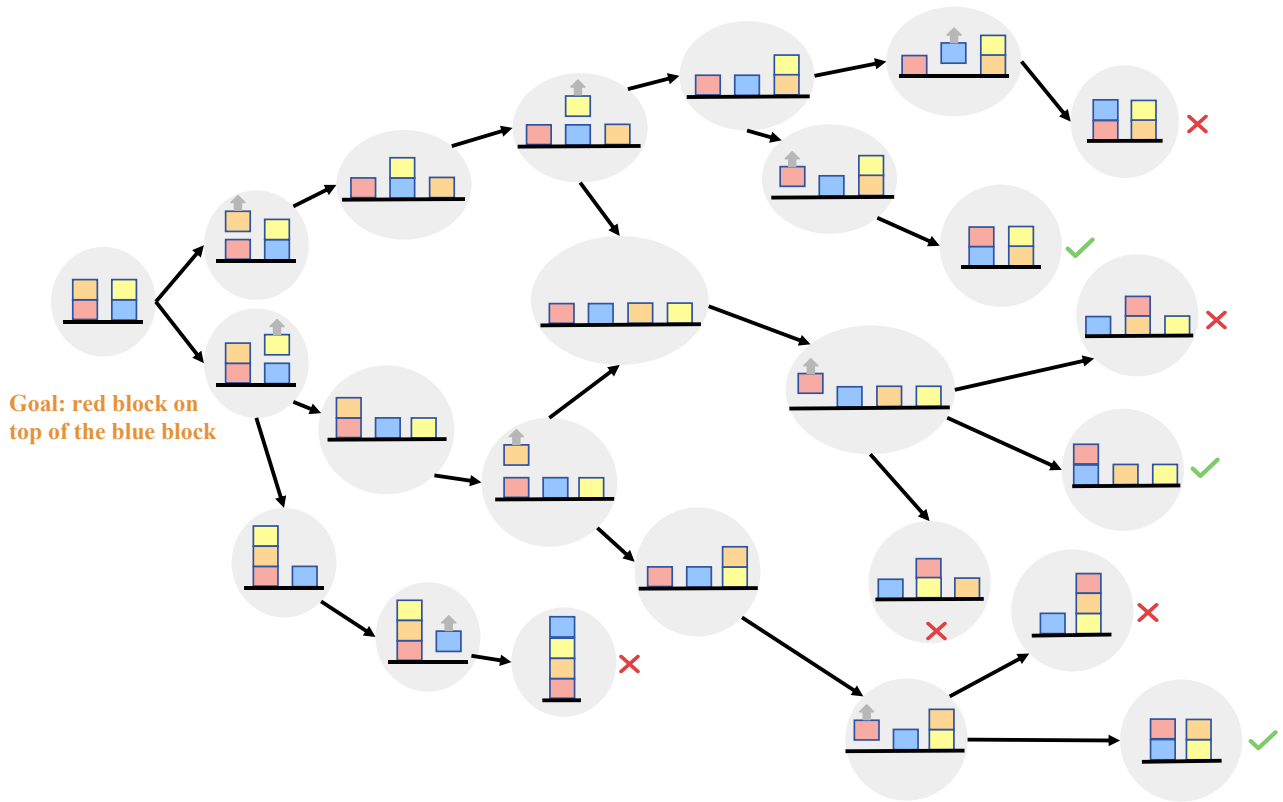


Figure 14: Example of BlocksWorld for 6-step planning.

