DyPO: Dynamic Policy Optimization for Multi-Turn Interactive Reasoning

Xiao Feng[1] **Bo Han**[1†] **Zhanke Zhou**[1] **Jiaqi Fan**[2]
Jiangchao Yao[3] **Ka Ho Li**[2] **Dahai Yu**[2] **Michael Kwok-Po Ng**[4]
[1]TMLR Group, Hong Kong Baptist University [2]TCL Corporate Research (HK) Co., Ltd
[3]CMIC, Shanghai Jiao Tong University [4]Department of Mathematics, Hong Kong Baptist University
`{xiaofeng, michael-ng}@hkbu.edu.hk` `{cszkzhou, bhanml}@comp.hkbu.edu.hk`
`{garyfan, karlli, dahai.yu}@tcl.com` `Sunarker@sjtu.edu.cn`

Abstract

Existing on-policy reinforcement learning methods, such as Group Relative Policy Optimization (GRPO) and its variants, have enhanced the reasoning capabilities of large language models (LLMs). However, these methods often rely on static, pre-trained knowledge to navigate partially observed contexts, limiting their effectiveness in dynamic and evolving environments. In such settings, LLMs must actively interact with the environment to gather critical information, necessitating further advancements in adaptive reasoning strategies. To mitigate this gap, we introduce **Dy**namic **P**olicy **O**ptimization (DyPO), which extends GRPO for multi-turn optimization in dynamic environments. In principle, DyPO guarantees the shifting of reasoning pattern from static to dynamic multi-turn reasoning and stablize the training process involving environmental information. DyPO incorporates four key innovations: (1) distinct thinking and action tokens that integrate real-time environmental feedback during rollouts, (2) removal of divergence regularization for dynamic reasoning transition, (3) masked intermediate observations with simplified advantage estimation for enhanced stability, and (4) auxiliary resampling with rejection sampling to mitigate over-generation noise. These enhancements enable DyPO to achieve adaptive alignment with multi-turn interactive reasoning. Evaluations on challenging simulated benchmarks, ALFWorld and WebShop, using two instantiations of DyPO with Qwen-2.5-3B-Instruct consistently demonstrate substantial improvements in both interactive decision-making and reasoning capabilities compared to existing approaches.

1 Introduction

Reasoning, the process of deriving conclusions from existing knowledge [1], is a cornerstone of Large Language Models (LLMs). These models demonstrate exceptional performance in domains such as mathematics [21] and coding [9], with reasoning capabilities often elicited through simple prompting techniques [36, 17, 35]. Integrating search methods like Tree-of-Thought [41] and Monte Carlo Tree Search (MCTS) [46] with post-training techniques [44, 22] further enhances LLMs' reasoning. This synergy enables step-by-step reasoning for tackling complex problems [42, 24].

Recent studies have shown that on-policy reinforcement learning (RL) methods can significantly enhance the reasoning capabilities of well-pretrained LLMs. Techniques such as Proximal Policy Optimization (PPO) [27] and Group Relative Policy Optimization (GRPO) [28] have proven particularly effective. These methods optimize policy models by sampling outputs and generating reward signals

†Correspondence to Bo Han (bhanml@comp.hkbu.edu.hk).

ICML 2025 Workshop on Programmatic Representations for Agent Learning.

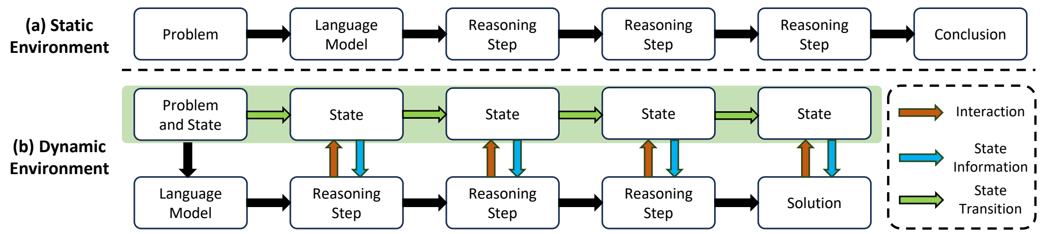

Figure 1: Comparison of reasoning in static environments and dynamic environments. In static environments, LLMs are initially prompted with full context and leverage the information of the context and their internal knowledge to derive the solution. In contrast, dynamic environments initially provide partial information about the context, and LLMs need to take actions, interacting with the environment to seek more information to derive the final solution.

through critical evaluation of the samples. GRPO, in particular, improves efficiency and stability by approximating advantage calculations via group sampling, eliminating the need for a separate critic model and demonstrating excellent scalability with increasing training data. Models trained with GRPO, such as DeepSeek-R1 [10], have achieved state-of-the-art performance on diverse reasoning tasks, earning widespread recognition in the research community.

While GRPO-based methods effectively train LLMs to reason using internal knowledge from pre-training, many real-world tasks require interaction with external environments to gather essential information. In dynamic settings, models must adaptively respond to evolving contexts, as illustrated in Fig. 1. For instance, in simulated environments like Minecraft [6], agents sequentially interact with initially unobservable objects (e.g., locating and chopping trees) to achieve complex objectives (e.g., building a house). Similarly, web agents [40] performing tasks like online shopping navigate sequential web pages, where future states are not immediately visible. These scenarios demand robust reasoning coupled with strategic, interactive decision-making, significantly increasing optimization complexity. Although GRPO has shown success in static domains like mathematics and programming, its efficacy in dynamic, partially observable environments remains largely unexplored. This gap raises a critical research question: *How can we design an RL-based algorithm that effectively elicits a model's multi-turn reasoning capabilities in dynamic environments?*

In this work, we address a key limitation of GRPO in multi-turn dynamic environments, where incomplete initial state information hinders task completion. To overcome this, we propose Dynamic Policy Optimization (DyPO), an extension of GRPO designed to optimize multi-turn interactive reasoning in dynamic settings. DyPO enhances reasoning transferability from static to dynamic environments and stabilizes the reasoning process amid external environmental feedback and multi-turn interactions. Specifically, DyPO introduces four key innovations: (1) involving specialized thinking and action tokens during rollout sampling, enabling multi-turn interactive reasoning with environmental feedback, (2) elimination of divergence regularization to facilitate the transition to dynamic reasoning, (3) masking of environmental observations and removal of reward deviation in advantage estimation for optimization stability, and (4) auxiliary resampling and rejection sampling strategies to mitigate noise from truncated overthinking. These modifications enable DyPO to achieve superior alignment with multi-turn interactive reasoning processes, resulting in enhanced reasoning capabilities across complex dynamic environments.

We introduce two instantiations, DyPO-Zero and DyPO, distinguished by whether the base model is fine-tuned with step-wise instruction data. We evaluate both through extensive experiments on two challenging dynamic benchmarks, ALFWorld [31] and WebShop [40], which simulate complex environments testing reasoning and interaction capabilities. Empirical results show that DyPO-Zero achieves an 18.41% improvement in success rate, while DyPO achieves a 3.43% improvement, both outperforming GRPO in generalizability and maintaining consistent performance as interactive turns increase. Additionally, DyPO demonstrates greater efficiency, with reduced degradation in utility as interactive turns decrease. Our findings confirm that DyPO significantly enhances LLM performance in dynamic settings and exhibits superior adaptivity compared to baselines. Case studies further reveal that DyPO-trained LLMs exhibit more consistent and stable reasoning processes.

In summary, our main contributions are three-fold as follows:

- We identify the limitation of GRPO that it overlooks the interaction ability of LLMs, which is crucial in multi-turn dynamic environments, while step-wise optimization struggles to elicit the reasoning capabilities of LLMs in dynamic environments. (Sec. 3)
- We propose DyPO, a method designed for multi-turn interactive reasoning that enhances adaptivity in dynamic environments and stabilizes optimization by bootstrapping the interactive rollout sampling process, addressing GRPO's limitations (Sec. 4).
- We implement DyPO-Zero and DyPO, two instantiations of our method, and evaluate them in two dynamic environments to demonstrate their effectiveness, generalizability, and efficiency in enhancing the multi-turn reasoning capabilities of LLMs (Sec. 5).

2 Related Work

In this section, we systematically review (1) benchmarks and methods for LLM reasoning in dynamic environments (Sec. 2.1) and (2) RL methods for enhancing LLM reasoning (Sec. 2.2).

2.1 LLM Reasoning in Dynamic Environments

LLMs perform exceptionally well in static environments, but their trustworthiness in dynamic settings, where they encounter unseen or irrelevant noise, remains a significant challenge [13, 12]. To address this, recent studies have developed diverse simulated environments to evaluate LLM reasoning in dynamic contexts, typically classified into two categories: (1) physical, embodied AI agents (e.g., robots) equipped with sensors and actuators [48, 5], and (2) game-based virtual environments like Minecraft [6] and StarCraft [32, 26]. Recent studies highlight LLMs' advanced reasoning capabilities in complex scenarios, indicating their potential for real-world dynamic applications [19, 45]. To assess LLM reasoning, numerous benchmarks have emerged, spanning simulated real-world settings [34, 31, 48], operating systems [37], web interfaces [40, 4], and strategic board games [14, 7, 11]. These environments pose unique challenges beyond traditional mathematical and coding tasks [3]: (1) partial observability, requiring active information-seeking due to incomplete state information; (2) large state and action spaces, demanding comprehension of states and available actions; and (3) constrained decision-making budgets, necessitating a balance between exploration and exploitation within limited turns. Together, these features underscore the intricate complexity of dynamic environments and highlight the critical need for advanced reasoning capabilities and adaptive planning.

2.2 RL for LLM Reasoning

Reinforcement learning (RL) [16] optimizes policy behavior to maximize predefined rewards and has recently shown significant promise in enhancing the reasoning capabilities of LLMs [39]. RL approaches for boosting LLM reasoning can be broadly categorized into two main paradigms:

Off-policy Algorithms optimize policies using data collected from behavior policies that differ from the one currently being trained. For example, Direct Preference Optimization (DPO) [25] fine-tunes LLMs on pairwise preference data and achieves substantial improvements when combined with MCTS [46, 38]. However, the acquisition of paired preference data is resource-intensive. Recent innovations have addressed this limitation: Kahneman-Tversky Optimization (KTO) [47] and Step-KTO [18] allow training using individual preference instances, while SimPO [23] reduces computational overhead by eliminating the reference model during training.

