Training Task Reasoning LLM Agents for Multi-turn Task Planning via Single-turn Reinforcement Learning

Hu, Hanjiang; Liu, Changliu; Li, Na; Wang, Yebin

TR2026-026   March 03, 2026

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in knowledge acquisition, reasoning, and tool use, making them promising candidates for au- tonomous agent applications. However, training LLM agents for complex multi-turn task planning faces significant challenges, including sparse episode-wise rewards, credit assignment across long horizons, and the computational overhead of reinforcement learning in multi-turn interaction settings. To this end, this paper introduces a novel approach that transforms multi- turn task planning into single-turn task reasoning problems, enabling efficient policy optimization through Group Relative Policy Optimization (GRPO) with dense and verifiable reward from expert trajectories. Our theoretical analysis shows that GRPO improvement on single-turn task reasoning results in a lower bound of the multi-turn success probability under the minimal turns, as well as the generalization to subtasks with shorter horizons. Experimental evaluation on the complex task planning benchmark demonstrates that our 1.5B parameter model trained with single-turn GRPO achieves superior performance compared to larger baseline models up to 14B parameters, with success rates of 70% for long-horizon planning tasks.

IEEE Control Systems Letters 2026

Training Task Reasoning LLM Agents for Multi-turn Task Planning via Single-turn Reinforcement Learning

Hanjiang Hu, Changliu Liu, Na Li, Yebin Wang

Abstract—Large Language Models (LLMs) have demonstrated remarkable capabilities in knowledge acquisition, reasoning, and tool use, making them promising candidates for autonomous agent applications. However, training LLM agents for complex multi-turn task planning faces significant challenges, including sparse episode-wise rewards, credit assignment across long horizons, and the computational overhead of reinforcement learning in multi-turn interaction settings. To this end, this paper introduces a novel approach that transforms multi-turn task planning into single-turn task reasoning problems, enabling efficient policy optimization through Group Relative Policy Optimization (GRPO) with dense and verifiable reward from expert trajectories. Our theoretical analysis shows that GRPO improvement on single-turn task reasoning results in a lower bound of the multi-turn success probability under the minimal turns, as well as the generalization to subtasks with shorter horizons. Experimental evaluation on the complex task planning benchmark demonstrates that our 1.5B parameter model trained with single-turn GRPO achieves superior performance compared to larger baseline models up to 14B parameters, with success rates of 70% for long-horizon planning tasks.

I. INTRODUCTION

Large Language Models (LLMs) as autonomous agents are important in modern AI-based systems, which can perceive environments, reason about plans, and execute actions to interact with the environments [1]. On top of the strong capabilities of LLMs, prompt-based agentic frameworks [2] are proposed by integrating observation for environment state, reasoning based on augmented LLM with tools and memory, and action execution that interacts with the environment through structured interfaces as a series of *single-turn* interactions with the environments. However, effort-costly prompt engineering is inevitable to build the LLM-based agent, and it is also computationally expensive for test-time scaling in the *multi-turn* interaction with the environment [3].

Therefore, training LLM agents through reinforcement learning (RL) for complex multi-turn task planning becomes a promising way to build effective agentic systems with low test-time cost [4], [5]. However, current RL approaches face critical challenges when applied to multi-turn interactions with the environment for LLMs. The reward of task completion is too sparse for effective exploitation of LLM policy in the multi-turn settings. Besides, the credit assignment becomes particularly difficult in long-horizon

planning problems, where determining which specific actions contribute to the eventual reward of success or failure is nontrivial [5]. The computational complexity of multi-turn RL scales poorly with sequence length, making it impractical for training on complex tasks that may require dozens of sequential decision turns [6].

In contrast, recent RL post-training successfully boosts LLM reasoning capability, particularly in math and coding domains based on on-policy methods, e.g., Group Relative Policy Optimization (GRPO) [7]. However, these successes have primarily focused on single-turn reasoning tasks with verifiable reward, where the model generates a complete solution in one step and receives the reward directly based on the correctness of the final answer [8]. The fundamental challenge lies in bridging the gap between single-turn reasoning capabilities and multi-turn task planning requirements, where LLM agents must be coherent along the interaction sequences while adapting to changing environmental states.

To this end, we aim to bridge the gap based on the insight that complex multi-turn task planning can be decomposed into a sequence of single-turn task reasoning problems, each requiring the agent to select the optimal next action given the current state. This decomposition enables us to leverage the provably effective and efficient RL post-training with dense and verifiable reward based on expert trajectory, with theoretical guarantees of the multi-turn decision-making capabilities for long-horizon planning. We compare and validate the effectiveness and generalizability of the proposed method on the challenging long-horizon task planning benchmark. The contributions are summarized below.

- We provide a theoretical analysis bridging single-turn task reasoning and multi-turn task completion, proving that the success probability after GRPO can be lower-bounded in multi-turn planning scenarios.
- We show empirical results on multiple tasks from the challenging ROBOTOUILLE benchmark, where our 1.5B model trained with single-turn GRPO consistently outperforms 14B baselines with fewer completion steps.
- We both theoretically and empirically show the strong cross-task generalization capabilities of the GRPO-optimized LLM agents, where models trained on complex tasks successfully transfer to simpler subtasks.

II. PROBLEM FORMULATION

We formulate the multi-turn task planning problem as two interconnected Markov Decision Processes (MDPs): a multi-turn MDP representing the complete long-horizon task

H. Hu and C. Liu are with the Robotics Institute, Carnegie Mellon University. N. Li is with School of Engineering and Applied Sciences, Harvard University. This work was done while H. Hu was a research intern and N. Li was a visiting scholar at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA. (email: `hanjianghu@cmu.edu`).
Y. Wang is with the MERL. (email: `yebinwang@ieee.org`).

planning and a single-turn MDP based on expert trajectories for training purposes.

