

MatchTIR: Fine-Grained Supervision for Tool-Integrated Reasoning via Bipartite Matching

Changle Qu^{1*}, Sunhao Dai¹, Hengyi Cai², Jun Xu^{1†}, Shuaiqiang Wang², Dawei Yin²

¹Gaoling School of Artificial Intelligence, Renmin University of China; ²Baidu Inc.

{changlequ, sunhaodai, junxu}@ruc.edu.cn, caihengyi@ict.ac.cn,
wangshuaiqiang@baidu.com, yindawei@acm.org

Abstract

Tool-Integrated Reasoning (TIR) empowers large language models (LLMs) to tackle complex tasks by interleaving reasoning steps with external tool interactions. However, existing reinforcement learning methods typically rely on outcome- or trajectory-level rewards, assigning uniform advantages to all steps within a trajectory. This coarse-grained credit assignment fails to distinguish effective tool calls from redundant or erroneous ones, particularly in long-horizon multi-turn scenarios. To address this, we propose MatchTIR, a framework that introduces fine-grained supervision via bipartite matching-based turn-level reward assignment and dual-level advantage estimation. Specifically, we formulate credit assignment as a bipartite matching problem between predicted and ground-truth traces, utilizing two assignment strategies to derive dense turn-level rewards. Furthermore, to balance local step precision with global task success, we introduce a dual-level advantage estimation scheme that integrates turn-level and trajectory-level signals, assigning distinct advantage values to individual interaction turns. Extensive experiments on three benchmarks demonstrate the superiority of MatchTIR. Notably, our 4B model surpasses the majority of 8B competitors, particularly in long-horizon and multi-turn tasks. Our codes are available at <https://github.com/quchangle1/MatchTIR>.

1 Introduction

Tool-Integrated Reasoning (TIR) has established itself as a critical paradigm for enhancing the capabilities of large language models (LLMs) by enabling them to interact with external tools during the reasoning process (Gou et al., 2024). By leveraging external tools, TIR allows LLMs to overcome limitations of static parametric knowledge (Jin et al.,

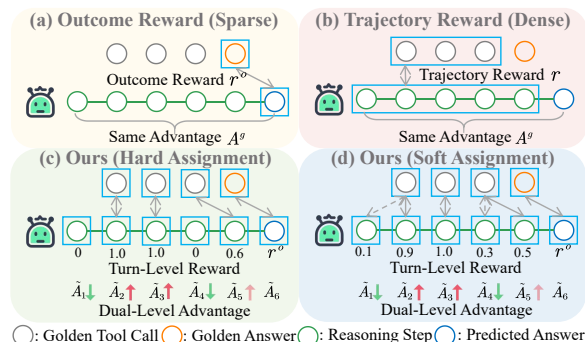


Figure 1: Comparison of reward and advantage assignment strategies for multi-turn TIR. (a) and (b) show traditional methods where all reasoning steps share the same reward and advantage. (c) and (d) illustrate our MatchTIR framework using hard and soft assignment to derive turn-level rewards and distinct advantages.

2025; Li et al., 2025b; Zhou et al., 2026), perform precise computations (Wang et al., 2024a; Das et al., 2024), and engage dynamically with external environments (Qin et al., 2024; Qu et al., 2024, 2025a,b; Wang et al., 2024b). Through multi-turn interactions, agents interleave reasoning, tool execution, and feedback acquisition, refining their trajectory toward solving complex real-world tasks.

To equip agents with such sophisticated tool-integrated reasoning capabilities, recent research has largely adopted reinforcement learning with verifiable rewards (RLVR) methods (Chang et al., 2025; Jiang et al., 2025; Zheng et al., 2025c; Lin and Xu, 2025), such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024). While early RLVR formulations primarily relied on sparse outcome-based rewards (Li et al., 2025a), subsequent works have incorporated trajectory-level signals (Qian et al., 2025; Zeng et al., 2025b) to provide denser supervision. However, as shown in Figure 1, even these trajectory-level formulations typically assign a uniform advantage value to every turn within a trajectory. This “one-size-fits-all” credit assignment fails to distinguish critical

*Work done during internship at Baidu.

†Corresponding author.

reasoning steps from redundant or erroneous tool calls, leading to inefficient optimization and consequently hindering the model from learning precise and efficient tool-use strategies.

Several studies have attempted to introduce fine-grained rewards via intrinsic signals or external reward models (Wang et al., 2025a; Zhang et al., 2025c). However, external reward models are inherently susceptible to bias and hallucination, while Monte Carlo estimation (Dong et al., 2025; Yang et al., 2026a) incurs prohibitive computational costs and high variance, especially in long-horizon tasks. Moreover, many of these efforts are confined to restricted tools such as search engines, where intermediate steps are characterized by high semantic variance due to the fact that multiple different queries may be equally valid, making it difficult to establish a unique ground truth for reliable step-level supervision. In contrast, general TIR scenarios naturally expose structured and verifiable signals, such as tool names, parameter names, and parameter contents, which enable explicit evaluation of the correctness of the tool use at each turn. This motivates us to formulate turn-level reward assignment as a bipartite matching problem between predicted and ground-truth tool interactions.

In this paper, we propose MatchTIR, a framework designed to assign distinct, precise advantages to individual turns within a reasoning trajectory. We formulate turn-level credit assignment as a bipartite matching problem, constructing a weighted bipartite graph based on similarity scores across tool names, parameter names, and parameter contents to align predicted calls with ground-truth references. Based on this formulation, we introduce both hard and soft credit assignment strategies to derive dense turn-level rewards: the hard variant enforces “one-to-one” matching, while the soft variant allows “one-to-many” alignment. Both strategies produce dense and precise turn-level rewards, enabling effective supervision of intermediate tool interactions. To ensure the model optimizes for both local accuracy and global success, MatchTIR synthesizes turn-level rewards with final outcome-based rewards. Furthermore, we introduce a dual-level advantage estimation mechanism that harmonizes local turn-level rewards with global trajectory-level outcomes. Specifically, the trajectory-level advantage aggregates all rewards across the entire path to evaluate global quality, while the turn-level advantage captures the t -th turn’s specific contribution via discounted accu-

mulation of all subsequent rewards. The policy is finally optimized using the GRPO objective with our integrated dual-level advantages.

Extensive experiments on both in-domain and out-of-domain benchmarks demonstrate the effectiveness and robustness of MatchTIR. In summary, our main contributions are as follows:

- We identify the uniform credit assignment problem in existing TIR methods as a key bottleneck preventing efficient optimization.
- We propose MatchTIR, which formulates turn-level credit assignment as a bipartite matching problem and introduces hard and soft matching strategies to provide dense, turn-level supervision.
- Extensive experiments on three benchmarks demonstrate the effectiveness of MatchTIR, with particularly strong gains on long-horizon scenarios, validating its robustness and generalizability.

2 Preliminaries

In this section, we first present the task formulation of TIR, and then introduce the reinforcement learning framework adopted in this work.

2.1 Task Formulation

Formally, given a user query q and a tool set $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$ consisting of n available tools, the goal of the agent is to generate an interaction trajectory $\tau = \{s_1, s_2, \dots, s_T\}$ to solve q , where T denotes the number of interaction turns. Specifically, each interaction turn $s_i = (n_i, c_i, o_i)$ is represented as a triplet, where n_i is the natural language reasoning in i -th turn, c_i denotes the set of invoked tools from \mathcal{T} along with their specific parameter names and parameter contents, and o_i is the observation returned by the environment. The trajectory terminates when the agent generates a final answer response (where $c_T = \emptyset$ and $o_T = \emptyset$) or reaches the pre-defined maximum turn limit L .

Goal. As discussed in §1, TIR is inherently a long-horizon, multi-turn process. However, existing outcome- or trajectory-level reward formulations assign uniform credit across an entire trajectory, failing to distinguish between effective and unnecessary tool invocations. This results in inefficient credit assignment and suboptimal tool-use behaviors. Therefore, our goal is to assign precise turn-level rewards, enabling fine-grained credit assignment that incentivizes effective tool usage contributing to the final solution while penalizing redundant or erroneous actions at each turn.