On-policy Algorithms optimize policies using data generated by the current policy under training. Methods such as PPO [27] and Reinforce++ [15] have been widely adopted to improve LLM reasoning. These techniques typically incorporate an additional reward model to evaluate the outputs produced by the policy, thereby mitigating the reliance on expert-level data annotation. However, they incur substantial computational costs due to the additional reward model training, and reward hacking [10] can also introduce unintended system vulnerabilities. To overcome these challenges, GRPO [28] employs group-based sampling and relative advantage estimation to approximate reward model functionality while demonstrating impressive efficacy on complex reasoning tasks [10]. Building on this framework, recent approaches such as DAPO [43] and Dr. GRPO [20] offer more sophisticated strategies for optimizing complex reasoning.

Despite the notable advancements achieved by the GRPO series, their applicability in dynamic environments remains underexplored. These algorithms are primarily designed for single-turn reasoning tasks with fully observable contexts, thus neglecting the challenges posed by multi-turn interactive reasoning. This gap limits the generalization of these advanced reasoning methods to real-world scenarios, where partial observability and sequential decision-making are prevalent.

3 Identifying the Limitation of GRPO in Multi-turn Dynamic Environments

This section presents a systematic analysis of the limitations of GRPO in dynamic environments and discusses two promising modifications for adaptively applying GRPO with multi-turn interactive reasoning: step-wise optimization and rollout optimization.

Problem Formulation. The formulation of a dynamic environments can be seen as a Markov Decision Process (MDP): $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma \rangle$, where $\mathcal{S} = \{s_1, s_2, ..., s_n\}$ is the state space, $\mathcal{A} = \{a_1, a_2, ..., a_n\}$ is the action space, $\mathcal{P}$ is the transition probability, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function, and γ is the discount factor. The goal of the policy model is to maximize the reward $r = \mathbb{E}[\sum_{j=1}^{n} \gamma^j r(s_j, a_j)]$. Dinstict from the static environments, in a specific state s_j the action a_j is taken based on the previous state s_{j-1} and the action a_{j-1}. Consequently, LLMs must make decisions based on limited environmental information, which may lack all the critical clues necessary for solving the task. Moreover, state transitions update the context in ways not present in static scenarios, challenging LLMs to understand these changes and adapt to the evolving information.

The objective of GRPO for static environments is shown as follows:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = &\mathbb{E}_{x\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}(\cdot|x)}\Bigg[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\Bigg(\frac{\pi_\theta(o_{i,t}|x,o_{i,<t};)}{\pi_{\text{old}}(o_{i,t}|x,o_{i,<t};)}\hat{A}_{i,t},\\ &\text{clip}\Big(\frac{\pi_\theta(o_{i,t}|x,o_{i,<t};)}{\pi_{\text{old}}(o_{i,t}|x,o_{i,<t};)},1-\epsilon,1+\epsilon\Big)\hat{A}_{i,t}\Bigg)-\beta\mathbb{D}_{\text{KL}}\left[\pi_\theta||\pi_{\text{ref}}\right]\Bigg],\end{aligned} \tag{1}$$

where π_θ, π_{old} are the current and old policy. π_{ref} denotes the reference model. G indicates the size of the sampling group, o_i is the i-th sample, $\text{clip}(\cdot, 1-\epsilon, 1+\epsilon)$ is the clip function that ensure the stability of optimization. $\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^{G})}{\text{std}(\{r_i\}_{i=1}^{G})}$ is the advantage function, which is determined by the reward function. The KL term is defined as follows:

$$\mathbb{D}_{\text{KL}}\left[\pi_\theta||\pi_{\text{ref}}\right] = \frac{\pi_{\text{ref}}(o_{i,t}|x,o_{i,<t})}{\pi_\theta(o_{i,t}|x,o_{i,<t})} - \log\frac{\pi_{\text{ref}}(o_{i,t}|x,o_{i,<t})}{\pi_\theta(o_{i,t}|x,o_{i,<t})} - 1. \tag{2}$$

In the original GRPO, each rollout o_i is generated solely based on the initial task x, with subsequent reasoning operating as a static process that relies only on internal knowledge. In dynamic environments, however, state transitions provide essential information for problem-solving, as the policy must actively interact with the environment to gather further insights. Consequently, conventional GRPO is not well-suited for optimizing multi-turn interactive reasoning.

The way to extend GRPO to multi-turn interactive reasoning are two-fold:

Step-wise Optimization. Each environmental state is treated as the complete context, with the policy's verification of an action serving as its reward, *i.e.*, $\mathcal{D} = \{s_t, a_t^*\}$ where $a_t^* = \text{argmax}_{a_t} r(s_t, a_t)$ is the annotated optimal action given the state s_t. This approach aligns with off-policy reinforcement learning that utilizes step-wise training data [29]. However, step-wise optimization has notable limitations. First, collecting step-wise training data is resource-intensive and requires expert annotations. Second, it overlooks multi-turn dynamics, potentially leading to suboptimal outcomes, such as overthinking in a single turn [2], which increases risks of reduced efficiency and context overflow.

Rollout Optimization. The complete trajectory $\tau = \{s_0, a_1, s_1, a_2, \ldots, s_n\}$, generated by the policy, constitutes a rollout. This optimization accounts for all reasoning and actions within the problem-solving sequence, allowing the reward function to be defined at both step and trajectory levels. This RL objective can be formulated as: $\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},\tau\sim\pi_\theta(\cdot|x)}\left[r(x,\tau)\right]$. Compared to step-wise optimization, rollout optimization enables straightforward goal-oriented reward function, which is

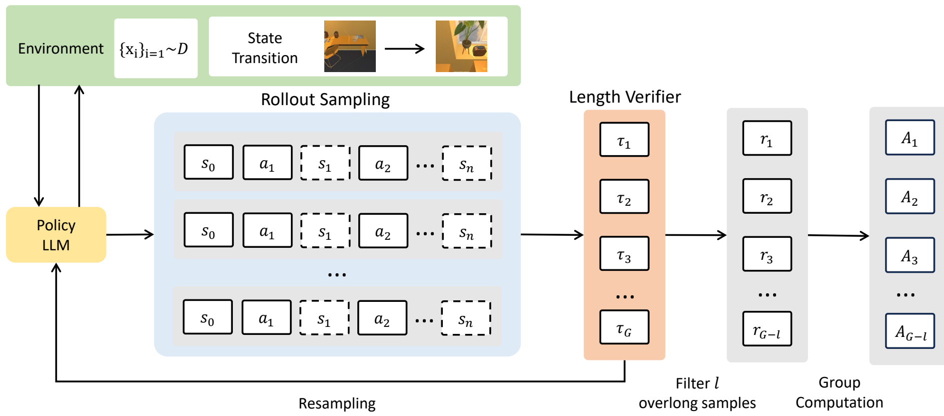

Figure 2: Illustrating the training procedure of DyPO. During each policy rollout, the model generates actions to interact with the environment. A length verifier then checks whether the trajectory exceeds a predefined maximum length; if it does, resampling or rejection sampling is employed. Additionally, during backpropagation, the initial context s_0 is preserved while subsequent observation tokens are masked to ensure optimization stability.

better aligned with the primary objective of the task. In addition, the trajectory-level rollout sampling offers greater flexibility as the policy can reason in the context across multiple states, enabling flexible decision-making such as exploration and exploitation.

4 Training Reasoning in Dynamic Environments

Based on the analysis above, we propose DyPO, a new RL-based post-training method designed for multi-turn interactive reasoning. The objective of DyPO is formulated as follows:

$$\mathcal{J}_{\text{DyPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_{\text{old}}(\cdot|x)} \sum_{i=1}^{G} \mathbb{I}(|y_i| < L_{\max}) \sum_{t=1}^{|y_i|}\sum_{j=1}^{n} \Bigg[\min\Bigg(\frac{\pi_\theta(y_{i,j,t}|x, y_{i,j,<t}; s_{i,<j})}{\pi_{\text{old}}(y_{i,j,t}|x, y_{i,j,<t}; s_{i,<j})} \hat{A}_{i,t},$$
$$\text{clip}\Bigg(\frac{\pi_\theta(y_{i,j,t}|x, y_{i,j,<t}; s_{i,<j})}{\pi_{\text{old}}(y_{i,j,t}|x, y_{i,j,<t}; s_{i,<j})}, 1-\epsilon, 1+\epsilon \Bigg) \hat{A}_{i,t} \Bigg) \Bigg], \text{ where } \hat{A}_{i,t} = r_i - \text{mean}(r). \tag{3}$$

Here, $y_{i,j,t}$ is the t-th token in the j-th turn of the sample i. Besides, $s_{i,<j}$ is the state of the previous turns of the sample i, while $L_{\max}$ is the pre-defined maximum length of the rollout sample. r_i is the reward of a individual sample, and mean(r) indicates the average of rewards of all samples. The overview of DyPO is presented in Fig. 2. Based on the Equ. (3), DyPO introduces several key enhancements for multi-turn interactive reasoning compared to the GRPO framework.

Multi-turn Interactive Rollout Sampling. DyPO introduces multi-turn rollout sampling, enabling state transitions and dynamic action adjustments during the reasoning process. In this framework, we employ interactive tokens, <action> and </action>, to highlight the actions taken by LLMs. In each turn, the action enclosed within these tokens is extracted and sent to the environment, at which point the reasoning process pauses to await an observation. The environment then updates the observation and returns it to the LLM's context, wrapped with <observation> and </observation>. Combined with the standard thinking tokens <think> and </think>, this rollout mechanism supports interactive reasoning that extends beyond static, internally-driven reasoning.

Reasoning Pattern Shifting. Prior knowledge encoded in the pre-training stage significantly influences RL training [8]. Because general LLMs are not tailored to dynamic environments, pattern shifting is inevitable. To promote multi-turn interactive reasoning and mitigate the influence of prior knowledge, we remove the KL divergence term. This term serves as a regularizer that encourages the policy to match the reference policy, thereby limiting the flexibility of DyPO optimization. Consequently, this modification enhances the adaptability and overall robustness of our method.