A. Multi-Turn Task Planning MDP

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, f, R, \mathcal{T}, s_0)$ be a finite-horizon MDP representing the multi-turn task planning problem, where $\mathcal{S}$ is the token state space representing the current task context and environment observation; $\mathcal{A}$ is the token action space of discrete possible single-step actions; $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is the deterministic but unknown token state transition function; $R : \mathcal{S} \times \mathcal{A} \rightarrow \{0, 1\}$ is the binary sparse reward function indicating whether the task is completed; $\mathcal{T} \subset \mathbb{N}$ is the finite time horizon; $s_0 \in \mathcal{S}$ is the initial state. Let $\pi : \mathcal{S} \rightarrow \mathcal{A}$ be a stochastic policy mapping token states to token action distributions. $\pi(a|s)$ denotes the probability of $a = \pi(s)$. A trajectory $\tau = [(s_0, a_0), (s_1, a_1), \ldots, (s_T, a_T)]$ is generated by following policy π for T steps, where $s_{i+1} = f(s_i, a_i)$. Then we define the successful trajectory induced by a stochastic policy as follows.

Definition 2.1 (Successful trajectory by stochastic policy): Given the MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, f, R, \mathcal{T}, s_0)$, a successful trajectory starting with $s_0^s = s_0$ is defined as:

$$\tau^s(T) = [(s_0^s, a_0^s), \ldots, (s_T^s, a_T^s)], \ R(s_T^s, a_T^s) = 1, \quad (1)$$

where $s_{i+1}^s = f(s_i^s, a_i^s)$ for $i = 0, 1, \ldots, T-1$; $a_i^s \sim \pi(\cdot|s_i^s)$ for $i = 0, 1, 2, \ldots, T$ given the stochastic policy π.

B. Expert Trajectory with Unique Optimality

On top of Definition 2.1, we further define the expert trajectory generated by the expert policy with the unique optimality of minimal trajectory length. We adopt superscript GT to denote "Ground Truth" for expert policy or trajectory.

Definition 2.2 (Expert trajectory by stochastic policy): Given the MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, f, R, \mathcal{T}, s_0)$, a trajectory $\tau^{GT}(T^{GT})$ from the stochastic policy π is defined as an expert trajectory if the following conditions hold:

1) $\tau^{GT}(T^{GT})$ is a successful trajectory generated from stochastic policy the π in Definition 2.1.
2) For any successful trajectory $\tau(T)$ generated by stochastic policy π with $R(s_T, a_T) = 1$, we have $T \geq T^{GT}$ and the equality holds if and only if $\tau = \tau^{GT}$.

Based on the expert trajectory τ^{GT} in Definition 2.2, we can further define the expert policy π^{GT} as follows.

Definition 2.3 (Expert policy from expert trajectory): Given the MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, f, R, \mathcal{T}, s_0)$ and an expert trajectory $\tau^{GT}(T^{GT})$, a policy $\pi^{GT} : \mathcal{S} \rightarrow \mathcal{A}$ is defined as the expert policy if it holds that

$$a_i^{GT} = \pi^{GT}(s_i^{GT}), \ \forall(s_i^{GT}, a_i^{GT}) \in \tau^{GT}(T^{GT}). \quad (2)$$

Remark 2.1: Even though expert trajectory $\tau^{GT}(T^{GT})$ is uniquely optimal with minimal length of turns T^{GT}, the induced expert policy π^{GT} is not necessarily unique or deterministic, as there is no state-action constraint beyond expert trajectories $\tau^{GT}(T^{GT})$. The unique optimal expert trajectory assumption ensures a well-defined single-step reward and serves as the foundation for the following results. We ensure

this assumption in the experiment and further conduct an ablation comparison of robustness under non-expert noisy training trajectories.

C. Single-Turn Task Reasoning MDP

Even though multi-turn task planning MDP $\mathcal{M}$ is the conventional one in the control and RL community, it is challenging and non-trivial to directly train an LLM policy for $\mathcal{M}$ due to intractable computational overhead with exponentially increasing token length of trajectory rollouts in the multi-turn setting [5], [6]. Therefore, to enable efficient LLM policy learning , we construct an additional single-turn MDP $\mathcal{M}_S = (\mathcal{S}, \mathcal{A}, \emptyset, r_{\pi^{GT}}, 1, s_0)$ based on the expert trajectory $\tau^{GT}(T^{GT})$ in Definition 2.2 and expert policy π^{GT} in Definition 2.3. This single-turn MDP can be viewed as a bandit problem as follows. $\mathcal{S}$ is the same token state space as in $\mathcal{M}$, $\mathcal{A}$ is the same token action space as in $\mathcal{M}$, $r_{\pi^{GT}} : \mathcal{S} \times \mathcal{A} \rightarrow \{0, 1\}$ is the binary reward function. There is no state transition function, which is denoted as $\emptyset$. The horizon 1 is one-step in the single-turn setting. Initial state is the same as in $\mathcal{M}$.

Definition 2.4 (Reward function from expert trajectory): The binary reward function $r_{\pi^{GT}}$ for single-turn MDP $\mathcal{M}_S = (\mathcal{S}, \mathcal{A}, \emptyset, r_{\pi^{GT}}, 1, s^{GT})$ is defined as:

$$r_{\pi^{GT}}(s, a) = \mathbb{1}\{a = \pi^{GT}(s)\} \in \{0, 1\}, \quad (3)$$

where π^{GT} is the expert policy defined in Definition 2.3.