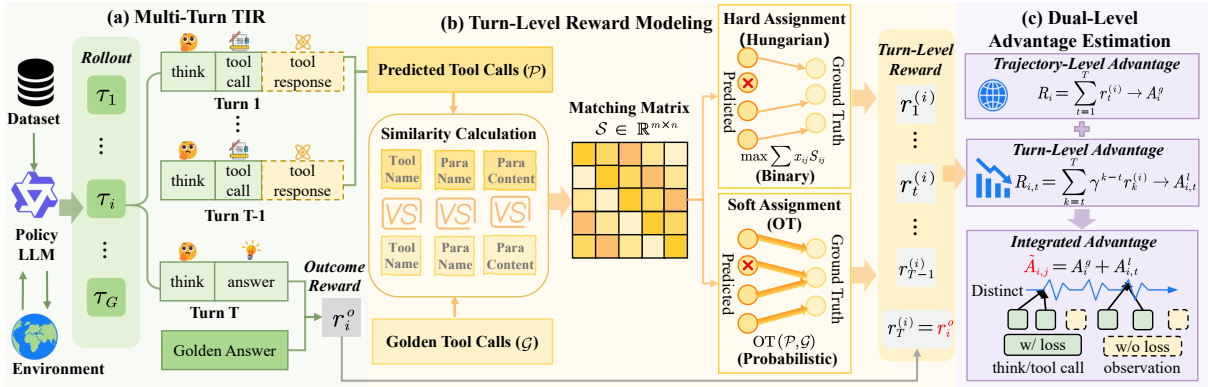


Figure 2: The illustration of our proposed MatchTIR. (a) illustrates the process of multi-turn TIR, where the policy LLM interacts with external tools over multiple turns and receives an outcome reward. (b) shows turn-level reward modeling by matching predicted and golden tool calls using hard or soft assignment strategies. (c) depicts the dual-level advantage estimation that integrates trajectory-level and turn-level signals for policy optimization.

2.2 Agentic Reinforcement Learning

To optimize the agent policy π_θ , we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a specialized reinforcement learning algorithm. Distinct from conventional methods, GRPO obviates the requirement for a learned value function by estimating advantages based on the relative performance within a group of G rollouts $\{\tau_i\}_{i=1}^G$ generated from the same prompt. The GRPO training objective can be formulated as

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim D, \{\tau_i\} \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \frac{1}{G} \sum_{i=1}^G \frac{1}{|\tau_i|} \sum_{j=1}^{|\tau_i|} \left[\min(w_{i,j} \hat{A}_{i,j}, \text{clip}(w_{i,j}, 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,j}) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta | \pi_{\text{ref}}) \right],$$

where π_{old} is the old policy, $w_{i,j} = \frac{\pi_\theta(\tau_{i,j}|q, \tau_{i,<j})}{\pi_{\theta_{\text{old}}}(\tau_{i,j}|q, \tau_{i,<j})}$ is the importance sampling ratio, $\hat{A}_{i,j}$ is the normalized advantage of the j -th token in the i -th rollout. The hyper-parameters ϵ and β denote the clipping threshold and the coefficient for KL divergence, respectively, where the latter regularizes the policy to remain close to the reference model π_{ref} . During training, tokens corresponding to tool responses are masked out when computing the objective, since these tokens are generated by the external environment rather than being controlled by the agent.

3 Our Approach: MatchTIR

In this section, we will provide a detailed introduction to our proposed MatchTIR. The overall framework is illustrated in Figure 2.

3.1 Reward Modeling

As shown in Figure 2(a), TIR is inherently a multi-turn process, where an agent interleaves reasoning with multiple tool invocations before arriving at a final answer. To optimize the agent’s policy, it is crucial to determine whether each predicted tool call contributes to the reasoning goal. However, simply comparing individual calls is insufficient, as the agent may generate redundant tool invocations. To address this, we reformulate turn-level credit assignment for TIR as a bipartite matching problem between two sets. Specifically, given a rollout trajectory, we extract a set of m predicted tool calls $\mathcal{P} = \{p_1, \dots, p_m\}$ and identify n ground-truth tool calls $\mathcal{G} = \{g_1, \dots, g_n\}$. Our goal is to derive precise rewards by finding an optimal alignment between \mathcal{P} and \mathcal{G} , thereby penalizing uninformative or repetitive actions while rewarding those that align with the expert reasoning path.

Matching Matrix Construction. We first construct a matching matrix $\mathcal{S} \in \mathbb{R}^{m \times n}$, where each entry S_{ij} represents the matching score between a predicted tool p_i and a ground-truth tool g_j . Inspired by the previous work (Qian et al., 2025), the matching score S_{ij} includes three components:

- **Tool Name Matching.** We first check the consistency of the tool name. Let p_i^{tn} and g_j^{tn} be the name of the predicted and golden tool, respectively:

$$S_{tn} = \mathbb{I}(p_i^{\text{tn}} = g_j^{\text{tn}}) \in \{0, 1\}, \quad (1)$$

where $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the predicted tool name matches the ground-truth tool name, and 0 otherwise.

- **Parameter Name Matching.** If the tool name matches, we calculate Jaccard similarity of param-

ter name sets N_{p_i} and N_{g_j} to measure the overlap:

$$S_{pn} = \frac{|N_{p_i} \cap N_{g_j}|}{|N_{p_i} \cup N_{g_j}|} \in [0, 1], \quad (2)$$

• **Parameter Content Matching.** We further assess the correctness of the values provided for each ground-truth parameter:

$$S_{pc} = \sum_{k \in N_{g_j}} \mathbb{I}(p_i[k] = g_j[k]) \in [0, |N_{g_j}|], \quad (3)$$

where $p_i[k]$ and $g_j[k]$ denote the parameter content associated with the parameter name k .

The final similarity score S_{ij} is normalized by the maximum possible score to ensure $S_{ij} \in [0, 1]$:

$$S_{ij} = S_{tn} \cdot \frac{S_{tn} + S_{pn} + S_{pc}}{1 + 1 + |N_{g_j}|}. \quad (4)$$

While the matrix \mathcal{S} provides pairwise alignment scores, directly using these scores as rewards can lead to reward hacking, such as repeatedly invoking similar tools to accumulate high similarity scores. To mitigate this, we introduce two assignment strategies to convert \mathcal{S} into turn-level rewards.

Hard Credit Assignment. The hard credit assignment strategy enforces a strict “one-to-one” alignment between predicted and ground-truth tool calls, preventing duplicated credit for the same ground-truth action. We formulate this process as a maximum weight bipartite matching problem and solve it using the Hungarian (Kuhn–Munkres) algorithm (Kuhn, 1956). Formally, let $x_{ij} \in \{0, 1\}$ be a binary assignment variable that $x_{ij} = 1$ implies that the predicted tool p_i is matched with the ground-truth tool g_j . The objective is to maximize the total matching score $\sum_{i,j} x_{i,j} S_{i,j}$, subject to the constraint that each predicted and ground-truth tool call is matched at most once. Based on the resulting matching, the reward for each predicted tool call p_i is defined as

$$r_{p_i} = \begin{cases} S_{ij}, & \text{if } \exists j, x_{ij} = 1; \\ -\lambda, & \text{otherwise,} \end{cases} \quad (5)$$

where λ is a fixed penalty (or zero) assigned to erroneous or redundant tool calls that fail to match any ground-truth tool call.