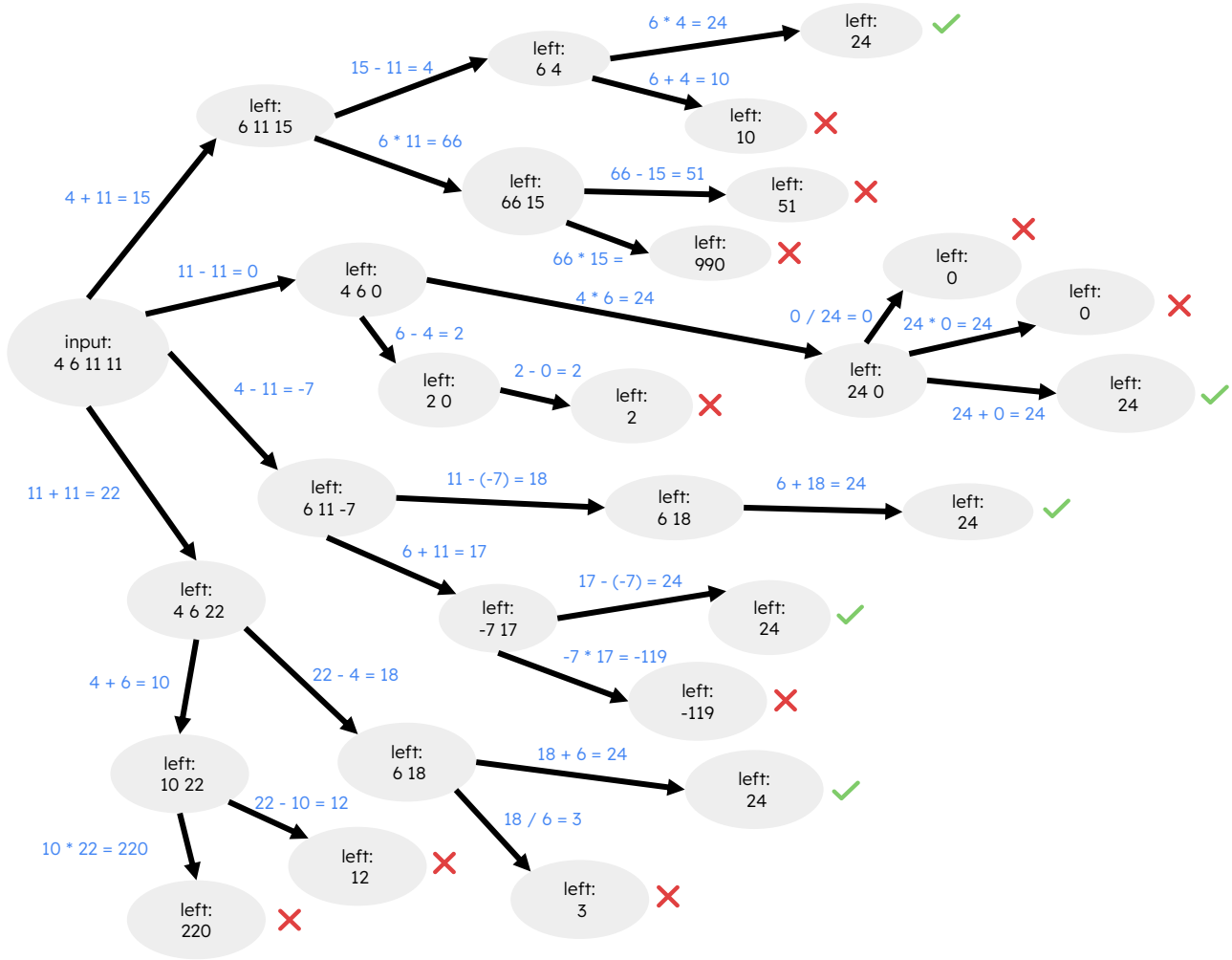


Figure 15: Example of game of 24.

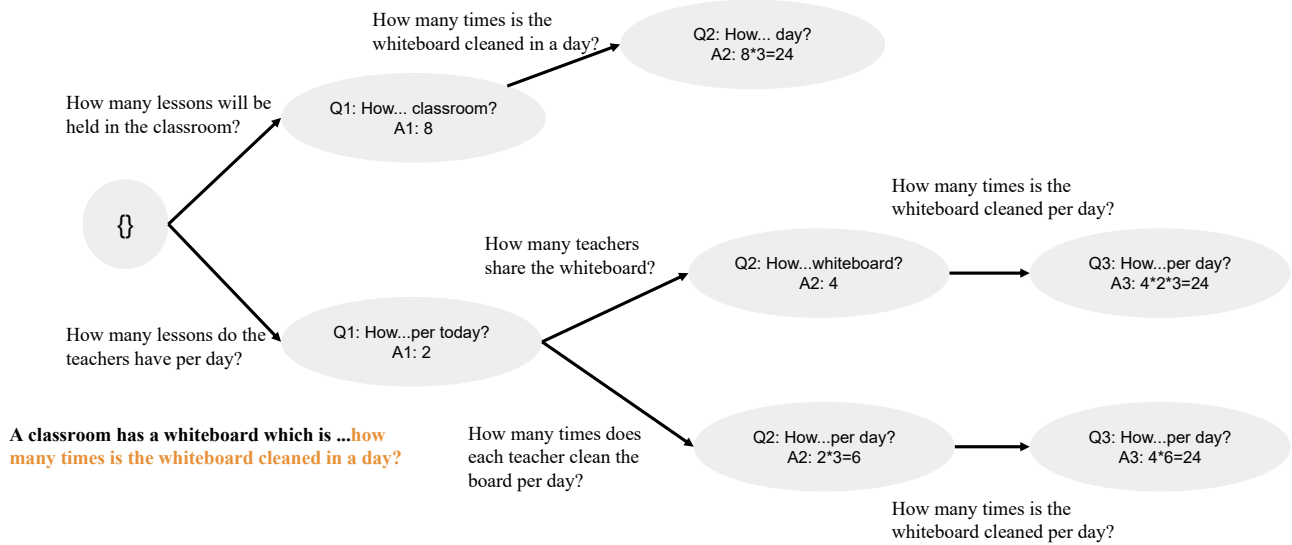


Figure 16: Example of GSM8K.