Optimization Stabilization. The interactive reasoning process incorporates environmental observations at each state. Since the policy model does not generate these observations but rather provides them externally, including their tokens in the loss computation introduces significant instability to the optimization process. To mitigate this issue, we implement a selective masking mechanism that excludes observation tokens from the rollout during loss computation, ensuring that the policy is only optimized based on the tokens it generates. This approach maintains training stability while preserving the interactive nature of the reasoning process. Furthermore, the standard deviation term in the advantage function disproportionately weights outlier samples (both extremely easy and difficult ones), introducing additional training instability [20]. We eliminate the standard deviation term in computing advantage, resulting in a more balanced reward signal across samples of varying difficulty.

Mitigating Overthinking. The overthinking phenomenon becomes particularly pronounced when the policy generates both thinking and actions in each turn. This issue is especially critical in multi-turn reasoning scenarios, where excessive generation in intermediate turns can lead to context length overflow, causing reasoning processes to be truncated before loss computation. This truncation exacerbates overthinking in later training stages, as GRPO inherently exhibits a length bias that favors longer responses [20]. As responses grow progressively longer across turns, training ultimately collapses when all samples exceed the truncation threshold.

To mitigate this risk, we eliminate the preference for longer responses by removing the normalization term $\frac{1}{\sum_{i=1}^{G}|o_i|}$ from the objective. Additionally, we implement both resampling and rejection sampling strategies in the rollout sampling process to further reduce overthinking. Specifically, when the policy generates a rollout exceeding the maximum allowable length, we resample a new rollout from the policy up to a predefined maximum number of attempts. If over-long rollouts persist after resampling, we reject these samples by assigning them zero weight in the loss computation, effectively preventing them from contributing to model updates.

5 Experiment

This section evaluates the efficacy of DyPO within dynamic environments. We proceed as follows: (1) describe the experimental setup (Sec. 5.1), (2) present the primary results and summarize the findings that DyPO demonstrates (Sec. 5.2).

5.1 Setup

Baselines. We employ the Qwen2.5-3B-Instruct model as our backbone in empirical studies. We evaluate our proposed DyPO method against two baselines: the training-free ReAct approach and the original GRPO method using step-wise training data:

- **ReAct:** This method prompts the base model to reason before taking actions in the environment. Observations from the environment are appended to the context for subsequent reasoning.
- **GRPO:** We curated trajectory-level correct interactive solutions and decomposed them into step-by-step actions for each turn and train the base model with the original GRPO objective.

To further demonstrate the effectiveness and adaptivity of DyPO, we train two instantiations: (1) **DyPO-Zero:** this variant applies DyPO directly to the base model without using step-wise instruction data, and (2) **DyPO:** this variant uses the checkpoint trained from the GRPO baseline on step-wise instruction data and further trains on the sampling at the trajectory level.

Datasets. We evaluate our method on two representative interactive environments: ALFWorld [31], and WebShop [40]. These environments are inherently dynamic with the partially observable nature, requiring LLMs to engage in strategic decision-making by taking actions and processing new observations to complete tasks, rather than relying solely on reasoning based on the initial context.

- **ALFWorld** is a synthetic text-based simulated environment which aligns with the embodied benchmark ALFRED [30]. It includes 6 distinct types of household tasks to systematically evaluate the multi-turn reasoning ability in real-world dynamic environments. In a specific task of ALFWorld, the model is initially provided with the description of seen objects (*e.g.*, a towelholder) and the goal of the task (*e.g.*, put some soapbar on cart). Then the model is required to explore the environment strategically and find target objects to complete the task.
- **WebShop** simulates an online shopping environment. This benchmark collects genuine web-shopping instructions and challenges models to follow these directions to purchase items that fulfill

Table 1: The main results of DyPO compared with ReAct and GRPO. All of the results are conducted using Qwen-2.5-3B-Instruct. We split the results into two parts: without step-wise data post-training and with step-wise data post-training. The success indicates the average task completion rate across five samples, and pass@5 indicates whether the model can complete the task within 5 samples.

Method	**ALFWorld**				**WebShop**		**Average**	
	valid_seen		valid_unseen					
	success	pass@5	success	pass@5	success	pass@5	success	pass@5
w/o step-wise data								
ReAct	6.96	25.93	7.61	28.35	15.61	69.23	9.86	40.42
DyPO-Zero	**13.71**	**30.00**	**14.63**	**41.00**	**56.48**	**91.50**	**28.27**	**54.17**
w/ step-wise data								
GRPO	21.57	51.43	15.52	44.03	53.36	96.50	30.15	63.65
DyPO	**31.29**	**55.71**	**25.37**	**56.72**	**57.21**	**97.00**	**37.95**	**69.81**

Table 2: We evaluate performance across varying rounds of interaction on both in-distribution and out-of-distribution splits of ALFWorld. By adjusting the maximum number of interactive turns from 10 to 30, we assess the success rate and pass@5 rate for ReAct and DyPO.

Benchmark	Method	Interactive Turn					
		10		20		30	
		success	pass@5	success	pass@5	success	pass@5
ALFWorld-unseen	ReAct	5.82	19.40	7.61	28.35	5.97	23.88
	DyPO-Zero	**8.51**	**17.91**	**14.63**	**41.00**	**12.69**	**33.58**
ALFWorld-seen	ReAct	6.85	23.57	6.96	25.93	6.71	22.86
	DyPO-Zero	**11.14**	**22.14**	**13.71**	**30.00**	**15.86**	**35.71**

specific user preferences (*e.g.*, "I need a 9.5 rubber-soled hiking shoe made of vinyl acetate"). The reward is determined by how well the purchased item aligns with the user's preference.

For ALFWorld, we use the train split tasks as the training data with 3553 samples, and the valid_seen and valid_unseen splits as the test set to assess the in-distribution and out-of-distribution performance. In WebShop, we manually split the whole 6000 tasks, with the former 3600 as the training split, and the latter 200 samples as the test split.

To bootstrap the reasoning before interactive actions, and help the policy understand the dynamic environment, we formulate the generation format in <think> and <action>. and modify the system prompt in both training and evaluation stages.

Training Details. We conducted all experiments on 8 NVIDIA H20 GPUs with 90 GB of memory. The global batch size is fixed as 14, with 2 gradient accumulation steps. The implementation of training and inference uses Transformer Reinforcement Learning and VLLM, respectively.

Evaluation Details. In the evaluation stage, we sample five reasoning trajectories per task and assess model performance using two metrics: success rate and pass@5. Success rate measures the average performance across all five samples, while pass@5 indicates whether the task is solved in at least one of the five attempts. For ALFWorld, both metrics directly reflect task completion. For WebShop, the success rate represents the average reward obtained upon task completion, as this benchmark employs fine-grained rewards measuring how well purchased products match user preferences. In this context, pass@5 indicates whether there are a reward greater than 0 in five samples.

5.2 Main Results

We show the main results in Tab. 1 and summarize the key findings:

Finding: DyPO significantly outperforms the baselines in dynamic environments. DyPO demonstrates superior performance across both ALFWorld and WebShop environments, consistently outperforming baseline methods across all evaluation metrics. Without step-wise data post-training, DyPO-Zero achieves substantial improvements over ReAct, with an 18.41% increase in success rate and a 13.7% gain in pass@5 rate. When applied as a post-training method to models already tuned with step-wise data, DyPO further enhances GRPO's performance by 3.43% in success rate and

Table 3: We report the average number of interactive turns required by the model in dynamic environments, with a maximum of 20 interactive turns. Results encompass all outcomes, including both successful and unsuccessful solutions.

Method	**ALFWorld**		**WebShop**	**Average**
	valid_seen	valid_unseen		
ReAct	19.18	19.24	15.25	17.89
DyPO-Zero	**18.52**	**18.46**	**11.03**	**16.00**
GRPO	17.20	18.05	6.81	14.02
DyPO	**16.57**	**17.76**	**6.46**	**13.60**

4.41% in pass@5 rate. These results indicate the effectiveness of DyPO in dynamic environments. and highlight DyPO's versatility as a flexible post-training algorithm that can effectively augment the capabilities of models previously optimized with step-wise data.

Finding: DyPO shows better adaptability to Out-of-Distribution task. Our experimental results demonstrate DyPO's robust generalization capabilities across both seen and unseen environments in ALFWorld. Without step-wise data post-training, DyPO-Zero maintains consistent performance between valid_seen and valid_unseen scenarios, while GRPO exhibits a notable performance disparity between in-distribution and out-of-distribution tasks. When applied as a post-training method, DyPO yields substantial improvements, achieving performance gains of 3.86% and 2.54% on valid_seen and valid_unseen environments respectively, compared to GRPO. These findings underscore DyPO method's effectiveness in enhancing model generalization, particularly valuable for real-world applications where agents frequently encounter novel, previously unseen environments.

To further evaluate DyPO's performance characteristics, we analyze its scalability across varying interaction rounds (Tab. 2) and examine the average number of interactions required to complete tasks across different methods (Tab. 3).

Finding: DyPO demonstrates scalable performance across different interactive turns. DyPO-Zero exhibits superior scalability compared to ReAct as the number of interaction rounds increases. Specifically, in the valid_unseen split of ALFWorld, DyPO achieves consistent performance improvements, demonstrating an 11.57% increase in pass@5 rate when scaling from 10 to 30 interaction rounds. In contrast, ReAct's performance deteriorates with increased rounds. These results suggest that DyPO training enhances the model's capacity for long-term interactive reasoning, enabling it to more effectively utilize environmental observations throughout extended interaction sequences.

Finding: DyPO Enhances Task Resolution Efficiency Training with DyPO significantly reduces the average number of interactive rounds compared to ReAct. Specifically, DyPO-Zero achieves a 1.89-turn reduction on average. When combined with GRPO, DyPO further decreases the number of turns required, highlighting its enhanced efficiency. Analysis of reasoning behaviors reveals (see Sec . D) that DyPO's trajectory-level optimization drives this improvement, as the model better leverages previously collected information. In contrast, GRPO-trained models often struggle to retain and utilize such information effectively.