Soft Credit Assignment. To facilitate a smoother optimization landscape and provide more nuanced feedback, we propose a soft relaxation of the assignment problem based on Optimal Transport

(OT) (Cuturi, 2013). Instead of enforcing a hard binary choice, this method treats the predicted and ground-truth tool sequences as discrete probability distributions and optimizes the cost of transforming one into the other. The matching matrix \mathcal{S} is transformed into a cost matrix \mathcal{C} via a decreasing transformation (e.g., $C_{ij} = -S_{ij}$), ensuring that high matching scores translate to minimal transportation costs. We then compute an optimal transport plan \mathcal{Z} that minimizes the total cost:

$$\min_{\mathcal{Z}} \sum_{i,j} Z_{ij} C_{ij}, \quad \text{s.t. } \mathcal{Z} \mathbf{1}_n = \mathbf{a}, \mathcal{Z}^\top \mathbf{1}_m = \mathbf{b}, \quad (6)$$

where $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^n$ are marginal distributions representing the importance weights of predicted and ground-truth tool calls, respectively.

The resulting transport plan \mathcal{Z} effectively functions as a probabilistic mapping between predicted and ground-truth interactions. Given the resulting transport plan, the reward r_{p_i} for each predicted tool call is then calculated as the weighted sum of its similarities across the ground-truth set:

$$r_{p_i} = \sum_j Z_{ij} \cdot S_{ij}. \quad (7)$$

Unlike the hard version, soft assignment allows a single ground-truth tool call reward to be distributed across multiple predicted turns. This mechanism ensures that the turn with the highest alignment receives the largest share of credit, while other relevant attempts still receive partial, dense feedback to guide the policy’s improvement.

Turn-Level Reward. The assignment mechanisms described above yield a specific reward r_{p_i} for each individual predicted tool call. However, within the multi-turn TIR framework, a single interaction turn may involve multiple tool calls, whereas policy optimization is typically performed at the turn level. Therefore, it is necessary to aggregate tool call rewards into a unified turn-level signal. To this end, we compute the turn-level reward by averaging the rewards of predicted tool calls within the same turn:

$$r_t = \frac{1}{|P_t|} \sum_{p_i \in P_t} r_{p_i}, \quad (8)$$

where $|P_t|$ denotes the number of predicted tool calls issued at turn t . This averaging strategy normalizes the reward across turns and discourages excessive tool usage, ensuring fair comparison between turns with different numbers of tool calls.

Outcome-Level Reward. To complement the fine-grained supervision provided by turn-level rewards, we incorporate an outcome-level reward r^o to evaluate the global objective fulfillment of the entire reasoning trajectory. Specifically, we use the F1 score between the predicted answer \hat{a} and the golden answer a as the outcome-level reward:

$$r^o = \frac{2|\hat{a} \cap a|}{|\hat{a}| + |a|} \in [0, 1]. \quad (9)$$

This outcome-level reward serves as a global constraint that complements the turn-level supervision, encouraging the model to not only make correct tool calls at each turn but also remain aligned with successful end-task completion.

3.2 Dual-Level Advantage Estimation

In conventional outcome-supervised RL frameworks like GRPO, all tokens within a sampled rollout are typically assigned a uniform advantage value derived from the final outcome reward. While this strategy is effective for single-turn or short-horizon tasks, it is fundamentally suboptimal for TIR, where a trajectory consists of multiple interaction turns with highly heterogeneous contributions. Assigning identical advantages to all tokens fails to distinguish between informative actions that advance task completion and uninformative or redundant tool-use behaviors, resulting in coarse credit assignment and noisy policy optimization. To address this limitation, we propose a dual-level advantage estimation framework that evaluates model actions from two complementary perspectives:

Trajectory-Level Advantage Estimation. We first evaluate the overall quality of each rollout within a group of G sampled trajectories. For the i -th rollout, we define the trajectory-level reward R_i as the sum of its turn- and outcome-level rewards:

$$R_i = \sum_{t=1}^T r_t^{(i)}, \quad (10)$$

where $r_t^{(i)}$ is the turn-level reward at turn t of the i -th rollout. For the final turn T , $r_T^{(i)}$ corresponds to the outcome-level reward r_i^o .

Following the GRPO framework, the trajectory-level advantage A_i^g is computed by normalizing the rewards across the group: $A_i^g = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^G)}{\text{std}(\{R_j\}_{j=1}^G)}$. This advantage reflects the relative performance of the entire reasoning path compared to others

generated for the same prompt, providing a global supervisory signal for policy optimization.

Turn-Level Advantage Estimation. While trajectory-level advantage provides a global performance signal, it assigns identical supervision to all interaction turns within the trajectory. To further differentiate the contributions of individual turns, we introduce local turn-level advantage estimation, which provides fine-grained supervision aligned with turn-level rewards. Specifically, to capture the impact of current decisions on future turns, we first compute a discounted cumulative reward $R_{i,j}$ for each turn t in the i -th rollout:

$$R_{i,t} = \sum_{k=t}^T \gamma^{k-t} r_k^{(i)}, \quad (11)$$

where γ is the discount factor. This formulation encourages the model to account for the long-term consequences of its tool-use behaviors.

To provide fine-grained supervision, we then compute the local turn-level advantage $A_{i,t}^l$ by comparing the cumulative reward against its group-level baseline: $A_{i,t}^l = \frac{R_{i,t} - \text{mean}(\{R_{j,t}\}_{j=1}^G)}{\text{std}(\{R_{j,t}\}_{j=1}^G)}$. This term captures fine-grained differences across interaction turns. This dual-view advantage estimation enables more precise credit assignment for TIR.

3.3 Policy Optimization

Advantage Integration. To jointly leverage the global task completion signal and local turn-level supervision, we integrate the trajectory-level and turn-level advantages into a unified supervisory signal. For a token at position j belonging to turn t within the i -th rollout, the integrated advantage $\tilde{A}_{i,j}$ is defined as $\tilde{A}_{i,j} = A_i^g + A_{i,t}^l$.

Objective Function. Using the integrated advantage, we optimize the policy π_θ under the GRPO framework by replacing the advantage with $\tilde{A}_{i,j}$:

$$\mathcal{J}(\theta) = \mathbb{E}_{q \sim D, \{\tau_i\} \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \frac{1}{G} \sum_{i=1}^G \frac{1}{|\tau_i|} \sum_{j=1}^{|\tau_i|} \left[\min(w_{i,j}, \tilde{A}_{i,j}, \text{clip}(w_{i,j}, 1 - \epsilon, 1 + \epsilon)\tilde{A}_{i,j}) - \beta \mathbb{D}_{\text{KL}}(\pi_\theta | \pi_{\text{ref}}) \right].$$

4 Experiments

In this section, we describe the experimental setups and present a comprehensive evaluation, with additional results provided in appendix C.

Methods	FTRL				BFCL							ToolHop	Avg.
	Solve-P	Solve-R	Solve-F1	Avg.	Multi-Turn			Agentic		Avg.	AC		
					Base	MF	MP	LC	Search			Memory	
Qwen3-4B													
Vanilla	30.78	29.65	25.85	28.76	41.50	31.00	26.50	27.50	9.00	19.57	22.96	31.63	27.78
GRPO	31.13	32.83	30.67	31.54	45.00	37.50	26.50	29.50	9.50	21.94	25.17	37.25	31.32
FTRL-S	34.10	34.07	31.54	33.24	43.00	35.50	31.00	28.00	6.50	23.44	24.67	38.63	32.18
FTRL-M	34.47	35.44	33.50	34.47	44.50	41.00	32.50	29.50	6.50	22.80	25.76	41.24	33.82
ToolRL-S	28.26	28.32	23.78	26.79	32.50	31.00	22.50	20.00	7.50	19.14	19.91	30.28	25.66
ToolRL-M	30.90	36.26	32.28	33.15	50.00	37.50	29.00	30.00	10.50	22.80	26.64	35.68	31.82
Ours (OT)	31.79	37.52	32.60	33.97	50.00	40.50	26.50	35.00	12.00	21.72	27.43	41.95	34.45
Ours (KM)	32.39	39.70	34.21	35.43	50.50	47.00	28.50	36.50	17.50	22.15	30.23	42.55	36.07
Qwen3-8B													
Vanilla	28.08	36.55	29.74	31.46	47.50	46.00	37.50	34.50	10.00	20.43	28.30	42.21	33.99
GRPO	31.59	39.75	32.54	34.63	52.50	45.50	34.50	36.50	14.50	18.92	29.48	40.64	34.92
FTRL-S	32.32	38.87	32.85	34.68	51.50	45.00	35.50	34.00	12.50	17.42	28.23	36.72	33.21
FTRL-M	33.28	40.60	34.20	36.03	50.50	51.00	35.00	33.50	15.00	21.94	30.49	43.32	36.61
ToolRL-S	25.57	35.31	26.72	29.20	41.00	39.50	31.50	25.00	7.50	17.63	23.41	32.93	28.51
ToolRL-M	30.64	40.07	32.51	34.41	50.00	47.00	39.50	32.00	15.50	16.77	29.13	42.55	35.36
Ours (OT)	33.61	42.56	33.61	36.59	55.50	52.00	38.50	36.00	18.00	23.87	33.22	45.80	38.54
Ours (KM)	36.33	44.18	37.33	39.28	60.00	49.00	39.00	40.50	17.00	26.02	34.32	46.16	39.92