III. METHODOLOGY

We first give the theoretical analysis of policy learning for single-turn MDP $\mathcal{M}_S$ based on dense reward in Definition 2.4. Then we define the success probability with the minimal turns for task planning MDP $\mathcal{M}$ and show that policy improvement from $\mathcal{M}_S$ can give a lower bound of success probability in $\mathcal{M}$, for even unseen task planning.

A. GRPO for Single-Turn Task Reasoning

GRPO is an on-policy model-free RL method that optimizes policies using verifiable rewards and replaces critic model with a group of multiple responses for a prompt to calculate a group-based advantage [7]. Based on the theoretical foundation of GRPO [8], we give the policy iteration and the success amplification in the single-turn MDP $\mathcal{M}_S$ defined with Definition 2.4.

1) GRPO Policy Iteration Dynamics: Let ρ_Q be a distribution over state query prompts $s \in \mathcal{S} \sim \rho_Q$. For the single-turn MDP $\mathcal{M}_S$, we consider the GRPO optimization problem which adopts the relative advantage within a group of policy outputs as the critic while maintaining proximity to a reference policy through KL regularization. The GRPO objective without clipping is formulated as:

$$\max_{\pi} \mathbb{E}_{s \sim \rho_Q} \mathbb{E}_{a \sim \pi_{\text{old}}(\cdot|s)} \frac{\pi(a|s)}{\pi_{\text{old}}(a|s)} A(s, a) - \beta \text{KL}(\pi || \pi^{\text{ref}}),$$

where the advantage function A based on the binary reward in Definition 2.4 is given by:

$$A_{\pi^{GT}}(s, a) = \frac{r_{\pi^{GT}}(s, a) - \mathbb{E}_{a' \sim \pi_{\text{old}}(\cdot|s)} r_{\pi^{GT}}(s, a')}{\sqrt{\text{Var}_{a' \sim \pi_{\text{old}}(\cdot|s)} r_{\pi^{GT}}(s, a')}}. \quad (4)$$

At iteration $n \geq 1$, the policy iteration of GRPO for $\mathcal{M}_S$ is,

$$\pi_n = \arg\max_{\pi} \mathbb{E}_{s \sim \rho_Q} \left\{ \mathbb{E}_{a \sim \pi_{n-1}(\cdot|s)} \left[\frac{\pi_n(a|s)}{\pi_{n-1}(a|s)} \cdot \right. \right. \tag{5}$$
$$\left. \left. \frac{r_{\pi^{GT}}(s,a) - \mathbb{E}_{a' \sim \pi_{n-1}(\cdot|s)} r_{\pi^{GT}}(s,a')}{\sqrt{\mathrm{Var}_{a' \sim \pi_{n-1}(\cdot|s)} r_{\pi^{GT}}(s,a')}} \right] - \beta \mathrm{KL}(\pi_n || \pi^{\mathrm{ref}}) \right\},$$

where initial policy is the reference policy $\pi_0 = \pi^{\mathrm{ref}}$.

2) GRPO Single-Turn Optimality and Success Amplification: Following the Theorem 3.1 from [8] below, we can have the optimality guarantee for GRPO iteration in Equation (5) in our single-turn task reasoning setting.

Theorem 3.1 (Theorem 3 from [8]): Let π^* be the fixed point of GRPO-optimized policy obtained from the policy iteration dynamics of Equation (5) and π^{ref} be the initial reference policy. Denote the success probabilities of the reference policy as $p^{\mathrm{ref}} = \mathbb{E}_{a \sim \pi^{\mathrm{ref}}(\cdot|s^{GT})}[\mathbb{1}\{a = \pi^{GT}(s^{GT})\}]$ and optimized policy as $p^* = \mathbb{E}_{a \sim \pi^*(\cdot|s^{GT})}[\mathbb{1}\{a = \pi^{GT}(s^{GT})\}]$, respectively. It holds that $p^* > p^{\mathrm{ref}}$.

Proof: The optimal policy at iteration n of the GRPO dynamics in Equation (5) satisfies $\pi_n(a|s) \propto \pi_{\mathrm{ref}}(a|s) \exp\left(\frac{1}{\beta} A_{\pi^{GT}}(s,a)\right)$, where the advantage function $A_{\pi^{GT}}$ upweights successful actions ($r_{\pi^{GT}}(s,a) = 1$) and downweights failures ($r_{\pi^{GT}}(s,a) = 0$). The variance normalization in $A_{\pi^{GT}}$ ensures positive advantages for successes and negative advantages for failures. At the fixed point π^*, this means $\exp(\frac{1}{\beta} A_{\pi^{GT}}) > 1$ for correct actions and $\exp(\frac{1}{\beta} A_{\pi^{GT}}) < 1$ for incorrect ones, systematically shifting probability mass toward successful outputs and yielding $p^* > p_{\mathrm{ref}}$. Check Appendix C.1 of [8] for full proof. ∎

Corollary 3.1 (GRPO single-Turn optimality): For the fixed point of GRPO-optimized policy π^* and the reference policy π^{ref}, the following holds for any state s from the expert trajectory τ^{GT},

$$\mathbb{E}_{a^* \sim \pi^*(\cdot|s)}[r_{\pi^{GT}}(s,a^*)] \geq \mathbb{E}_{a^{\mathrm{ref}} \sim \pi^{\mathrm{ref}}(\cdot|s)}[r_{\pi^{GT}}(s,a^{\mathrm{ref}})]. \tag{6}$$

Proof: It holds based on the single-turn reward formulation in Definition 2.4 based on Theorem 3.1. ∎

Corollary 3.1 establishes that GRPO training on the single-turn MDP $\mathcal{M}_S$ with expert trajectories leads to policies that achieve higher success rates than the reference policy before GRPO training, providing the foundation for analyzing multi-turn performance.