5.3 Training Dynamics

Figure 3 illustrates the reward trends for Qwen-2.5-3B-Instruct and Qwen-2.5-1.5B-Instruct during training on ALFWorld and WebShop with GRPO and DyPO. Key findings are summarized below:

Finding: DyPO consistently improves performance during training, whereas GRPO yields only marginal gains. The reward curves for DyPO demonstrate consistently higher and more stable improvements across training. On the WebShop dataset, both models achieve approximately a 0.45 increase in completion rate after 200 training steps. Similarly, on the ALFWorld dataset, the reward improvement reaches around 0.15. In contrast, GRPO provides limited improvement in selecting optimal actions. For ALFWorld, the training curve exhibits an oscillating trend with no significant performance gains. For WebShop, although performance improves after training, the process is less stable, as evidenced by the oscillating training curves.

Finding: Larger models perform better in DyPO, while GRPO offers equally despite the model size. In DyPO, the performance gap between Qwen-2.5-3B and Qwen-2.5-1.5B during training is pronounced, with the larger model showing a greater reward increase. In contrast, GRPO training curves remain similar regardless of model size. This suggests that DyPO better leverages model scaling, enabling larger base models to achieve superior performance during training, while GRPO's performance gains are less dependent on model size.

Figure 3: To illustrate changes in reward trends during training, we present training curves for four settings: GRPO in ALFWorld (**top left**), GRPO in WebShop (**top right**), DyPO-Zero in ALFWorld (**bottom left**), and DyPO-Zero in WebShop (**bottom right**), using the Qwen-2.5-1.5B-Instruct and Qwen-2.5-3B-Instruct models. For GRPO, we report the exact match reward, while for DyPO-Zero, we report the task completion reward.

6 Conclusion and Further Discussion

Conclusion. Current cutting-edge RL reasoning methods often overlook the importance of interactive capabilities in dynamic environments. To address this limitation, we introduce a novel approach, DyPO, which enables LLMs to interact with environments and continuously reason over multi-turn interactions. To ensure high-quality and stable RL training, we modify the GRPO objective function and incorporate both a length verifier and environmental observation token masking. We present two variants, DyPO-Zero and DyPO depending on whether the base model is trained with step-wise instruction data. Empirical studies on ALFWorld and WebShop demonstrate that DyPO not only improves performance but also enhances adaptability in dynamic settings. These results highlight DyPO as a promising method for advancing both the reasoning and interactive capabilities of LLMs in complex, multi-turn environments.

Limitations. Despite the effectiveness of DyPO, our work has several limitations. First, our experiments were conducted only on two environments, ALFWorld and WebShop. Future work should evaluate DyPO in more complex and diverse settings, such as intricate video game scenarios. Second, our empirical study was limited to the Qwen-2.5-3B-Instruct model; extending the evaluation to larger and more powerful models (e.g., Qwen-2.5-72B) would provide further validation of DyPO's effectiveness. Finally, we have not explored DyPO in the context of visual-language models. Investigating its performance on such models would further establish the method's versatility.

Broader Impacts. Dynamic environments are common in real-world applications, like online shopping, GUIs, and gaming, challenging GRPO methods in multi-turn tasks. Our proposed DyPO framework offers a principled approach to enhance both the reasoning and interactive abilities of models in these settings. This work highlights the potential of optimization algorithms tailored for dynamic environments as a promising research direction, potentially unlocking new pathways for LLMs to surpass the limitations of static reasoning tasks. Our investigation thus provides a novel perspective on advancing large language models, particularly in domains that require continuous interaction with evolving environments.

References

[1] Léon Bottou. From machine learning to machine reasoning: An essay. *Machine learning*, 94: 133–149, 2014.

[2] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

[3] Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*, pages 41–75. Springer, 2019.

[4] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. In *NeurIPS*, 2023.

[5] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

[6] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022.

[7] Xidong Feng, Yicheng Luo, Ziyan Wang, Hongrui Tang, Mengyue Yang, Kun Shao, David Mguni, Yali Du, and Jun Wang. Chessgpt: Bridging policy learning and language modeling. In *NeurIPS*, 2023.

[8] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

[9] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

[10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[11] Akshat Gupta. Are chatgpt and gpt-4 good poker players?–a pre-flop analysis. *arXiv preprint arXiv:2308.12466*, 2023.

[12] Bo Han, Jiangchao Yao, Tongliang Liu, Bo Li, Sanmi Koyejo, and Feng Liu. Trustworthy machine learning: From data to models. *Foundations and Trends® in Privacy and Security*, 7(2-3):74–246, 2025. ISSN 2474-1558. doi: 10.1561/3300000043. URL http://dx.doi.org/10.1561/3300000043.

[13] Alex Havrilla and Maia Iyer. Understanding the effect of noise in llm training data with algorithmic chains of thought. *arXiv preprint arXiv:2402.04004*, 2024.

[14] Amy K Hoover, Julian Togelius, Scott Lee, and Fernando de Mesentier Silva. The many ai challenges of hearthstone. *KI-Künstliche Intelligenz*, 34:33–43, 2020.

[15] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

[16] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

[17] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022.

[18] Yen-Ting Lin, Di Jin, Tengyu Xu, Tianhao Wu, Sainbayar Sukhbaatar, Chen Zhu, Yun He, Yun-Nung Chen, Jason Weston, Yuandong Tian, et al. Step-kto: Optimizing mathematical reasoning through stepwise binary feedback. *arXiv preprint arXiv:2501.10799*, 2025.

[19] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886*, 2024.

[20] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[21] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jian-Guang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. In *The Thirteenth International Conference on Learning Representations*, 2025.

[22] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 2024.

[23] Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[24] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[25] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

[26] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim GJ Rudner, Chia-Man Hung, Philip HS Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*, 2019.

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[28] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[29] Wentao Shi, Mengqi Yuan, Junkang Wu, Qifan Wang, and Fuli Feng. Direct multi-turn preference optimization for language agents. *arXiv preprint arXiv:2406.14868*, 2024.

[30] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.

[31] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *ICLR*, 2021.

[32] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019.

[33] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning, 2020.

[34] Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? 2022.

[35] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[37] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *NeurIPS*, 2024.

[38] Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*, 2024.

[39] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*, 2025.

[40] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In *NeurIPS*, 2022.

[41] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2023.

[42] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

[43] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[44] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

[45] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*, 2023.

[46] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. ReST-MCTS*: LLM self-training via process reward guided tree search. In *NeurIPS*, 2024.

[47] Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, et al. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. In *ICLR*, 2024.

[48] Zhanke Zhou, Xiao Feng, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, and Bo Han. From passive to active reasoning: Can large language models ask the right questions under incomplete information? In *ICML*, 2025.

Appendix

A Dataset Introduction 13

B Implementation Details 15

C Further Experiments Analysis 18

D Case Studies 18

D.1 ALFWorld Task 18

D.2 Case Using ReAct in ALFWorld 19

D.3 Case Using GRPO in ALFWorld 20

D.4 Case Using DyPO-Zero in ALFWorld 23

D.5 Case Using DyPO in ALFWorld 24

D.6 WebShop Task 24

D.7 Case Using ReAct in WebShop 25

D.8 Case Using GRPO in WebShop 28

D.9 Case Using DyPO-Zero in WebShop 29

D.10 Case Using DyPO in WebShop 31

A Dataset Introduction

Table 4: The detailed statistics of ALFWorld dataset.

Split	Pick & Place	Examine in Light	Clean & Place	Heat & Place	Cool & Place	Pick Two & Place	All
#Train	790	308	650	459	533	813	3553
#Valid_seen	35	13	27	16	25	24	140
#Valid_unseen	24	18	31	23	21	17	134

ALFWorld [31] is a synthetic text-based simulated environment which aligns with the embodied benchmark ALFRED [30]. It systematically evaluates the multi-turn reasoning ability in real-world dynamic environments, which involves 6 distinct types of household tasks: Pick & Place, Examine in Light, Clean & Place, Heat & Place, Cool & Place, Pick Two & Place. An example of the reasoning in ALFWorld is presented in Fig. 4. The details of each task are shown as follows:

- **Pick & Place.** (*e.g.,* put a toiletpaper in toiletpaperhanger.) The LLMs need to find the target object and pick it up, move to the target receptacle and place it.
- **Examine in Light.** (*e.g.,* examine the pencil with the desklamp.) The LLMs need to find the target object, pick it up, find a place has light, and turn on the light to examine the object on hand.
- **Clean & Place.** (*e.g.,* clean some soapbar and put it in cart.) The LLMs need to find the target object, pick it up, find a place has water (*e.g.*, sink and basin), wash the object, and then find the correct place to put it down.

- **Heat & Place.** (*e.g.,* heat some egg and put it in sidetable.) The LLMs need to find the target object, pick it up, go to a somewhere has heat source (*e.g.*, stove and microwave),heat the object, and find the correct place to put it down.
- **Cool & Place.** (*e.g.,* cool some pot and put it in diningtable.) The LLMs need to find the target object, pick it up, find a place has cool source (*e.g.*, fridge), and cool the object. Then move to the target receptacle and place it.
- **Pick Two & Place.** (*e.g.,* put two kettle in cabinet.) The LLMs need to find one target object, pick it up, and move to the target receptacle and place it, then find another target object, pick it up, and move to the target receptacle and place it.

An example of ALFWorld reasoning

-= Welcome to TextWorld, ALFRED! =-
You are in the middle of a room. Looking quickly around you, you see a cabinet 13, a cabinet 12, a cabinet 11, a cabinet 10, a cabinet 9, a cabinet 8, a cabinet 7, a cabinet 6, a cabinet 5, a cabinet 4, a cabinet 3, a cabinet 2, a cabinet 1, a coffeemachine 1, a countertop 1, a drawer 5, a drawer 4, a drawer 3, a drawer 2, a drawer 1, a fridge 1, a garbagecan 1, a microwave 1, a sinkbasin 1, a stoveburner 4, a stoveburner 3, a stoveburner 2, a stoveburner 1, and a toaster 1.
Your task is to: put some lettuce on countertop.

> go to drawer 2

You arrive at drawer 2. The drawer 2 is closed.

> take lettuce 2 from sinkbasin 1

You pick up the lettuce 2 from the sinkbasin 1.

> go to coffeemachine 1

You arrive at coffeemachine 1. On the coffeemachine 1, you see nothing.

> move lettuce 2 to countertop 1

You move the lettuce 2 to the countertop 1.