Table 1: Performance comparison between MatchTIR and the baselines on three benchmarks with two LLMs. For our method, KM and OT denote the hard credit assignment and the soft credit assignment, respectively. For BFCL, MF, MP, and LC correspond to the Miss Function, Miss Parameter, and Long Context subsets, respectively. We highlight the best performance of Qwen3-4B models and Qwen3-8B models with the red and blue, respectively.

4.1 Experimental Setups

Training Dataset. We train the model on the FTRL dataset (Ye et al., 2025b), which comprises over 2,000 automatically constructed tool-use environments with verifiable feedback, covering single-hop, multi-hop, and parallel reasoning scenarios.

Evaluation Benchmarks. We evaluate MatchTIR on three benchmarks: FTRL is used for in-domain evaluation, while BFCL (Patil et al., 2025) and ToolHop (Ye et al., 2025a) are adopted for out-of-domain evaluation to assess generalization. Details of these benchmarks are shown in Appendix A.

Baselines. To evaluate the effectiveness of our proposed MatchTIR, we compare it against comprehensive baselines: (1) Vanilla: We evaluate the vanilla Qwen3 (Yang et al., 2025) models without reinforcement learning. (2) GRPO (Shao et al., 2024): The standard GRPO framework that optimizes the policy solely based on outcome rewards. (3) ToolRL (Qian et al., 2025): ToolRL assigns rewards based on the correctness of tool names, parameter names, and parameter content. While originally designed for single-turn tool use, we extend it to the multi-turn interaction setting by applying its scoring mechanism at the trajectory level. The single-turn and multi-turn variants are denoted as ToolRL-S and ToolRL-M, respectively.

(4) FTRL (Ye et al., 2025b): FTRL computes rewards using the ratio of successful tool calls to total attempts. Similar to ToolRL, we include both the single-turn baseline (FTRL-S) and its adaptation to the multi-turn setting (FTRL-M).

Implementation Details. We adopt Qwen3-4B and Qwen3-8B as backbone models and optimize them using the GRPO objective implemented in the verl framework (Sheng et al., 2025). All experiments are conducted under the Zero-RL setting, with a batch size of 256 and 16 rollouts sampled per query. The maximum number of turns is set to 10, the penalty scale λ is set to 0, and the discount factor γ is set to 0.9. The reinforcement learning phase spans 3 epochs, conducted on 8 NVIDIA A800-80G GPUs. Please refer to Appendix B for more details about hyperparameter configurations.

4.2 Experimental Results

We present the experimental results in Table 1, from which we derive the following observations:

- Firstly, compared to vanilla models, most post-training methods generally achieve consistent performance gains across all benchmarks, demonstrating the effectiveness of reinforcement learning for optimizing complex agentic behaviors in TIR.
- Secondly, compared to single-turn training, models trained in a multi-turn setting exhibit signif-

Reward Design	FTRL			
	Solve-P	Solve-R	Solve-F1	Avg.
<i>Trajectory-Level Advantage Estimation</i>				
Outcome	31.59	39.75	32.54	34.63
Turn-Level	32.70	40.70	35.08	36.16
Integrated Reward	32.31	43.16	35.11	36.86
<i>Turn-Level Advantage Estimation</i>				
Turn-Level	29.50	42.06	32.01	34.52
Integrated Reward	32.88	41.67	34.65	36.40
<i>Dual-Level Advantage Estimation</i>				
Turn-Level	32.45	43.08	35.06	36.96
MatchTIR	36.33	44.18	37.33	39.28

Table 2: Ablation study on reward design and advantage estimation on the FTRL dataset using Qwen3-8B.

icantly superior performance under similar reward designs, confirming the necessity of end-to-end optimization for complex interaction sequences.

- Furthermore, compared to all baseline methods, both variants of MatchTIR achieve the highest performance across three benchmarks, demonstrating strong generalization that overcomes the potential exploration constraints imposed by ground-truth alignment. Notably, our 4B models even surpass most 8B baselines. This improvement can be attributed to our turn-level reward design and dual-level advantage estimation, which effectively distinguishes useful tool calls from mistakes.

- Finally, compared to the soft variant (OT), the hard variant (KM) consistently yields better results. This suggests that soft alignment may assign partial credit to near-miss tool calls (e.g., parameter errors) which are actually fatal in execution, leading to sub-optimal policy updates. In contrast, KM provides a rigorous binary signal essential for exact tool usage by enforcing a strict one-to-one mapping.

4.3 Ablation Study

To evaluate the contribution of each component within our framework, we conduct a comprehensive ablation study on the Qwen3-8B model across two dimensions: **reward design** and **advantage estimation strategies**. As shown in Table 2, turn-level rewards consistently outperform outcome-based rewards, indicating the importance of supervising intermediate tool interaction steps. Furthermore, combining turn-level and outcome rewards yields additional performance gains, indicating that effective optimization for TIR benefits from jointly modeling intermediate tool execution quality and final task correctness. Regarding advantage estimation, using either trajectory-level or turn-level advantage alone leads to suboptimal performance.

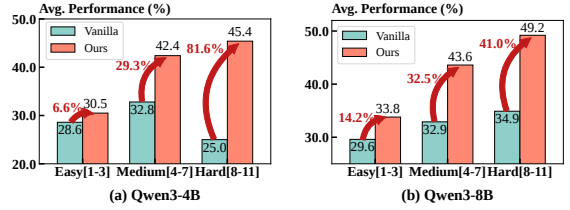


Figure 3: Performance comparison on subsets of varying task complexity, where queries are grouped based on the number of tools required to solve each task.

Model	Method	Tool Calls ↓	Success ↑	Failure ↓
Qwen3-4B	Vanilla	1444	15.44%	84.56%
	Ours	1297	27.83%	72.17%
Qwen3-8B	Vanilla	1593	19.08%	80.92%
	Ours	1373	27.90%	72.10%

Table 3: Analysis of tool-use efficiency and accuracy.

In contrast, integrating both consistently yields superior results across all evaluation metrics. This suggests that assigning token-level credit requires balancing fine-grained local execution signals with global trajectory-level context. Overall, MatchTIR achieves the best performance by effectively synthesizing these components, confirming that fine-grained turn-level signals effectively complement global trajectory-level supervision, enabling more precise and reliable optimization.

4.4 Robustness to Task Complexity

To evaluate how MatchTIR adapts to varying reasoning depths, we categorize the evaluation queries into three difficulty levels, namely Easy, Medium, and Hard, based on the number of tool invocations required to resolve each task. As illustrated in Figure 3, MatchTIR consistently outperforms the vanilla baseline across all complexity levels for both the 4B and 8B model scales, demonstrating robust generalizability. Notably, the performance gains become increasingly pronounced as task complexity scales, with the most significant gains observed on the Hard subset. This trend indicates that long-horizon, multi-turn tool-interaction tasks benefit more substantially from turn-level reward assignment. In these complex scenarios, individual tool calls contribute unevenly to the final outcome, and a uniform advantage to all tokens within such a long trajectory inevitably obscures these critical distinctions. By providing fine-grained supervision, MatchTIR enables more effective credit assignment, thereby yielding superior performance, especially on long-horizon, multi-turn tasks.