B. Success Probability for Multi-Turn Task Planning

To analyze how GRPO improvements on the single-turn MDP $\mathcal{M}_S$ can influence the performance in the multi-turn MDP $\mathcal{M}$, we introduce the success probability with the minimal turns that captures the probability of achieving the optimal expert trajectory, by assuming the existence of minimal completion steps given any state below.

Definition 3.1 (Minimal turns for task completion): For any state $s \in \mathcal{S}$ in the multi-turn MDP $\mathcal{M}$, denote the unique minimal length of turns as $T^*(s) \geq 0$ among all successful trajectories defined in Definition 2.1 from s, i.e., $T^*(s) := \min\{T \geq 0 : \exists \tau(T) \text{ s.t. } R(s_T, a_T) = 1, s_0 = s\}$.

$T^*(s)$ exists uniquely based on the existence and uniqueness of the expert trajectory in Definition 2.2 by taking any states as initial states. We then define the probability of the successful trajectory with minimal steps.

Definition 3.2 (Success probability with minimal steps): With $T^*(s)$ in Definition 3.1, the probability of achieving the task planning goal in the minimal number of steps starting from state s_t at time t under policy π in the multi-turn MDP $\mathcal{M}$ is defined as,

$$P_t^\pi(s_t) = \mathbb{P}_\pi\left(R(s_{t+T^*(s_t)}, a_{t+T^*(s_t)}) = 1 | s_t\right). \tag{7}$$

We then present the recursion condition for the success probability in Definition 3.2 based on the reward function from the expert trajectory in Definition 2.4.

Theorem 3.2 (Recursion equation of success probability): The success probability defined in Definition 3.2 satisfies the following recursion condition,

$$P_t^\pi(s_t) = \mathbb{E}_{a \sim \pi(\cdot|s_t)}[r_{\pi^{GT}}(s_t, a) \cdot P_{t+1}^\pi(s_{t+1})], \tag{8}$$

where π^{GT} is the expert policy in Definition 2.3 by expert trajectory with $P_{t+T^*(s_t)}^\pi(s_{t+T^*(s_t)}) = 1$ for any state s_t.

Proof: For any state s_t as the initial state, based on Definition 2.3, π^{GT} is induced by the expert trajectory $\tau^{GT} = [(s_t, a_t), \ldots, (s_{T^{GT}}, a_{T^{GT}})], s_{k+1} = f(s_k, \pi^{GT}(s_k)), k = t, t+1, \ldots, T^{GT} - 1$ with minimal turns of task completion. By $P_{t+T^*(s_t)}^\pi(s_{t+T^*(s_t)}) = 1$, we have $R(s_{t+T^*(s_t)}, a_{t+T^*(s_t)}) = 1$. Combining the uniqueness of expert trajectory in Definition 2.2 and minimal completion steps $T^*(s)$ in Definition 3.1, it holds that

$$T^*(s_{T^{GT}}) = 0, T^{GT} = T^*(s_k) + k, k = t, \ldots, T^{GT}. \tag{9}$$

For any state that deviates from the expert trajectory $s'_{k+1} \neq f(s_k, \pi^{GT}(s_k))$, by the minimal completion turns,

$$T^*(s'_{k+1}) > T^*(s_k) - 1, \forall k = t, t+1, \ldots, T^{GT} - 1. \tag{10}$$

By the law of total expectation over the next action $t+1$:

$$P_t^\pi(s_t) = \mathbb{P}_\pi(R(s_{t+T^*(s_t)}, a_{t+T^*(s_t)}) = 1 | s_t) \tag{11}$$
$$= \mathbb{E}_{a \sim \pi(\cdot|s_t)}[\mathbb{P}_\pi(R(s_{t+T^*(s_t)}, a_{t+T^*(s_t)}) = 1 | s_t, a)].$$

For $a \sim \pi(\cdot|s_t)$, if $a \neq \pi^{GT}(s_t)$, i.e., $r_{\pi^{GT}}(s_t, a) = 0$, we have $s'_{t+1} \neq f(s_t, \pi(s_t))$. Then based on Equation (10), it takes strictly more than $T^*(s_t) - 1$ steps for completion, i.e.,

$$\mathbb{P}_\pi(R(s_{t+1+T^*(s_t)-1}, a_{t+1+T^*(s_t)-1}) = 1 | s'_{t+1}) = 0. \tag{12}$$

If $a = \pi^{GT}(s_t)$, i.e., $r_{\pi^{GT}}(s_t, a) = 1, s_{t+1} = f(s_t, a)$, then based on Equation (9), we have

$$\mathbb{P}_\pi(R(s_{t+T^*(s_t)}, a_{t+T^*(s_t)}) = 1 | s_t, a) \tag{13}$$
$$= \mathbb{P}_\pi(R(s_{t+1+T^*(s_{t+1})}, a_{t+1+T^*(s_{t+1})}) = 1 | s_{t+1})$$
$$= P_{t+1}^\pi(s_{t+1}).$$

With the values of $r_{\pi^{GT}}(s_t, a) \in \{0, 1\}$ and combining Equations (11) to (13), it holds that

$$P_t^\pi(s_t) = \mathbb{E}_{a \sim \pi(\cdot|s_t)}[r_{\pi^{GT}}(s_t, a) \cdot P_{t+1}^\pi(s_{t+1})], \tag{14}$$

which concludes the proof. ∎

Please note that the recursion equation in Theorem 3.2 is not the same as Bellman recursion in value function iteration, and the key difference lies in that the multiplication of the binary single-turn reward $r_{\pi^{GT}}(s_t, a)$ is involved in the recursion of expectation, instead of addition. This recursion bridges the final multi-turn success in Equation (7) with single-turn reward improvements at each step in Equation (6), which is crucial to the following analysis on how single-turn GRPO gives a lower bound of success probability of the multi-turn task planning.