Figure 4: In the ALFWorld reasoning trajectory, the **bold** text denotes the action taken in each turn, while the *italicized* text represents the environment's observation in response to the action.

WebShop [40] is a large-scale interactive web-shopping environment featuring 1,181,436 products. These products are organized into five primary categories (beauty, food, fashion, furniture, and electronics), encompassing 313 fine-grained subcategories (e.g., "Women's Loafers & Slip-Ons" under fashion, "Pendants and Chandeliers" under furniture). Each product is defined by multiple attributes (e.g., target gender, age group) and customizable options (e.g., color, size). In a shopping task, LLMs navigate this structured product space to identify items matching the specified requirements and add them to the shopping cart. An example reasoning trajectory is provided in Fig. 5. Upon completing a shopping task (i.e., clicking the [buy now] button), the environment assigns a reward reflecting how well the selected items align with the task requirements. This reward is computed based on the attributes and options specified in the shopping task, as detailed below:

$$r = r_{\text{type}} \cdot \frac{|\mathbf{U}_{\text{att}} \cap \mathbf{Y}_{\text{att}}| + |\mathbf{U}_{\text{opt}} \cap \mathbf{Y}_{\text{opt}}| + \mathbf{1}\left[y_{\text{price}} \leq u_{\text{price}}\right]}{|\mathbf{U}_{\text{att}}| + |\mathbf{U}_{\text{opt}}| + 1}, \tag{4}$$

$$\text{where } r_{\text{type}} = \begin{cases} 0, & \text{if TextMatch}(\hat{y}, y^*) = 0 \\ 0.1, & \text{if TextMatch}(\hat{y}, y^*) < 0.1 \\ 0.5, & \text{if TextMatch}(\hat{y}, y^*) > 0.2 \text{ and } c = 1 \text{ and } f = 1 \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

Table 5: Hyperparameter settings used in our experiments for both training and evaluation phases. Since ReAct is a training-free method, its training hyperparameters are denoted with $\backslash$.

Stage	Hyper-parameter	ReAct	GRPO	DyPO
Training	**Global Batch Size**	\	14	14
	Gradient Accumulation Steps	\	2	2
	Warmup Ratio	\	0.1	0.1
	Learning Rate	\	3.0e-06	3.0e-06
	Learning Rate Scheduler	\	cosine	cosine
	Max Training Steps	\	200	200
	Max Completion Length	\	4096	6000
	Max Interaction Steps	\	20	20
	Inference Temperature	\	0.9	0.9
Evaluation	**Max Interaction Steps**	20	20	20
	Max Tokens	32768	32768	32768
	Inference Temperature	0.7	0.7	0.7
	Sampling Size	5	5	5

Here, $\mathbf{U}_{\text{att}}$ and $\mathbf{U}_{\text{opt}}$ represent the sets of required attributes and options in the shopping task, while $\mathbf{Y}_{\text{att}}$ and $\mathbf{Y}_{\text{opt}}$ denote the corresponding sets for the purchased items. y_{price} is the price of the purchased items, and u_{price} represents the price requirement specified in the task. The variables c and f indicate whether the main category ($c = 1$ if matched) and fine-grained sub-category ($f = 1$ if matched) of the purchased items match the task requirements. Finally, $\hat{y}$ represents the title of the items chosen by the LLM, and y^* is the title of the ground truth items for the shopping task.

B Implementation Details

In this section, we explicitly introduce the detailed implementation of DyPO.

Framework Implementation. In this work, we implement DyPO using the Transformer Reinforcement Learning [33] (TRL) framework, version 0.15.0. As discussed in Section 3, the original GRPO trainer in TRL is incompatible with multi-turn reasoning tasks. To address this, we modified both the sampling and training processes of the GRPO trainer to support multi-turn reasoning and developed DyPO based on this adapted trainer. For inference, we utilize vLLM, version 0.7.2, with the default V0 engine across all experiments. For the baseline, we directly implemented the GRPO method using the original GRPO trainer in TRL.

Hyperparameters. Table 5 presents the comprehensive hyperparameter settings used in DyPO. During training, we carefully constrain the maximum completion length to prevent Out Of Memory (OOM) errors, as longer token sequences consume proportionally more GPU memory. For evaluation, we set a generous maximum token limit of 32,768 to accommodate even the longest reasoning trajectories. To ensure robust and reliable experimental results, we conduct 5 sampling runs for each model configuration.

Prompt Template. We customize the system prompt and environmental information format for each specific task to enhance the policy's task comprehension and environmental awareness, ultimately improving training stability and performance. For ALFWorld (Tab. 6), the system prompt provides a comprehensive description of the environment, interaction protocols between the model and environment, and valid action templates. We structure environmental information using `<observation>` `</observation>` tags to encapsulate environmental observations. When the model attempts an invalid action, we provide feedback by replacing the environmental information with a description of valid actions. For WebShop (Tab. 7), we employ a similar approach but with additional emphasis on explicitly delineating valid actions at each step, given the complexity of web page environmental information. This explicit action guidance helps ensure the policy maintains a clear understanding of available actions throughout the interaction process.

Reward Function. The reward function plays a crucial role in training policies for both DyPO and GRPO approaches. For DyPO, we evaluate the entire trajectory of sampled actions and assign rewards

An example of WebShop reasoning

Instruction: Find me butt lifting, light weight women's shorts with high waist, tummy control with color: black, and size: 3x-large, and price lower than 50.00 dollars

> search[butt lifting, light weight women's shorts with high waist, tummy control]

[B09QCP4579]: LWLW High Waist Yoga Shorts for Women,Tummy Control Biker Shorts Exercise Workout Butt Lifting Tights Women's Short Pants $13.99
[B09QCVCYVY]: Burband Womens High Waist Bike Shorts Tummy Control Workout Yoga Pants Running Exercise Running Shorts with Side Pockets $6.82
[B09Q37JQZ6]: Women's Sexy Swimsuit One Piece High Neck Halter Bikini Floral Stiching See Through Monokini Tummy Control Beachwear $10.99 to $18.99
[B085S5P7WB]: Women High Waist Tummy Control Fitness Yoga Pants Mesh Patchwork Leggings Sexy Lace Stitching Tights $0.01 to $2.99
[B09PL5W9PD]: Women's One Piece Swimsuit Halter Plunge Neck Tummy Control Bathing Suits Push Up Tankini Sets Plus Size Beachwear $7.8
[B09S632DT3]: Sleepwear Womens Chemise Nightgown Full Slip Lace Lounge Dress with Briefs Mesh Chemise V Neck Soft Pajama Dress Nightdress $11.99 to $14.99
[B09PBPZ24Z]: myhehthw Women's High Waisted Jeans for Women Distressed Ripped Jeans Slim Fit Butt Lifting Skinny Stretch Jeans Trousers $22.99 to $25.99
[B09PVNLVRW]: Women's V-Neck Rompers Printed Jumpsuit Long Sleeve Homewear Butt Flap Pajamas One-Piece Onesies Nightwear Sexy Bodysuit $17.4 to $28.67

> click[B09QCP4579]

LWLW High Waist Yoga Shorts for Women,Tummy Control Biker Shorts Exercise Workout Butt Lifting Tights Women's Short Pants
color: [black], [blue], [gray], [purple], [wine]
size: [small], [medium], [large], [x-large], [xx-large], [3x-large]
Price: $13.99
Rating: N.A.
[Description], [Features], [Reviews], [Buy Now]

> click[3x-large]

LWLW High Waist Yoga Shorts for Women,Tummy Control Biker Shorts Exercise Workout Butt Lifting Tights Women's Short Pants
color: [black], [blue], [gray], [purple], [wine]
size: [small], [medium], [large], [x-large], [xx-large], [3x-large]
Price: $13.99
Rating: N.A.
[Description], [Features], [Reviews], [Buy Now]

> click[black]

LWLW High Waist Yoga Shorts for Women,Tummy Control Biker Shorts Exercise Workout Butt Lifting Tights Women's Short Pants
color: [black], [blue], [gray], [purple], [wine]
size: [small], [medium], [large], [x-large], [xx-large], [3x-large]
Price: $13.99
Rating: N.A.
[Description], [Features], [Reviews], [Buy Now]

> click[buy now]

Thank you for shopping with us!

Figure 5: In the WebShop reasoning trajectory, **bold** text indicates the action taken in each turn, *italicized* text represents the environment's response to the action, and [button] denotes interactive buttons on the shopping website that LLMs can engage with.

Table 6: The prompt templates for ALFWorld. For valid actions, the system provides environmental observations within <observation> </observation> tags. For invalid actions, the system provides feedback by listing the valid action templates instead of environmental observations.

System Prompt	A conversation between user and assistant. The user describes an environment, and the assistant complete tasks in it through acting and observing. The assistant thinks about the reasoning process, takes an action, and receives observations from the environment provided by system, and continue thinking, taking action based on the received observation from system. Assistant should keep thinking and taking action until the system says the task is completed. The assistant's reasoning process, action, should be enclosed within example reasoning per turn: <think> the task is to move the apple to the table, I need to find the apple first</think> <action>go to fridge</action>. After reasoning the environment provides the observation enclosed within <observation> </observation> tags. Valid actions: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>.
Valid Feedback	<observation>[information about the environment]</observation>.
Invalid Feedback	<observation>Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object></observation>.

Table 7: The prompt templates used in WebShop, different from ALFWorld, we do not check whether the action is valid or not, we just provide the valid action templates with current environmental information.

System Prompt	A conversation between user and assistant. The user describes an online shopping task, and the assistant completes the task through acting and observing. The assistant thinks about the reasoning process, takes an action, and receives observations from the shopping website provided by system, and continue thinking, taking action based on the received observation from system. Assistant should keep thinking and taking action until the system says the task is completed. The assistant's reasoning process, action, should be enclosed within example reasoning per turn: <think>the task is to find a red shirt under $25, I need to search for shirts first</think> <action>search[red shirt]</action>After reasoning the environment provides the observation enclosed within <observation> </observation> tags. Valid actions include: search[QUERY] to search for products, and click[OPTION] to click on a specific option or product.
Feedback	<observation>[information about the environment]. Valid actions include: [Explicitly valid actions in this turn]</observation>.

based on task completion. Specifically, in ALFWorld, we use a binary reward scheme, assigning a reward of 1 for trajectories that successfully complete the task, and 0 otherwise. For WebShop, we utilize the environment's built-in reward function mentioned in Sec. A to evaluate the final outcome. In contrast, GRPO employs a simpler reward mechanism that checks for exact matches between the model's actions and the ground truth actions from the training data.