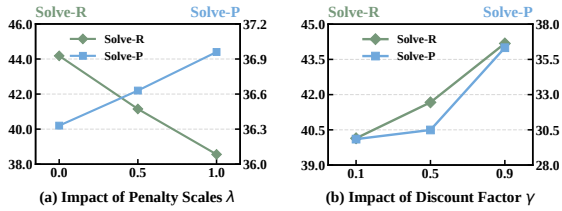


Figure 4: Sensitivity analysis of performance to hyper-parameters. (a) shows the impact of penalty scales λ . (b) illustrates the effect of discounted factor γ .

4.5 Tool-Use Efficiency and Accuracy

To assess how turn-level reward assignment influences the tool-use behavior of the agent, we analyze the total number of tool calls as well as the proportions of successful and failed tool invocations. As shown in Table 3, models trained with our approach consistently make fewer tool calls while achieving substantially higher tool-call success rates across both Qwen3-4B and Qwen3-8B. Specifically, our method reduces the overall frequency of tool invocations and significantly lowers the failure rate, indicating that agents learn to invoke tools more selectively and accurately rather than relying on excessive or redundant calls. These results suggest that fine-grained, turn-level supervision not only improves task-level performance but also leads to more efficient and reliable tool-use strategies.

4.6 Hyper-parameter Analysis

Impact of Penalty Scales. As illustrated in Figure 4(a), we observe a distinct trade-off between precision and success rate as the penalty scale λ increases. Specifically, intensifying the penalty for unmatched or redundant tool calls leads to a consistent improvement in Solve-P, indicating that the model becomes more conservative and effectively suppresses redundant or erroneous tool invocations. However, this increased stringency results in a marginal decline in Solve-R, suggesting that overly strong penalties may inhibit exploratory yet necessary tool calls in high-uncertainty scenarios. These results highlight the importance of balancing penalty strength to encourage precise tool usage without excessively constraining the model’s ability to explore and resolve complex queries.

Impact of Discount Factor. Figure 4(b) demonstrates the performance sensitivity to the discount factor γ . We observe a consistent and significant improvement across all evaluation metrics as γ increases from 0.1 to 0.9. This trend underscores the long-horizon nature of multi-turn TIR, where early

tool-calling decisions critically shape subsequent interactions and ultimately determine task success. A higher γ effectively enables the turn-level advantage to capture these long-term dependencies, ensuring that intermediate actions are optimized not just for immediate correctness, but for their contribution to the overall task completion.

5 Related Work

5.1 Tool-Integrated Reasoning

Recently, RLVR, such as GRPO-based methods (Shao et al., 2024; Yu et al., 2025a; Zheng et al., 2025a), has been widely adopted to enhance TIR by directly optimizing models against task-level feedback (Wang et al., 2025c; Xue et al., 2025; Qu et al., 2025c). Early RL-based approaches rely on rewards derived solely from final outcomes, resulting in sparse and delayed supervision (Li et al., 2025a; Feng et al., 2025a). Subsequent methods mitigate this issue by assigning rewards at the trajectory level, providing denser training signals (Qian et al., 2025; Zhang et al., 2025a; Zeng et al., 2025b; Wei et al., 2025; Yang et al., 2026b; Chen et al., 2025; Gan et al., 2026). Despite these advances, most existing methods assign a single reward to an entire trajectory, causing all actions to share identical advantage signals regardless of their individual quality. Such coarse-grained feedback prevents models from distinguishing effective from ineffective tool calls and obscures turn-level contributions in multi-turn scenarios (Lightman et al., 2024). In contrast, our work introduces a turn-level reward assignment and dual-level advantage estimation framework, which enables fine-grained supervision.

5.2 Fine-Grained Credit Assignment

Recent studies have explored turn-level or process-level rewards to perform fine-grained credit assignment (Zeng et al., 2025a; Yu et al., 2025b; Yue et al., 2025; Zheng et al., 2025b; Zhang et al., 2025b), which can be broadly categorized into sampling-based and model-based methods. Sampling-based approaches typically employ Monte Carlo rollouts to approximate turn-level credit (Feng et al., 2025b; Tran et al., 2025). While theoretically unbiased, these approaches suffer from high variance and substantial computational overhead, especially in multi-turn interaction settings. Model-based methods instead leverage intrinsic reward functions or external reward models to estimate the utility of intermediate steps (Wang et al., 2025a,b). However,

the accuracy of these rewards is not guaranteed, as these signals are often model-dependent, difficult to calibrate, and prone to limited generalization across domains. In this paper, we exploit the structured nature of TIR by formulating turn-level credit assignment as a bipartite matching problem, enabling fine-grained and precise turn-level rewards.

6 Conclusion

In this paper, we propose MatchTIR, a novel framework designed to provide fine-grained supervision for Tool-integrated Reasoning (TIR) via bipartite matching. By introducing a turn-level reward design and dual-level advantage estimation, our method assigns distinct advantages to individual tokens, effectively distinguishing high-quality tool calls from redundant or uninformative ones. Extensive experiments on both in-domain and out-of-domain benchmarks demonstrate the effectiveness of MatchTIR across various model scales, with particularly pronounced improvements in long-horizon, multi-turn interaction scenarios.

Limitations

Despite the significant performance gains achieved on the Qwen3-4B and Qwen3-8B model series, our work has certain limitations. First, due to constrained computational resources, we were unable to conduct experiments on larger-scale models. We hope that the publication of this work will encourage further exploration of our framework on more massive architectures. Second, our approach relies on the availability of ground-truth trajectories for turn-level supervision. This poses a challenge in highly open-ended scenarios such as “deep research” tasks, where ground-truth parameters and reasoning steps are inherently difficult to determine or verify. Consequently, adapting our method to such problems remains a promising but challenging direction for future research.

Acknowledgements

This work was funded by National Key Research and Development Program of China (2023YFA1008704), National Natural Science Foundation of China (62472426, 62276248), fund for building world-class universities (disciplines) of Renmin University of China. Work partially done at Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, and Beijing Key Labo-

ratory of Research on Large Models and Intelligent Governance.

References

- Qikai Chang, Zhenrong Zhang, Pengfei Hu, Jun Du, Jiefeng Ma, Yicheng Pan, Jianshu Zhang, Quan Liu, and Jianqing Gao. 2025. Thor: Tool-integrated hierarchical optimization via rl for mathematical reasoning. *arXiv preprint arXiv:2509.13761*.
- Cong Chen, Kaixiang Ji, Hao Zhong, Muzhi Zhu, Anzhou Li, Guo Gan, Ziyuan Huang, Cheng Zou, Jiajia Liu, Jingdong Chen, Hao Chen, and Chunhua Shen. 2025. [Gui-shepherd: Reliable process reward and verification for long-sequence gui tasks](#). *Preprint*, arXiv:2509.23738.
- Marco Cuturi. 2013. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13*, page 2292–2300, Red Hook, NY, USA. Curran Associates Inc.
- Debrup Das, Debopriyo Banerjee, Somak Aditya, and Ashish Kulkarni. 2024. Mathsensei: a tool-augmented large language model for mathematical reasoning. In *In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.
- Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, and 1 others. 2025. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*.
- Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. 2025a. Retool: Reinforcement learning for strategic tool use in llms. *arXiv preprint arXiv:2504.11536*.
- Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025b. [Group-in-group policy optimization for LLM agent training](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Guo Gan, Yuxuan Ding, Cong Chen, Yuwei Ren, Yin Huang, and Hong Zhou. 2026. [Android coach: Improve online agentic training efficiency with single state multiple actions](#). *Preprint*, arXiv:2604.07277.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Minlie Huang, Nan Duan, Weizhu Chen, and 1 others. 2024. Tora: A tool-integrated reasoning agent for mathematical problem solving. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.
- Dongfu Jiang, Yi Lu, Zhuofeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou,