C. Lower Bound of Multi-Turn Success Probability

We now analyze how GRPO improvements on single-turn task reasoning in $\mathcal{M}_S$ will give a lower bound of success probability in the multi-turn task planning MDP $\mathcal{M}$. First, we assume the bounded Total Variation (TV) distance of actions between the proximal states in the training and test data. Then we show the lower bound of single-turn GRPO performance as Lemma 3.1.

Assumption 3.1 (Bounded Policy with State Proximity): For any execution state $s \in \mathcal{S}_{\text{exec}}$ encountered during policy deployment, there exists a training state $s' \in \mathcal{S}_{\text{train}}$ with the same expert action $\pi^{GT}(s') = \pi^{GT}(s)$ such that for any policy π, $\|\pi(\cdot|s) - \pi(\cdot|s')\|_{\text{TV}} \leq \delta$, where $\delta > 0$ is a bounded constant and $\|\cdot\|_{\text{TV}}$ denotes the TV distance.

Lemma 3.1 (Lower bound of GRPO policy deployment): Under Assumption 3.1, for any execution state $s \in \mathcal{S}_{\text{exec}}$, $\mathbb{E}_{a\sim\pi^*(\cdot|s)}[r_{\pi^{GT}}(s, a)] \geq \mathbb{E}_{a\sim\pi^{\text{ref}}(\cdot|s)}[r_{\pi^{GT}}(s, a)] - 4\delta$, where π^* is the GRPO-optimized policy and π^{ref} is the reference policy.

Proof: Let $s \in \mathcal{S}_{\text{exec}}$. By Assumption 3.1, there exists $s' \in \mathcal{S}_{\text{train}}$ such that $\|\pi^*(\cdot|s) - \pi^*(\cdot|s')\|_{\text{TV}} \leq \delta$, s$\|\pi^{\text{ref}}(\cdot|s) - \pi^{\text{ref}}(\cdot|s')\|_{\text{TV}} \leq \delta$, and $\pi^{GT}(s') = \pi^{GT}(s)$. With $r_{\pi^{GT}}(s, a) = r_{\pi^{GT}}(s', a) \in \{0, 1\}$, for any policy π and states s, s' with $\|\pi(\cdot|s) - \pi(\cdot|s')\|_{\text{TV}} \leq \delta$, we have:

$$\left|\mathbb{E}_{a\sim\pi(\cdot|s)}[r_{\pi^{GT}}(s, a)] - \mathbb{E}_{a\sim\pi(\cdot|s')}[r_{\pi^{GT}}(s', a)]\right| \quad (15)$$
$$\leq \sum_{a\in\mathcal{A}} |\pi(a|s) - \pi(a|s')| \cdot |r_{\pi^{GT}}(s, a)|$$
$$\leq \sum_{a\in\mathcal{A}} |\pi(a|s) - \pi(a|s')| = 2\|\pi(\cdot|s) - \pi(\cdot|s')\|_{\text{TV}} \leq 2\delta.$$

By Corollary 3.1, since $s' \in \mathcal{S}_{\text{train}}$:

$$\mathbb{E}_{a\sim\pi^*(\cdot|s')}[r_{\pi^{GT}}(s', a)] \geq \mathbb{E}_{a\sim\pi^{\text{ref}}(\cdot|s')}[r_{\pi^{GT}}(s', a)]. \quad (16)$$

Applying (15) to π^{ref}, π^* and combining (16) will complete the proof. ∎

Finally, the step-dependent lower bound of success probability is shown below, which becomes tighter as the time step is closer to the minimal length of turns T^{GT} of expert trajectories. It is consistent with the intuition that the task is more likely to complete from the middle step compared to initial step.

Theorem 3.3 (Lower bound of GRPO success probability): Let π^* be the fixed point of the GRPO-optimized policy from the policy iteration dynamics of Equation (5) and π^{ref}

be the initial reference policy. Under Assumption 3.1 and Equation (6), it holds for any given state s_t at step t that,

$$P_t^{\pi^*}(s_t) \geq P_t^{\pi^{\text{ref}}}(s_t) - (T^{GT} - t) \cdot 4\delta, \quad (17)$$