Table 8: We compared the performance of DyPO with ReAct and GRPO using two model sizes: Qwen-2.5-1.5B-Instruct and Qwen-2.5-3B-Instruct. The experimental setup follows the configuration outlined in Tab. 1.

Model	Method	ALFWorld				WebShop		Average	
		valid_seen		valid_unseen					
		success	pass@5	success	pass@5	success	pass@5	success	pass@5
Qwen-2.5-1.5B	ReAct	1.43	6.43	0.60	2.99	5.32	53.50	2.46	20.97
	GRPO	10.57	32.86	8.66	26.87	2.06	30.0	7.10	29.91
	DyPO	**26.86**	**50.71**	**25.67**	**50.00**	**62.67**	**95.5**	**38.40**	**65.40**
Qwen-2.5-3B	ReAct	6.96	25.93	7.61	28.35	15.61	69.23	9.86	40.42
	GRPO	21.57	51.43	15.52	44.03	53.36	96.50	30.15	63.65
	DyPO	**31.29**	**55.71**	**25.37**	**56.72**	**57.21**	**97.00**	**37.95**	**69.81**

C Further Experiments Analysis

Table 8 presents the empirical results of DyPO using Qwen-2.5-1.5B-Instruct and Qwen-2.5-3B-Instruct. Compared to ReAct and GRPO, DyPO with Qwen-2.5-1.5B-Instruct achieves significant performance gains, including a 31.3% improvement in average success rate and a 35.49% improvement in pass@5 rate over GRPO. These results highlight DyPO's strong scalability across model sizes and its effectiveness in multi-turn dynamic environments.

D Case Studies

In this section, we present case studies to analyze behavioral differences across reasoning methods. We examine reasoning trajectories extracted from the evaluation logs of our main results (Table 1), generated using the base or training checkpoints of Qwen-2.5-3B-Instruct. Below, we outline the cases studied:

- A task from the ALFWorld valid_unseen split (Lst. D.1):
  - ReAct (Lst. D.2 in Red)
  - GRPO (Lst. D.3 in Yellow)
  - DyPO-Zero (Lst. D.4 in Green)
  - DyPO (Lst. D.5 in Cyan)
- A task from the WebShop test split (Lst. D.6):
  - ReAct (Red Lst. D.7 in Red)
  - GRPO (Lst. D.8 in Yellow)
  - DyPO-Zero (Lst. D.9 in Green)
  - DyPO (Lst. D.10 in Cyan)

Each case is presented in detail, with key findings summarized below.

D.1 ALFWorld Task

ALFWorld Task

-= Welcome to TextWorld, ALFRED! =-

You are in the middle of a room. Looking quickly around you, you see a bed 1, a desk 2, a desk 1, a drawer 6, a drawer 5, a drawer 4, a drawer 3, a drawer 2, a drawer 1, a garbagecan 1, a laundryhamper 1, a safe 1, a shelf 6, a shelf 5, a shelf 4, a shelf 3, a shelf 2, and a shelf 1.

Your task is to: put a pencil in shelf.

This task requires the model to find a pencil in the room, pick up the pencil, find an empty shelf, and put the pencil in the shelf.

D.2 Case Using ReAct in ALFWorld

ReAct Response in ALFWorld

Assistant: <think>First, I need to find a pencil. Pencils are often found on desks or in drawers. I should start by looking at desk 1.</think><action>go to desk 1</action>

System: You arrive at desk 1. On the desk 1, you see a desklamp 1, and a pencil 1.

Assistant: <think>I've found a pencil on desk 1. Next, I need to take the pencil and put it on the shelf.</think><action>take pencil 1 from desk 1</action>

System: You pick up the pencil 1 from the desk 1.

Assistant: <think>I have the pencil. Now, I need to move the pencil to a shelf.</think><action>go to shelf 1</action>

System: You arrive at shelf 1. On the shelf 1, you see nothing.

Assistant: <think>Shelf 1 is empty. I will try a few more shelves to find an empty one.</think><action>go to shelf 2</action>

System: You arrive at shelf 2. On the shelf 2, you see a mug 1.

Assistant: <think>Shelf 2 has a mug. Let me try shelf 3.</think><action>go to shelf 3</action>

System: You arrive at shelf 3. On the shelf 3, you see a desklamp 1.

Assistant: <think>The shelf 3 already has a desklamp, so I will try the next one. Shelf 4.</think><action>go to shelf 4</action>

System: You arrive at shelf 4. On the shelf 4, you see a alarmclock 1.

Assistant: <think>Solid, shelf 4 also has an alarmclock. I will try shelf 5.</think><action>go to shelf 5</action>

System: You arrive at shelf 5. On the shelf 5, you see nothing.

Assistant: <think>Shelf 5 is empty. I will try the next one.</think><action>go to shelf 6</action>

System: You arrive at shelf 6. On the shelf 6, you see nothing.

Assistant: <think>Shelf 6 is empty. I will check one more shelf.</think><action>go to shelf 7</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>Shelf 7 does not exist. I will now try to find a pencil in the available shelves.</think><action>go to shelf 1</action>

System: You arrive at shelf 1. On the shelf 1, you see nothing.

Assistant: <think>Shelf 1 is empty. I will try the next available shelf, shelf 2.</think><action>go to shelf 2</action>

System: You arrive at shelf 2. On the shelf 2, you see a mug 1.

Assistant: <think>Shelf 2 has a mug. Let me try shelf 3.</think><action>go to shelf 3</action>

System: You arrive at shelf 3. On the shelf 3, you see a desklamp 1.

Assistant: <think>The shelf 3 already has a desklamp, so I will try the next one. Shelf 4.</think><action>go to shelf 4</action>

System: You arrive at shelf 4. On the shelf 4, you see a alarmclock 1.

Assistant: <think>Solid, shelf 4 also has an alarmclock. I will try shelf 5.</think><action>go to shelf 5</action>

System: You arrive at shelf 5. On the shelf 5, you see nothing.

Assistant: <think>Shelf 5 is empty. I will try the next one.</think><action>go to shelf 6</action>

System: You arrive at shelf 6. On the shelf 6, you see nothing.

Assistant: <think>Shelf 6 is empty. I will check one more shelf.</think><action>go to shelf 7</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>Shelf 7 does not exist. I will now try to find a pencil in the available shelves.</think><action>go to shelf 1</action>

System: You arrive at shelf 1. On the shelf 1, you see nothing.

Assistant: <think>Shelf 1 is empty. I will try the next available shelf, shelf 2.</think><action>go to shelf 2</action>

System: You arrive at shelf 2. On the shelf 2, you see a mug 1.

Assistant: <think>Shelf 2 has a mug. Let me try shelf 3.</think><action>go to shelf 3</action>

System: You arrive at shelf 3. On the shelf 3, you see a desklamp 1.

Assistant: <think>The shelf 3 already has a desklamp, so I will try the next one. Shelf 4.</think><action>go to shelf 4</action>

System: You arrive at shelf 4. On the shelf 4, you see a alarmclock 1.

ReAct is unable to complete this task effectively. Initially, the model correctly understands the task, identifying pencil 1 on desk 1 and beginning the search for an empty shelf. However, despite checking all available shelves, the model fails to place the pencil on an empty shelf, even when one is found. Furthermore, the model searches for a nonexistent "shelf 7," which is not present in the room. These observations suggest that, while the base model with ReAct can gather environmental information, it struggles to extract critical details and effectively utilize environmental feedback.

D.3 Case Using GRPO in ALFWorld

GRPO Response in ALFWorld

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: You arrive at desk 1. On the desk 1, you see a desklamp 1, and a pencil 1.

Assistant: <think>I found the pencil, now I need to move it to the shelf. The closest shelf is shelf 1.</think> <action>take pencil 1 from desk 1</action>

System: You pick up the pencil 1 from the desk 1.

Assistant: <think>Now that I have the pencil, I need to move it to the shelf.</think> <action>go to shelf 1</action>

System: You arrive at shelf 1. On the shelf 1, you see nothing.

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: You arrive at desk 1. On the desk 1, you see a desklamp 1.

Assistant: <think>I found the pencil, now I need to move it to the shelf. The closest shelf is shelf 1.</think> <action>take pencil 1 from desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

Assistant: <think>I need to find a pencil first. The most likely places to find a pencil are the desk and the drawer.</think> <action>go to desk 1</action>

System: Invalid action, valid actions should be enclosed in tags <action> </action> and the valid actions should be in the following template: look, inventory, go to <receptacle>, open <receptacle>, close <receptacle>, take <object> from <receptacle>, move <object> to <receptacle>, examine <something>, use <object>, heat <object> with <receptacle>, clean <object> with <receptacle>, cool <object> with <receptacle>, slice <object> with <object>

The GRPO model demonstrates significant execution shortcomings in the ALFWorld task. Although it correctly interprets the task and successfully picks up the pencil, the model becomes trapped in an invalid action loop, repeatedly attempting to navigate to desk 1 and failing to utilize the identified empty shelf (e.g., shelf 1) to complete the task. The model's lack of state memory prevents it from recalling that it already holds the pencil or has explored specific environmental information, resulting in poor adaptability to system feedback, such as invalid action prompts. These deficiencies underscore GRPO's limitations in effective state management and feedback utilization in dynamic environments, leading to reduced execution efficiency and task incompletion.

D.4 Case Using DyPO-Zero in ALFWorld

DyPO-Zero Response in ALFWorld

Assistant: <action> look </action>

System: You are in the middle of a room. Looking quickly around you, you see nothing.

Assistant: <action> inventory </action>

System: You are not carrying anything.

Assistant: <action> go to desk 1 </action>

System: You arrive at desk 1. On the desk 1, you see a desklamp 1, and a pencil 1.

Assistant: <action> take pencil 1 from desk 1 </action>

System: You pick up the pencil 1 from the desk 1.

Assistant: <action> go to shelf 1 </action>

System: You arrive at shelf 1. On the shelf 1, you see nothing.