- Chao Du, and 1 others. 2025. Verltool: Towards holistic agentic reinforcement learning with tool use. *arXiv preprint arXiv:2509.01055*.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Serkan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-rl: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*.
- Harold W Kuhn. 1956. Variants of the hungarian method for assignment problems. *Naval research logistics quarterly*, 3(4):253–258.
- Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025a. Torl: Scaling tool-integrated rl. *arXiv preprint arXiv:2503.23383*.
- Yuchen Li, Hengyi Cai, Rui Kong, Xinran Chen, Jiamin Chen, Jun Yang, Haojie Zhang, Jiayi Li, Jiayi Wu, Yiqun Chen, and 1 others. 2025b. Towards ai search paradigm. *arXiv preprint arXiv:2506.17188*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *The Twelfth International Conference on Learning Representations*.
- Heng Lin and Zhongwen Xu. 2025. Understanding tool-integrated reasoning. *arXiv preprint arXiv:2508.19201*.
- Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. The berkeley function calling leaderboard (bfc): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*.
- Cheng Qian, Emre Can Acikgoz, Qi He, Hongru WANG, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2025. [ToolRL: Reward is all tool learning needs](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2024. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.
- Changle Qu, Sunhao Dai, Hengyi Cai, Yiyang Cheng, Jun Xu, Shuaiqiang Wang, and Dawei Yin. 2025a. Uplift-rag: Uplift-driven knowledge preference alignment for retrieval-augmented generation. *Findings of the Association for Computational Linguistics: EMNLP*.
- Changle Qu, Sunhao Dai, Xiaoqi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2024. Towards completeness-oriented tool retrieval for large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*.
- Changle Qu, Sunhao Dai, Xiaoqi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025b. [From exploration to mastery: Enabling LLMs to master tools via self-driven interactions](#). In *The Thirteenth International Conference on Learning Representations*.
- Changle Qu, Sunhao Dai, Xiaoqi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025c. Tool learning with large language models: A survey. *Frontiers of Computer Science*, 19(8):198343.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.
- Hieu Tran, Zonghai Yao, and Hong Yu. 2025. Exploiting tree structure for credit assignment in rl training of llms. *arXiv preprint arXiv:2509.18314*.
- Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. 2025a. Information gain-based policy optimization: A simple and effective approach for multi-turn llm agents. *arXiv preprint arXiv:2510.14967*.
- Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. 2025b. Spa-rl: Reinforcing llm agents via stepwise progress attribution. *arXiv preprint arXiv:2505.20732*.
- Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. 2025c. Otc: Optimal tool calls via reinforcement learning. *arXiv e-prints*, pages arXiv–2504.
- Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2024a. [Mathcoder: Seamless code integration in LLMs for enhanced mathematical reasoning](#). In *The Twelfth International Conference on Learning Representations*.
- Shuai Wang, Weiwen Liu, Jingxuan Chen, Yuqi Zhou, Weinan Gan, Xingshan Zeng, Yuhan Che, Shuai Yu, Xinlong Hao, Kun Shao, and 1 others. 2024b. Gui agents with foundation models: A comprehensive survey. *arXiv preprint arXiv:2411.04890*.
- Yifan Wei, Xiaoyan Yu, Yixuan Weng, Tengfei Pan, Angsheng Li, and Li Du. 2025. Autotir: Autonomous tools integrated reasoning via reinforcement learning. *arXiv preprint arXiv:2507.21836*.

- Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. 2025. Simpletir: End-to-end reinforcement learning for multi-turn tool-integrated reasoning. *arXiv preprint arXiv:2509.02479*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Shuo Yang, Caren Han, Yihao Ding, Shuhe Wang, and Eduard Hovy. 2026a. [Tooltree: Efficient LLM tool planning via dual-feedback monte carlo tree search and bidirectional pruning](#). In *The Fourteenth International Conference on Learning Representations*.
- Shuo Yang, Soyeon Caren Han, Xueqi Ma, Yan Li, Mohammad Reza Ghasemi Madani, and Eduard Hovy. 2026b. Evotool: Self-evolving tool-use policy optimization in llm agents via blame-aware mutation and diversity-aware selection. *arXiv preprint arXiv:2603.04900*.
- Junjie Ye, Zhengyin Du, Xuesong Yao, Weijian Lin, Yufei Xu, Zehui Chen, Zaiyuan Wang, Sining Zhu, Zhiheng Xi, Siyu Yuan, and 1 others. 2025a. Toolhop: A query-driven benchmark for evaluating large language models in multi-hop tool use. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2995–3021.
- Junjie Ye, Changhao Jiang, Zhengyin Du, Yufei Xu, Xuesong Yao, Zhiheng Xi, Xiaoran Fan, Qi Zhang, Tao Gui, Xuanjing Huang, and 1 others. 2025b. Feedback-driven tool-use improvements in large language models via automated build environments. *arXiv preprint arXiv:2508.08791*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025a. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Yuanqing Yu, Zhefan Wang, Weizhi Ma, Shuai Wang, Chuhan Wu, Zhiqiang Guo, and Min Zhang. 2025b. Steptool: Enhancing multi-step tool usage in llms via step-grained reinforcement learning. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 3952–3962.
- Chuhuai Yue, Chengqi Dong, Yinan Gao, Hang He, Jiajun Chai, Guojun Yin, and Wei Lin. 2025. Promoting efficient reasoning with verifiable stepwise reward. *arXiv preprint arXiv:2508.10293*.
- Siliang Zeng, Quan Wei, William Brown, Oana Frunza, Yuriy Nevmyvaka, Yang Katie Zhao, and Mingyi Hong. 2025a. Reinforcing multi-turn reasoning in LLM agents via turn-level credit assignment. In *ICML 2025 Workshop on Computer Use Agents*.
- Yirong Zeng, Xiao Ding, Yutai Hou, Yuxian Wang, Li Du, Juyi Dai, Qiuyang Ding, Duyu Tang, Dandan Tu, Weiwen Liu, and 1 others. 2025b. Tool zero: Training tool-augmented llms via pure rl from scratch. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9135–9147.
- Shaokun Zhang, Yi Dong, Jieyu Zhang, Jan Kautz, Bryan Catanzaro, Andrew Tao, Qingyun Wu, Zhiding Yu, and Guilin Liu. 2025a. Nemotron-research-tool-n1: Tool-using language models with reinforced reasoning. *arXiv preprint arXiv:2505.00024*.
- Wenlin Zhang, Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. 2025b. Process vs. outcome reward: Which is better for agentic RAG reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Yaocheng Zhang, Haohuan Huang, Zijun Song, Yuanheng Zhu, Qichao Zhang, Zijie Zhao, and Dongbin Zhao. 2025c. Criticsearch: Fine-grained credit assignment for search agents via a retrospective critic. *arXiv preprint arXiv:2511.12159*.
- Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025a. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.
- Xuhui Zheng, Kang An, Ziliang Wang, Yuhang Wang, and Yichao Wu. 2025b. [StepSearch: Igniting LLMs search ability via step-wise proximal policy optimization](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21816–21841, Suzhou, China. Association for Computational Linguistics.
- Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025c. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*.
- Yuqi Zhou, Sunhao Dai, Changle Qu, Liang Pang, Jun Xu, and Ji-Rong Wen. 2026. Learning to retrieve from agent trajectories. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Appendix

A Benchmark Details

In this section, we introduce the detailed description of the benchmarks used in this paper. The statistics of these datasets are shown in Table 4.