Proof: By the existence and uniqueness of the expert trajectory in Definition 2.2, denote the expert trajectory starting from s_t as $\tau^{GT} = [(s_t, a_t), \ldots, (s_{T^{GT}}, a_{T^{GT}})]$ with minimal turns of task completion $T^*(s_t) = T^{GT} - t$ in Definition 3.1. We prove by backward induction on the length of remaining turns to task completion. In the final case when $t = T^{GT}$, we have $P_t^{\pi^*}(s_t) = P_t^{\pi^{\text{ref}}}(s_t) = 1$ since the task is already completed. Now for the inductive case at step $t + 1, t + 1 \leq T^{GT}$, we assume $P_{t+1}^{\pi^*}(s') \geq P_{t+1}^{\pi^{\text{ref}}}(s') - (T^{GT} - t - 1) \cdot 4\delta$ for all $s' \in \mathcal{S}$. Then for any state s at step t, by Theorem 3.2:

$$P_t^{\pi^*}(s) = \mathbb{E}_{a\sim\pi^*(\cdot|s)}[r_{\pi^{GT}}(s, a) \cdot P_{t+1}^{\pi^*}(f(s, a))] \quad (18)$$

$$\geq \mathbb{E}_{a\sim\pi^*}[r_{\pi^{GT}}(s, a)\{P_{t+1}^{\pi^{\text{ref}}}(f(s, a)) - (T^{GT} - t - 1)4\delta\}] \quad (19)$$

$$= \mathbb{E}_{a\sim\pi^*(\cdot|s)}[r_{\pi^{GT}}(s, a) \cdot P_{t+1}^{\pi^{\text{ref}}}(f(s, a))]$$
$$\quad - (T^{GT} - t - 1) \cdot 4\delta \cdot \mathbb{E}_{a\sim\pi^*(\cdot|s)}[r_{\pi^{GT}}(s, a)] \quad (20)$$

$$\geq \mathbb{E}_{a\sim\pi^*(\cdot|s)}[r_{\pi^{GT}}(s, a) \cdot P_{t+1}^{\pi^{\text{ref}}}(f(s, a))]$$
$$\quad - (T^{GT} - t - 1) \cdot 4\delta \quad (21)$$

$$\geq \mathbb{E}_{a\sim\pi^{\text{ref}}(\cdot|s)}[r_{\pi^{GT}}(s, a) \cdot P_{t+1}^{\pi^{\text{ref}}}(f(s, a))]$$
$$\quad - 4\delta - (T^{GT} - t - 1) \cdot 4\delta \quad (22)$$

$$= P_t^{\pi^{\text{ref}}}(s) - (T^{GT} - t) \cdot 4\delta, \quad (23)$$

where (19) applies the inductive hypothesis, (21) uses $\mathbb{E}_{a\sim\pi^*(\cdot|s)}[r_{\pi^{GT}}(s, a)] \leq 1$, (22) applies Lemma 3.1 (noting that since $0 \leq P_{t+1}^{\pi^{\text{ref}}}(s') \leq 1$, the weighted expectation difference is bounded by the marginal difference), and (23) uses Theorem 3.2 for π^{ref}. ∎

Furthermore, we show that the policy trained with expert trajectories with single-turn GRPO can also generalize to subtask defined below.

Definition 3.3 (Subtask and its success probability): Given the expert trajectory $\tau^{GT}(T^{GT}) = [(s_0, a_0), \ldots, (s_{T^{GT}}, a_{T^{GT}})]$ in Definition 2.2 for the full task, a subtask is defined by a completion step $k^* < T^{GT}$ where the subtask is completed upon reaching (s_{k^*}, a_{k^*}), i.e., $R_{\text{sub}}(s_{k^*}, a_{k^*}) = 1$. The success probability with the fewest turns for subtask at $s_t, t \leq k^*$ is defined as,

$$P_t^{\text{sub},\pi}(s_t) = \mathbb{P}_{\pi}\left(R_{\text{sub}}(s_{k^*}, a_{k^*}) = 1 | s_t\right), \quad (24)$$

which is the probability of achieving subtask goal in the minimal number of steps starting from state s_t at time t under policy π.

Based on the success probability of subtask, we present the following corollary of Theorem 3.3.

Corollary 3.2 (GRPO generalization to subtasks): For any subtask defined by completion step $k^* < T^{GT}$ in Definition 3.3, the GRPO-optimized policy π^* achieves better performance on the subtask than the reference policy π^{ref} for any given state s_t at step $t \leq k^*$, i.e., $P_t^{\text{sub},\pi^*}(s_t) \geq P_t^{\text{sub},\pi^{\text{ref}}}(s_t) - (k^* - t) \cdot 4\delta$.

Proof: Since the subtask uses the same expert trajectory prefix as the full task, Lemma 3.1 also holds for subtask goal with the same single-turn reward function $r_{\pi^{GT}}(s,a)$. Besides, the recursion condition of $P_t^{\text{sub},\pi}(s_t)$ also applies. Therefore, following the proof of Theorem 3.3, by letting $T^{GT} = k^*$, π^* achieves better success probability than π^{ref} over $P_t^{\text{sub}}(s_t)$ for any given state s_t at step $t \le k^*$. ∎

IV. EXPERIMENTS

In the experiments, we aim to answer the following two questions empirically: *1) How is the performance of multi-turn task planning after GRPO over single-turn task reasoning? 2) Are GRPO-trained agents generalizable to unseen planning tasks and robust to noisy non-expert training trajectories?* We answer the first question in Section IV-B and the second one in Section IV-C. Prior to the answers, we first introduce the experimental setup of environment, model training and evaluation metrics.

A. Experimental Setup

a) Environment and Tasks: We evaluate our approach on the ROBOTOUILLE benchmark [9], a challenging multi-turn task planning environment designed for testing LLM agents' reasoning capabilities in cooking scenarios. ROBOTOUILLE provides a deterministic long-horizon complex environment where agents must coordinate multiple actions across different kitchen stations to complete recipes. We focus on three progressively challenging cooking tasks with different maximum steps: (1) *Burger* - at most 10 steps; (2) *Cheese Burger* - at most 15 steps; and (3) *Double Cheese Burger* - at most 23 steps. These tasks were selected to provide increasing complexity in both planning horizon and reasoning requirements while sharing the in-distribution states and actions, allowing us to evaluate our method's scalability and generalization across different task difficulties.