Assistant: <action> move pencil 1 to shelf 1 </action>

System: You move the pencil 1 to the shelf 1.

DyPO-Zero exhibits marked improvements over ReAct and GRPO in the ALFWorld task. Unlike ReAct, which accurately identifies pencil 1 on desk 1 but fails to complete the task due to repeated shelf exploration and attempts to access a nonexistent "shelf 7," DyPO-Zero executes a streamlined action sequence with high efficiency. It initiates the task with a "look" action to assess the environment, checks its inventory, locates pencil 1 on desk 1, picks it up, and promptly navigates to shelf 1, confirming its availability before successfully placing the pencil. In contrast, GRPO becomes trapped in an invalid action loop after retrieving the pencil and fails to leverage known information. DyPO-Zero's robust state management prevents repetitive or erroneous actions, demonstrating superior integration of environmental feedback, sustained task focus, and efficient task completion, underscoring DyPO-Zero's enhancement of decision-making and adaptability in dynamic environments.

D.5 Case Using DyPO in ALFWorld

DyPO Response in ALFWorld

Assistant: <think>I need to find a pencil first. Pencils are often found on desks or in drawers.</think> <action>go to desk 1</action>

System: You arrive at desk 1. On the desk 1, you see a desklamp 1, and a pencil 1.

Assistant: <action>take pencil 1 from desk 1</action>

System: You pick up the pencil 1 from the desk 1.

Assistant: <think>I need to move the pencil to a shelf.</think> <action>go to shelf 1</action>

System: You arrive at shelf 1. On the shelf 1, you see nothing.

Assistant: <action>move pencil 1 to shelf 1</action>

System: You move the pencil 1 to the shelf 1.

DyPO exhibits significant advancements over DyPO-Zero in the ALFWorld task. Notably, DyPO finds the optimal solution trajectory of this task. Unlike DyPO-Zero, which relies on diminishing thinking processes during its reasoning trajectory, DyPO adaptively engages in reasoning, optimizing its action sequence. It avoids unnecessary actions, such as redundant "look" or "inventory" observed in DyPO-Zero's reasoning process, thereby enhancing efficiency. These improvements highlight DyPO's superior ability to dynamically adjust its reasoning and execute precise, purposeful actions.

D.6 WebShop Task

WebShop Task

Instruction: Find me home office furniture sets for dining room, living room with color: taupe, and item shape: round, and size: 21 in x 35 in, and price lower than 60.00 dollars

This task presents a shopping scenario where a user provides instructions to purchase an office furniture set according to their preferences. The model must execute the appropriate actions to complete the purchase successfully.

D.7 Case Using ReAct in WebShop

ReAct Response in WebShop

assistant: <think>According to the instruction, I need to find dining room and living room furniture sets with a taupe color, round shape, size 21 in x 35 in, and a price lower than 60.00 dollars. First, I should search for dining room and living room furniture sets.</think>
<action>search[dining room furniture taupe round 21 in x 35 in price less than 60.00]</action>

System: [Back to Search] Page 1 (Total results: 50) [Next >]
[B07FKGQKZ1]: Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4" $44.99 to $33.98
[B09RQ22QJY]: Buffet Sideboard Kitchen Break Room Lunch Coffee Kitchenette Model 8004 BREAKTIME 2 pc Espresso – Factory Assembled (Furniture Items Purchase ONLY) $1997.5
[B09JC84P1M]: RUIXFLR Round End Table with Fabric Storage Basket, Marble Look Tray Side Table, Chic 2 Tier Coffee Table for Living Room Bedroom Office $121.99
[B07S8F7SHK]: Southern Enterprises Kempsey Convertible Console Dining Table, white $458.84
[B0854JTZMT]: BAYCHEER Tiffany Style Stained Glass Bowl Shade Pendant Light Chandelier Decorative Hanging Lamp Pendant Lighting Adjustable Ceiling Fixture with Pull Chain 3 Lights for Living Room Dining Room $169.99
[B09LHM4WQS]: Lift Top Coffee Tables for Living Room Coffee Table with Storage, Lift Cable Coffee Table, Dining Table with Hidden Storage Compartment Chestnut $82.99
[B07FVHPK5X]: wall26 Canvas Print Wall Art Window View of Yellow Poppies & Field Nature Wilderness Photography Realism Rustic Scenic Colorful Relax/Calm Ultra for Living Room, Bedroom, Office - 16"x24" $32.99
[B09PGQQQDL]: XLBHLH Black LED Chandelier Circular Dimmable 40W 1 Linear Aluminum Pendant Lighting Hanging Ceiling Light for Contemporary Dining Table Entry Kitchen Island $297.33
[B0943TLW3L]: Artificial Areca Palm Tree-Faux Tropical Fake Plant - Floor Dypsis Lutescens Silk Plants in Pot - 5 Ft Tall Fake Tree for Home Decor - Living Room, Patio, Office Indoor/Outdoor, K160 $99.99
[B088WSDHTW]: 4-Tier Ladder Bookshelf Organizer,Iron Open Bookcase Organizer (Black) $36.94

assistant: <think>After reviewing the search results, it appears that not all the items meet the criteria. Let's start by clicking on the 'B07FKGQKZ1' item to check if it's a dining room and living room furniture set, and if it's within the price range.</think>
<action>click[b07fkgqkz1]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige]
navy-multi taupe Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
Description [Features] [Reviews] [Buy Now]

assistant: <think>After checking the item with the size 21 in x 35 in which is a round area rug, it does not match the item shape and color criteria. I will now click 'next >' to see if there are other options available.</think>
<action>click[next >]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]

color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige]
navy-multi taupe Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

assistant: <think>After clicking 'next >', the page is still showing the area rug. Let's check the options available to see if there are any dining room and living room furniture sets that match the requirements. I will click 'next >' again to move to the next page.</think>
<action>click[next >]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige]
navy-multi taupe Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

assistant: <think>After checking the page with the area rug, it still does not meet the requirement of being a dining room and living room furniture set, round shape, size 21 in x 35 in, and a price lower than 60.00 dollars. I will click 'next >' to move to the next page.</think>
<action>click[next >]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige]
navy-multi taupe Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

assistant: <think>After checking the page with the area rug, it still does not meet the requirement of being a dining room and living room furniture set, round shape, size 21 in x 35 in, and a price lower than 60.00 dollars. I will click 'next >' again to move to the next page.</think>
<action>click[next >]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige]
navy-multi taupe Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

assistant: <think>After clicking 'next >' one more time, it still seems the page is showing the same area rug. Let's try to refine the search by clicking 'next >' one more time to see if we can get new options.</think>

<action>click[next >]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige]
navy-multi taupe Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

assistant: <think>After clicking 'next >' one more time, it still seems the page is showing the same area rug. It looks like the area rug is the only option that matches the criteria. Let's check the area rug to see if it is a dining room and living room furniture set, round shape, size 21 in x 35 in, and a price lower than 60.00 dollars.</think>
<action>click[next >]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige]
navy-multi taupe Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

assistant: <think>The area rug that is being shown is the Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4" with a price range of $44.99 to $33.98. The color is blue/gray, black/black. This matches the criteria of taupe color, but the item shape is a runner and not round. The size is 5'2"x7'4" which is not 21 in x 35 in. Therefore, it does not fully meet the criteria of round shape and size 21 in x 35 in.</think>
<action>click[buy now]</action>

System: Thank you for shopping with us! Reward: 0.5

The ReAct approach exhibits notable limitations in this shopping scenario, revealing several key inefficiencies. First, the method demonstrates repetitive behavior by cyclically viewing and interacting with the same area rug listing, indicating limited adaptive capabilities when confronted with suboptimal search results. Second, although the model identifies early that the available items failed to meet the specified criteria (taupe color, round shape, 21in x 35in dimensions, and under $60), it persists with an ineffective navigation pattern rather than employing alternative strategies such as search term refinement. Third, the model's final decision to purchase an area rug that deviated substantially from the requirements, failing to match the furniture set category, shape, size, and color specifications—suggests fundamental deficiencies in its decision-making framework when handling imperfect matches. This behavior indicates that ReAct lacks robust strategies for scenarios where exact requirement fulfillment is unattainable. The resultant low reward score (0.5 / 1.0) quantitatively validates these observed limitations in meeting user specifications.

D.8 Case Using GRPO in WebShop

GRPO Response in WebShop

Assistant: <action>search[taupe home office furniture round 21 in x 35 in price < 60.00]</action>

System: [Back to Search] Page 1 (Total results: 50) [Next >]
[B07FKGQKZ1]: Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4" $44.99 to $33.98
[B09JC84P1M]: RUIXFLR Round End Table with Fabric Storage Basket, Marble Look Tray Side Table, Chic 2 Tier Coffee Table for Living Room Bedroom Office $121.99
[B01N9RL0AX]: ClosetMaid 1312 4-Tier Wood Ladder Shelf Bookcase, Natural $73.99
[B07FVHPK5X]: wall26 Canvas Print Wall Art Window View of Yellow Poppies & Field Nature Wilderness Photography Realism Rustic Scenic Colorful Relax/Calm Ultra for Living Room, Bedroom, Office - 16"x24" $32.99
[B09RQ22QJY]: Buffet Sideboard Kitchen Break Room Lunch Coffee Kitchenette Model 8004 BREAKTIME 2 pc Espresso – Factory Assembled (Furniture Items Purchase ONLY) $1997.5
[B0784G74Z9]: Daonanba Durable Stable Coffee Table Teak Resin Handmade Unique Furniture Elegant Style Home Decoration 23.6" Round Tabletop $156.92
[B09688C4XM]: Lifestyle Furniture Sectional Recliner Sofa Set Living Room Reclining Couch with Drop Down Table (Single Piece or Combination) (GS2900, 3PCS) $1847.0
[B08GFNJN5R]: Byootique Classic Black Makeup Train Case Soft Sided Barber Cosmetic Backpack Organize Storage Carry on Travel with Side Pocket Removable Bag $79.9
[B073KYWX93]: MJL Furniture MAX Button Tufted Upholstered Square Blue Ottoman Black $268.99
[B085RBZCHH]: MEETWARM 16 inch Table Legs Wood Furniture Legs Tapered Round for Coffee End Tables Side Table Chair Mid-Century Modern DIY Furniture Leg Natural M8 Hanger Bolts, Set of 4 $23.99 to $21.59

Assistant: <action>click[b07fkgqkz1]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <action>click[round]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <action>click[taupe]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"

Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <action>click[taupe]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <action>click[taupe]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <action>click[buy now]</action>

System: Thank you for shopping with us! Reward: Your score (min 0.0, max 1.0) 0.83

GRPO demonstrates measurable improvements over ReAct in executing the shopping task, exhibiting enhanced search efficiency and strategic behavior. The method demonstrates a more sophisticated search strategy by constructing comprehensive initial queries that incorporated all critical parameters (taupe coloration, round configuration, dimensional constraints, and price limitations) simultaneously. Furthermore, GRPO displays superior utilization of the platform's filtering mechanisms, actively engaging with shape and color refinement options to optimize search results contrast to ReAct's cyclical interaction patterns. This enhanced performance is quantitatively reflected in a substantially higher reward score (0.83 compared to ReAct's 0.5). Nevertheless, GRPO's approach reveals certain limitations: the model exhibits perseverative behavior in repeatedly activating the 'taupe' filter without implementing alternative strategies when the initial approach proves suboptimal. Moreover, paralleling ReAct's shortcomings, GRPO ultimately selects a purchase (an area rug) that fails to satisfy all specified criteria, notably diverging from the core requirement of a comprehensive furniture set for dining and living spaces.