FTRL.¹ The Feedback-Driven Tool-Use Reinforcement Learning (FTRL) (Ye et al., 2025b) benchmark is designed for evaluating tool-integrated reasoning under automatically constructed local execution environments. To mirror the complexity of real-world interactions, FTRL categorizes tool-use scenarios into four distinct levels of difficulty:

- **Single-Hop:** Contains a single query that can be resolved with a one-time tool invocation.
- **Parallel Single-Hop:** Involves multiple independent sub-queries, allowing the agent to invoke multiple tools in parallel to gather information.
- **Multi-Hop:** Consists of a sequence of dependent sub-queries, where the execution of a subsequent tool depends on the output of previous steps.
- **Parallel Multi-Hop:** A hybrid and complex structure that includes both independent and interdependent sub-queries, requiring advanced logical planning and orchestration.

To provide a multidimensional assessment of model performance, FTRL employs the following verifiable metrics:

- **Solve-P:** Measures the precision of the tool invocations. It is defined as the ratio of valid tool calls to the total number of calls made:

$$\text{Solve-P} = \begin{cases} \frac{q}{p}, & \text{if } p > 0 \\ 1, & \text{if } p = 0 \end{cases} \quad (12)$$

where p is the number of tools invoked, q is the number of sub-questions successfully solved.

- **Solve-R:** Evaluates the completeness of task execution by measuring the proportion of successfully resolved sub-tasks. The formula is:

$$\text{Solve-R} = \frac{q}{n} \quad (13)$$

where n is the number of sub-questions in total.

Solve-F1: Evaluates the harmonic mean of Solve-P and Solve-R, providing an overall performance measure:

$$\text{Solve-F1} = \frac{2 \cdot \text{Solve-P} \cdot \text{Solve-R}}{\text{Solve-P} + \text{Solve-R}} \quad (14)$$

¹<https://github.com/bytedance/FTRL>

In this work, we use FTRL as the in-domain benchmark for both training and evaluation. Its structured feedback and multi-turn interaction setting make it particularly suitable for studying fine-grained credit assignment and turn-level supervision in tool-integrated reasoning.

BFCL.² The Berkeley Function Calling Leaderboard (BFCL) (Patil et al., 2025) is a comprehensive and rigorous benchmark designed to evaluate the function-calling capabilities of LLMs across a wide range of scenarios. To assess the performance of models in complex agentic settings, BFCL V3 and V4 introduce several critical task dimensions:

- **Multi-Turn:** This dimension evaluates the model’s ability to maintain state and handle dynamic decision-making over multiple conversational turns. It consists of four subsets: **Base**, which provides standard multi-turn interactions; **Missing Parameters** and **Missing Functions**, which test robustness when information is incomplete; and **Long-Context**, which requires extracting tool information from extensive distractor text.
- **Agentic:** This dimension focuses on the model’s capacity for autonomous reasoning and interaction within a stateful environment. It includes two subsets: **Web Search**, requiring real-time information retrieval, and **Memory**, testing the ability to recall and utilize information from previous interactions.

BFCL employs a multi-layered evaluation strategy to ensure execution rigor:

- **Multi-turn:** Requires passing two per-turn checks for overall success: (1) **State-based Evaluation**, which verifies backend system state changes, and (2) **Response-based Evaluation**, which validates the execution path against minimal viable result paths for read-only requests.

- **Agentic:** Employs an exact-match metric isolated to the “answer” field. This prevents false positives from incidental phrasing in unstructured text, ensuring the model commits to a precise and intended response.

ToolHop.³ ToolHop (Ye et al., 2025a) is a query-driven benchmark specifically designed to evaluate LLMs in multi-hop tool-use scenarios. It features 995 user queries and 3,912 locally executable tools, focusing on the model’s ability to handle complex

²<https://github.com/ShishirPatil/gorilla/blob/main/berkeley-function-call-leaderboard>

³<https://huggingface.co/datasets/bytedance-research/ToolHop>

Datasets	Train					Test						
	FTRL	FTRL				BFCL						ToolHop
		Single-Hop	Parallel SH	Multi-Hop	Parallel MH	Base	Miss Func	Miss Param	Long Context	Web Search	Memory	
#Number	2,215	50	50	50	50	200	200	200	200	200	465	995

Table 4: Detailed statistics of the experimental datasets, including training data from FTRL and test sets across FTRL, BFCL, and ToolHop benchmarks.

interdependencies.

ToolHop focuses on the accuracy of the final answer through a verifiable execution environment: **Answer Correctness (AC)**: The primary metric that measures the proportion of queries for which the model provides a correct final answer, verified against ground-truth labels.

B More Implementation Details

The detailed hyper-parameter configurations for our experiments are summarized in Table 5.

Category	Hyper-parameter
Data Configuration	
Train Batch Size	256
Validation Batch Size	256
Max Prompt Length	7000
Max Response Length	23000
Optimization	
Learning Rate	1e-6
PPO Mini Batch Size	32
KL-Divergence loss coefficient	0.001
Rollout Configuration	
Rollout Name	vllm
Rollout Temperature	1.0
GPU Memory Utilization	0.7
Number of Rollouts	16
Training & Logging	
Save Frequency	8
Test Frequency	4
Total Epochs	3

Table 5: Training Configuration.

In addition to the standard training setup, we provide further clarifications on the implementation of our reward and evaluation mechanisms. Our turn-level advantage estimation is performed by comparing rollouts within the same group at each specific turn t . Given that trajectories in a group may have varying lengths, we implement the following logic to ensure robust estimation: For a given turn t , the advantage is computed only among the subset of rollouts that actually reach or exceed t turns. In scenarios where only a single rollout within a group contains the t -th turn, we provide a fallback

Advantage Estimation	FTRL			
	Solve-P	Solve-R	Solve-FI	Avg.
Weighted Product (KM)	33.10	42.38	34.62	36.70
Weighted Sum (KM)	29.95	40.71	32.77	34.48
MatchTIR (KM)	36.33	44.18	37.33	39.28

Table 6: Additional analysis on advantage estimation on the FTRL dataset using Qwen3-8B.

by setting the mean to 0 and the variance to 1 for normalization purposes. This prevents numerical instability and ensures that the single available data point does not lead to skewed gradients. For the evaluation of the **Web Search** subset within the BFCL benchmark, we utilize the Bing Search ⁴ API as the underlying search engine.

C More Experiments

We further perform additional experiments on FTRL with Qwen3-8B for deeper analysis, which are presented here due to space constraints.

C.1 Analysis on Advantage Estimation

To further investigate the mechanism of turn-level credit assignment, we conduct supplementary experiments using the Qwen3-8B model on the FTRL benchmark. We compare MatchTIR with an alternative intuition-based approach, referred to here as the Weighted Intra-trajectory Method.

Weighted Intra-trajectory Method. To further explore the mechanisms of turn-level credit assignment, we investigate two alternative variants that estimate advantages within an individual trajectory (intra-trajectory) rather than across different rollouts (inter-group). For both variants, we first compute a local turn-level advantage $A_{i,t}^{loc}$ by treating each turn as an independent sample within the same rollout: $A_{i,t}^{loc} = \frac{R_{i,t} - \text{mean}(\{R_{i,j}\}_{j=1}^T)}{\text{std}(\{R_{i,j}\}_{j=1}^T)}$. Based on $A_{i,t}^{loc}$, we propose two weighting strategies to adjust the global trajectory advantage A_i^g :

- **Weighted Product:** This method uses the local advantage as a scaling factor to modulate the

⁴<https://serpapi.com/bing-search-api>

Cost Matrix	FTRL			
	Solve-P	Solve-R	Solve-F1	Avg.
Exponential Transformation	30.96	41.50	33.38	35.28
Normalized Inversion	32.77	42.88	34.91	36.85
Linear Inversion	33.61	42.56	33.61	36.59

Table 7: Additional analysis on cost matrix construction of OT on the FTRL dataset using Qwen3-8B.

magnitude of the global signal. The final adjusted advantage $\hat{A}_{i,t}$ for each token in turn t is defined as $\hat{A}_{i,t} = \left(1 + 0.1 \cdot \text{sgn}(A_i^g) \cdot A_{i,t}^{loc}\right) A_i^g$, where $\text{sgn}(\cdot)$ ensures the weighting direction aligns with the trajectory performance.