TABLE I: Performance Comparison on Different Tasks. Best results are in **bold**. "—" indicates no successful trajectories.

Results of Burger	SR (↑)	ASAT (↓)	ASST (↓)
Llama3-70b (expert)	0.53±0.15	12.13±0.57	9.53±0.50
Qwen2.5-7b	0.07±0.06	14.97±0.06	14.0±0.0
Qwen2.5-14b	0.10±0.10	14.87±0.15	13.67±0.29
Qwen1.5b (w/ SFT)	0.30±0.10	12.37±1.17	8.64±0.13
Qwen1.5b (w/ SFT & GRPO)	**0.73±0.06**	**10.07±0.25**	**8.18±0.16**
Results of Cheeseburger	SR (↑)	ASAT (↓)	ASST (↓)
Llama3-70b (expert)	0.73±0.06	17.23±0.38	15.12±0.13
Qwen2.5-7b	0.0±0.0	23.0±0.0	—
Qwen2.5-14b	0.10±0.10	22.63±0.32	19.33±1.15
Qwen1.5b (w/ SFT)	0.57±0.06	16.97±1.36	14.73±0.49
Qwen1.5b (w/ SFT & GRPO)	**0.70±0.00**	**15.53±0.23**	**12.30±0.35**
Results of Double Cheeseburger	SR (↑)	ASAT (↓)	ASST (↓)
Llama3-70b (expert)	0.20±0.10	32.50±0.89	21.77±2.25
Qwen2.5-7b	0.0±0.0	35.0±0.0	—
Qwen2.5-14b	0.0±0.0	35.0±0.0	—
Qwen1.5b (w/ SFT)	0.13±0.06	22.87±0.85	27.17±2.75
Qwen1.5b (w/ SFT & GRPO)	**0.3±0.0**	**30.5±0.0**	**20.0±0.0**

b) Data Collection and Model Training: Expert trajectories are collected using rejection sampling and minimum-step trajectory filtering from Llama3.3-70B-Instruct [10],

TABLE II: Performance comparison on Burger task with different RL post-training methods. Best results are in **bold**.

RL post-training methods	SR (↑)	ASAT (↓)	ASST (↓)
SFT & PPO	0.60±0.10	11.40±0.61	8.53±0.29
SFT & RLOO	0.70±0.00	10.37±0.06	8.38±0.08
SFT & REINFORCE++	0.67±0.12	10.80±1.11	8.30±0.05
SFT & GRPO	**0.73±0.06**	**10.07±0.25**	**8.18±0.16**

TABLE III: Cross-task generalization of SR and ASAT.

Evaluation Tasks (→) Model Training Tasks (↓)	Burger SR / ASAT	Cheese Burger SR / ASAT	Double Cheese Burger SR / ASAT
Burger	0.73±0.06 / 10.07±0.25	0.0±0.0 / 23.0±0.0	0.0±0.0 / 35.0±0.0
Cheese Burger	0.60±0.10 / 10.90±0.17	0.7±0.0 / 15.50±0.23	0.0±0.0 / 35.0±0.0
Double Cheese Burger	0.5±0.0 / 11.60±0.10	0.37±0.06 / 19.13±0.55	0.3±0.0 / 30.5±0.0

which serves as our expert policy π^{GT}. To ensure diversity and robustness, we generate expert trajectories across varied kitchen layouts and ingredient placements, creating approximately 100 successful trajectories per task type. For the ablation of training on noisy non-expert trajectories, we randomly choose 10%, 30%, 50% of the states along all trajectories that are perturbed to have sub-optimal actions but keep the task completion successful. The state-action tokens follow the ReAct agentic framework [2], where each action a_i^{GT} consists of structured reasoning followed by a specific action selection from the valid action list indicated from state tokens. ReAct format and single-turn correctness reward of Definition 2.4 are weighted by 0.1 and 0.9, respectively. Our training pipeline consists of two sequential post-training stages using the VeRL implementation [11] of RL post-training. First, we perform supervised fine-tuning (SFT) on Qwen2.5-1.5B-Instruct [12] using the collected expert trajectories to establish a strong initialization. The SFT takes 8 epochs with batch size of 8 and the RL post-training uses a batch size of 256, learning rate of 1×10^{-6}, and KL regularization coefficient $\beta = 0.001$ for 50 epochs.

c) Evaluation Metrics and Baselines: We evaluate all models on 10 unseen kitchen layouts per task type, ensuring that the test environments differ from training data in terms of object positions, kitchen topology, and ingredient availability. The timeout lengths of attempt turns are 15, 23, 35 for tasks of *Burger, Cheese Burger, Double Cheese Burger*, respectively. We assess the model performance of effectiveness and efficiency using three key metrics: (1) *Success Rate (SR)* - the proportion of episodes where the agent successfully completes the task within the step timeout limit; (2) *Average Steps of All Trajectories (ASAT)* - the mean number of steps taken across all episodes, including timeout failures; and (3)

TABLE IV: Performance comparison on Burger task with different ratio of noisy trajectories. Best results are in **bold**.