D.9 Case Using DyPO-Zero in WebShop

DyPO-Zero Response in WebShop

Assistant: <think>I need to start my search with the given query "home office furniture sets for dining room, living room with color: taupe, and item shape: round, and size: 21 in x 35 in, and price lower than $60."</think><action>search*[home office furniture sets for dining room, living room color: taupe shape: round size: 21 in x 35 in price: below $60]*</action>

System: [Back to Search] Page 1 (Total results: 50) [Next >]
[B07FKGQKZ1] Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4" $44.99 to $33.98

[B09JC84P1M] RUIXFLR Round End Table with Fabric Storage Basket, Marble Look Tray Side Table, Chic 2 Tier Coffee Table for Living Room Bedroom Office $121.99
[B0854JTZMT] BAYCHEER Tiffany Style Stained Glass Bowl Shade Pendant Light Chandelier Decorative Hanging Lamp Pendant Lighting Adjustable Ceiling Fixture with Pull Chain 3 Lights for Living Room Dining Room $169.99
[B0943TLW3L] Artificial Areca Palm Tree-Faux Tropical Fake Plant - Floor Dypsis Lutescens Silk Plants in Pot - 5 Ft Tall Fake Tree for Home Decor - Living Room, Patio, Office Indoor/Outdoor, K160 $99.99 [B09LHM4WQS] Lift Top Coffee Tables for Living Room Coffee Table with Storage, Lift Cable Coffee Table, Dining Table with Hidden Storage Compartment Chestnut $82.99
[B01N9RL0AX] ClosetMaid 1312 4-Tier Wood Ladder Shelf Bookcase, Natural $73.99
[B08LNFNXSD] Armchair Makeup Stool Lounge Stools Home Office Desk Chair Dining Chair Home Office Chair Floor Chair Computer Chair Ergonomic Learning Back Desk $381.81
[B09HXDPKJF] GXP Solid Reclaimed Wood Bedside Cabinet w/3 Drawers Nightstand Side Table $100.0
[B07FVHPK5X] wall26 Canvas Print Wall Art Window View of Yellow Poppies & Field Nature Wilderness Photography Realism Rustic Scenic Colorful Relax/Calm Ultra for Living Room, Bedroom, Office - 16"x24" $32.99
[B09688C4XM] Lifestyle Furniture Sectional Recliner Sofa Set Living Room Reclining Couch with Drop Down Table (Single Piece or Combination) (GS2900, 3PCS) $1847.0

Assistant: <think>I have received a search result with product information. The first product in the search results has the ASIN B07FKGQKZ1. I should start my search by clicking on this product to check its details.</think><action>click[B07FKGQKZ1]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi], [taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4" Price: $44.99 to $33.98 Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <think>I have been provided with a list of product sizes to search for. The size I'm looking for is '21 in x 35 in'. I need to start my search with this specific size.</think><action>click[21 in x 35 in]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi], [taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4" Price: $44.99 to $33.98 Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <think>I have selected the size '21 in x 35 in'. Next, I need to filter the search results to find products with a round shape.</think><action>click[round]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi], [taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4" Price: $44.99 to $33.98 Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <think>I have selected the size '21 in x 35 in' and filtered the search results to include products with a round shape. Next, I need to filter the search results to find products with the

color taupe.</think><action>click[taupe]</action>

System: [Back to Search] [< Prev]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi], [taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4" Price: $44.99 to $33.98 Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <think>I have selected the size '21 in x 35 in' and filtered the search results to include products with a round shape and color taupe. Next, I need to find products within the price range of less than $60.</think><action>click[buy now]</action>

System: Thank you for shopping with us! Reward: 1.0

DyPO-Zero exhibits superior performance characteristics compared to both ReAct and GRPO in the WebShop task, demonstrating significant advancements in search strategy and decision-making capabilities. The agent's most distinctive feature is its implementation of a systematic, hierarchical approach to filtering, incorporating explicit reasoning steps before each action execution. In contrast to ReAct's cyclical behavioral patterns and GRPO's repetitive filter interactions, DyPO-Zero employs a methodologically sound approach, sequentially applying filters in a logically structured manner while maintaining comprehensive feature selection accuracy. This enhanced strategic framework translates into quantitatively superior performance, as evidenced by DyPO-Zero achieving the maximum reward score (1.0), substantially outperforming both ReAct and GRPO.

D.10 Case Using DyPO in WebShop

DyPO Response in WebShop

Assistant: <action>search[taupe dining room living room furniture round 21x35 size 80]</action>

System: [Back to Search] Page 1 (Total results: 50) [Next >]
[B07FKGQKZ1]: Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4" $44.99 to $33.98
[B09JC84P1M]: RUIXFLR Round End Table with Fabric Storage Basket, Marble Look Tray Side Table, Chic 2 Tier Coffee Table for Living Room Bedroom Office $121.99
[B0943TLW3L]: Artificial Areca Palm Tree-Faux Tropical Fake Plant - Floor Dypsis Lutescens Silk Plants in Pot - 5 Ft Tall Fake Tree for Home Decor - Living Room, Patio, Office Indoor/Outdoor, K160 $99.99
[B07S8F7SHK]: Southern Enterprises Kempsey Convertible Console Dining Table, white $458.84
[B0854JTZMT]: BAYCHEER Tiffany Style Stained Glass Bowl Shade Pendant Light Chandelier Decorative Hanging Lamp Pendant Lighting Adjustable Ceiling Fixture with Pull Chain 3 Lights for Living Room Dining Room $169.99
[B09LHM4WQS]: Lift Top Coffee Tables for Living Room Coffee Table with Storage, Lift Cable Coffee Table, Dining Table with Hidden Storage Compartment Chestnut $82.99
[B0957XW92M]: Giantex Faux Fur Vanity Stool Chair, Round Footstool Ottoman with Metal Legs for Living Room, Fluffy Chair with Furry Padded Seat, Makeup Stool for Bedroom, Decorative Furniture Footrest White $52.99
[B09GBF12CY]: Corner Floor Lamp, Smart RBG Floor Lamp with App Control, Timer, Magiacous Music Sync Modern Floor Lamp, Full Spectrum Color Changing Standing Lamp, for Living Room, Bedroom, Game, Party, Reading $69.99
[B009EEVDSQ]: Queen Size Tan Traditional Japanese Floor Futon Mattresses, Foldable Cushion Mats, Yoga, Meditaion 60" Wide X 80" Long $149.0
[B09PGQQQDL]: XLBHLH Black LED Chandelier Circular Dimmable 40W 1 Linear Aluminum Pendant Lighting Hanging Ceiling Light for Contemporary Dining Table Entry Kitchen Island [SEP] $297.33

Assistant: <action>click[b07fkgqkz1]</action>

System: [Back to Search] Page 1 (Total results: 50) [Next >]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <action>click[round]</action>

System: [Back to Search] Page 1 (Total results: 50) [Next >]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <action>click[taupe]</action>

System: [Back to Search] Page 1 (Total results: 50) [Next >]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <action>click[taupe]</action>

System: [Back to Search] Page 1 (Total results: 50) [Next >]
size: [1 ft 9 in x 7 ft 2 in], [3 ft 7 in x 5 ft 2 in], [5 ft 2 in x 7 ft 4 in], [5 ft 3 in x 7 ft 5 in], [7 ft 8 in], [7 ft 8 in x 10 ft 7 in], [7 ft 9 in x 10 ft 8 in], [9 ft 2 in x 12 ft 5 in], [21 in x 35 in]
item shape: [runner], [rectangular], [round]
color: [black-gray], [blue | gray], [brown | black], [blue], [navy | red], [taupe | orange], [black | gray], [brown | beige], [navy-multi taupe]
Home Dynamix Lyndhurst Rotana Modern Area Rug, Contemporary Blue/Gray/Ivory 5'2"x7'4"
Price: $44.99 to $33.98
Rating: N.A.
[Description] [Features] [Reviews] [Buy Now]

Assistant: <action>click[buy now]</action>

System: Thank you for shopping with us! Reward: 0.83

The DyPO model, originating from a GRPO-trained checkpoint, exhibits reasoning capabilities with notable similarities to GRPO in their fundamental characteristics. Both approaches demonstrate a persistent limitation in action selection, manifesting as repetitive behavior patterns. This observation highlights a critical constraint in GRPO's architecture: while it shows enhanced capability in processing and interpreting single-turn interactions, its capacity for integrating information across multiple sequential interactions is suboptimal. The model's difficulty in maintaining and utilizing a comprehensive understanding of the interaction history indicates limitations in its multi-turn reasoning capabilities. Nevertheless, the DyPO-trained model demonstrates measurable improvements in mitigating cyclical interaction patterns, achieving greater operational efficiency compared to its GRPO counterpart. This indicates that while DyPO inherits certain limitations from its GRPO foundation, it successfully addresses some of the base model's behavioral inefficiencies.