Ratio of noisy trajectories	SR (↑)	ASAT (↓)	ASST (↓)
Noisy, 50%	0.50±0.10	11.90±0.26	8.23±0.21
Noisy, 30%	0.56±0.06	11.43±0.55	8.23±0.21
Noisy, 10%	0.67±0.06	10.47±0.29	8.19±0.17
Clean, 0%	**0.73±0.06**	**10.07±0.25**	**8.18±0.16**

Average Steps of Successful Trajectories (ASST) - the mean number of steps for successfully completed episodes only. The temperature is set to be 0.7 and all values are in *Mean ± Standard Deviation* from 3 random trials. We compare our approach against prompt-based larger agents (7B, 14B models of Qwen2.5), 1.5B models only with SFT, and 1.5B models with SFT and other RL post-training methods (PPO [13], RLOO [14], REINFORCE++ [15]).

B. Results of Task Planning with Single-turn GRPO

In Table I, the models with SFT and GRPO consistently outperform much larger-size baseline models and perform better or keep on par with expert policy. The SFT-alone model provides limited success compared to the one followed by GRPO, validating the effectiveness of single-turn RL for task reasoning in Theorem 3.3. Beyond success rates, the model with GRPO post-training consistently achieves lower average step counts for successful episodes, demonstrating that GRPO learns more optimal policies with minimal steps of task completion in Definition 3.2.

From Table II, the proposed GRPO-based approach achieves the best performance in all metrics compared to other RL post-training methods. The PPO-based agent performs the worst due to its reliance on a separate critic network for advantage estimation, which introduces additional approximation errors. Compared to critic-free RL methods like RLOO and REINFORCE++, group-relative advantage computation of GRPO provides more stable gradient estimates by normalizing within sampled responses.

C. Ablation Study and Generalization Analysis

a) Cross-task Generalization Analysis: To evaluate generalizability, we train agents on individual tasks and assess zero-shot performance across all tasks. The results in Table III empirically validate Corollary 3.2: agents trained on complex tasks generalize effectively to simpler ones, while the reverse fails. The Double Cheese Burger-trained agent achieves non-zero success rates on all tasks, whereas Burger-trained and Cheese Burger-trained agents completely fail on more complex task of Double Cheese Burger. These findings demonstrate that our single-turn GRPO approach enables meaningful cross-task generalization from complex to simpler tasks, aligning with our theoretical predictions, though highlighting that complex reasoning capabilities cannot be readily acquired from simpler task training alone.

b) Robustness to Noisy Non-expert Trajectories: To assess robustness to imperfect demonstrations, we evaluate the Burger task with varying ratios of noisy trajectories. Table IV shows that our method exhibits strong robustness: even with 50% noisy data, success rate only drops to 0.50 from 0.73, while efficiency metrics (ASAT and ASST) remain stable. Performance degrades gradually as noise increases. This robustness stems from GRPO's group-relative advantage computation, which effectively filters suboptimal demonstrations

during optimization. These results indicate that our single-turn GRPO approach can successfully learn from imperfect expert trajectories, valuable for practical deployment where obtaining perfectly optimal demonstrations is challenging.

V. CONCLUSION

In this paper, we present a theoretical framework demonstrating that GRPO improvements on single-turn task reasoning provide a lower bound on multi-turn success probability under minimal steps for challenging multi-turn interaction scenarios. Empirical results validate our approach, with our small-parameter model consistently outperforming larger baselines while requiring fewer steps for task completion. We also show cross-task generalization capabilities both theoretically and empirically. While limitations remain regarding the reliance on expert trajectories, our work shows how to leverage single-turn RL successes for complex multi-turn task planning, opening promising directions for more capable and efficient LLM-based agents.

REFERENCES

[1] J. Luo, W. Zhang, Y. Yuan, Y. Zhao *et al.*, "Large language model agent: A survey on methodology, applications and challenges," *arXiv preprint arXiv:2503.21460*, 2025.

[2] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *International Conference on Learning Representations (ICLR)*, 2023.

[3] S. Guan, H. Xiong, J. Wang, J. Bian, B. Zhu, and J.-g. Lou, "Evaluating llm-based agents for multi-turn conversations: A survey," *arXiv preprint arXiv:2503.22458*, 2025.

[4] Z. Wang, K. Wang, Q. Wang, P. Zhang, L. Li, Z. Yang, X. Jin, K. Yu, M. N. Nguyen, L. Liu *et al.*, "Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning," *arXiv preprint arXiv:2504.20073*, 2025.

[5] L. Feng, Z. Xue, T. Liu, and B. An, "Group-in-group policy optimization for llm agent training," *arXiv preprint arXiv:2505.10978*, 2025.

[6] F. Peiyuan, Y. He, G. Huang, Y. Lin, H. Zhang, Y. Zhang, and H. Li, "Agile: A novel reinforcement learning framework of llm agents," *Advances in Neural Information Processing Systems*, vol. 37, pp. 5244–5284, 2024.

[7] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[8] Y. Mroueh, "Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification," *arXiv preprint arXiv:2503.06639*, 2025.

[9] G. Gonzalez-Pumariega, L. S. Yean, N. Sunkara, and S. Choudhury, "Robotouille: An asynchronous planning benchmark for llm agents," in *The Thirteenth International Conference on Learning Representations*.

[10] A. Dubey, A. Jauhri, A. Pandey, A. Kadian *et al.*, "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.

[11] G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu, "Hybridflow: A flexible and efficient rlhf framework," in *Proceedings of the Twentieth European Conference on Computer Systems*, 2025, pp. 1279–1297.

[12] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, and et al, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[14] A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker, "Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms," *arXiv preprint arXiv:2402.14740*, 2024.

[15] J. Hu, J. K. Liu, H. Xu, and W. Shen, "Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models," *arXiv preprint arXiv:2501.03262*, 2025.