- **Weighted Sum:** This method directly incorporates the local turn-level signal by adding it to the global advantage: $\hat{A}_{i,t} = A_i^g + A_{i,t}^{loc}$

Results and Discussion. As shown in Table 6, the results indicate that this Weighted Intra-trajectory Method approach yields inferior performance compared to MatchTIR. We attribute this to the fact that intra-trajectory normalization inherently overlooks the temporal interdependencies between turns. In multi-turn TIR tasks, the success of a later turn is often contingent upon the precise execution of earlier tool calls. By solely focusing on the internal variance of a single rollout, the weighted method fails to capture the relative quality of the reasoning path across different potential strategies, whereas our inter-group discounted approach provides a higher-resolution signal for mastering long-horizon interactions.

C.2 Analysis on Cost Matrix Construction

Since the cost matrix C in our Optimal Transport (OT) framework is derived from the similarity matrix S , the transformation function applied to S significantly influences the final turn-level reward assignment. To evaluate this impact, we compare three distinct construction methods:

- **Linear Inversion:** $C = -S$. This maintains the original linear distribution of similarity scores.

- **Normalized Inversion:** $C = 1 - S_{\text{norm}}$, where $S_{\text{norm}} = \frac{S - \min(S)}{\max(S) - \min(S) + \epsilon}$.

- **Exponential Transformation:** $C = -e^S$. This non-linear mapping amplifies the differences between high-similarity pairs and others.

Results and Discussion. As shown in Table 7, the results indicate that the model performance is relatively insensitive to the choice between Linear Inversion and Normalized Inversion. This consis-

Methods	FTRL			
	Solve-P	Solve-R	Solve-F1	Avg.
<i>FTRL</i>				
Single-Turn	32.32	38.87	32.85	34.68
Expanded Single-Turn	35.34	43.13	35.30	37.92
Multi-Turn	33.28	40.60	34.20	36.03
<i>ToolRL</i>				
Single-Turn	25.57	35.31	26.72	29.20
Expanded Single-Turn	30.23	27.73	27.28	28.41
Multi-Turn	30.64	40.07	32.51	34.41
<i>MatchTIR</i>				
Multi-Turn	36.33	44.18	37.33	39.28

Table 8: Performance comparison between multi-turn training and expanded single-turn baselines on FTRL and ToolRL using Qwen3-8B.

tency suggests that the raw similarity scores generated by our matching module are already well-distributed, and linear re-scaling does not alter the relative matching order. In contrast, we observe a performance drop with the Exponential Transformation. We hypothesize that this is primarily due to numerical instability and noise amplification. The exponential function excessively magnifies minor fluctuations in similarity scores, creating a cost matrix with values spanning several orders of magnitude. This not only hinders the convergence of the Sinkhorn solver due to precision errors during iterative normalization but also leads to high-variance advantage that destabilize the policy gradient.

C.3 Multi-Turn vs. Expanded Single-Turn

While our main results demonstrate the superiority of multi-turn training, one could argue that this advantage stems from increased computational expenditure during inference and training. To ensure a fair comparison, we conduct an experiment by decomposing multi-turn trajectories into expanded single-turn training sets. Specifically, a trajectory consisting of T turns is decomposed into $T - 1$ independent training samples. For each sample at turn t , the model is provided with the conversation history of the preceding t turns and tasked with predicting the $(t + 1)$ -th tool-calling action. This data expansion ensures that the single-turn baseline is exposed to the same volume of state-transition information as the multi-turn approach.

Results and Discussion. As shown in Table 8, while increasing the data volume via decomposition improves performance on the **FTRL** benchmark, it still falls short of our end-to-end multi-turn optimization. Interestingly, on the **ToolRL** bench-

```
Prompt Template for Training

System Prompt:
# Tools
You may call one or more functions to assist with the
user query.
You are provided with function signatures within
<tools></tools> XML tags:
<tools>
{Tool List}
</tools>
For each function call, return a json object
with the function name and arguments within
<tool_call></tool_call> XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-
json-object>}
</tool_call>
User Prompt:
Please call the given tools to answer the question.
Please note that all your information must be obtained
by calling tools and not by answering the question
directly.
If the call fails, you need to try to correct it and con-
tinue until you arrive at an answer.
Only output the final answer (in words, numbers, or
phrases) inside the <answer></answer> tag, without
any explanations or extra information.
Question: {question}
Output:
```

provided in FTRL (Ye et al., 2025b), the prompt template for training is shown in Figure 5.

Figure 5: Prompt template used in our experiments.

mark, we observe that increasing data through single-turn decomposition actually leads to a performance decline. We attribute this to the inherent limitations of the single-turn paradigm: Single-turn models rely heavily on the immediate reward design for each isolated step. Without the context of the entire trajectory, they struggle to reconcile local actions with long-term goals. The single-turn format restricts the potential for sophisticated reward and advantage estimation, which requires a holistic view of the interaction. These findings suggest that end-to-end multi-turn optimization is fundamentally more effective, as it allows the model to capture inter-turn dependencies that simple data expansion in a single-turn setting cannot replicate.

D Case Study

In this section, we present a case study to demonstrate the rationality of the turn-level rewards generated by MatchTIR. As shown in Table 9, this example specifically highlights the distinction between the hard and soft variants.

E Prompt

In this section, we present the prompts used in our experiments. Specifically, we follow the template

Question: What is the common building material of the town where the architect who designs buildings inspired by the architecture of the castle on the hill overlooking the valley?

Ground Truth Tool Calls: [{"name": "valley_hill_analyzer", "parameters": {}}, {"name": "landmark_locator", "parameters": {"landmark_type": "castle", "geographic_feature": "Yosemite Valley", "position_relation": "on the hill"}}, {"name": "architect_inspiration_analyzer", "parameters": {"landmark": "hearth castle"}}, {"name": "person_locator", "parameters": {"name": "Lily Chen"}}, {"name": "building_material_analyzer", "parameters": {"location": "Santa Barbara"}}]

Ground Truth Answer: Stone.

Turn 1:

<think> Okay, let me try to figure out how to answer this question. </think>

<tool_call> {"name": "landmark_locator", "arguments": {"landmark_type": "castle", "geographic_feature": "valley", "position_relation": "overlooking"}} </tool_call>

<tool_response> {"result": "No specific landmark found matching the criteria"} </tool_response>

[WRONG]

Reward Given by MatchTIR (KM): 0

Reward Given by MatchTIR (OT): 0.025

Turn 2:

<think> Okay, the first tool call didn't return a specific landmark. Maybe I need to </think>

<tool_call> {"name": "valley_hill_analyzer", "arguments": {}} </tool_call>

<tool_response> {"result": "Yosemite Valley has a hill overlooking it."} </tool_response>

[CORRECT]

Reward Given by MatchTIR (KM): 1.0

Reward Given by MatchTIR (OT): 0.1667

Turn 3:

<think> Okay, so the valley_hill_analyzer returned Yosemite Valley with a hill overlooking it. Now, the next step is to </think>

<tool_call> {"name": "landmark_locator", "arguments": {"landmark_type": "castle", "geographic_feature": "Yosemite Valley", "position_relation": "on the hill"}} </tool_call>

<tool_response> {"result": "Hearth Castle"} </tool_response>

[CORRECT]

Reward Given by MatchTIR (KM): 1.0

Reward Given by MatchTIR (OT): 0.1581

Turn 4:

<think> Okay, so the Hearst Castle is located on the hill overlooking Yosemite Valley. Now, the next step is to </think>

<tool_call> "name": "architect_inspiration_analyzer", "arguments": {"landmark": "Hearth Castle"} </tool_call>

<tool_response> Lily Chen </tool_response>

[CORRECT]

Reward Given by MatchTIR (KM): 1.0

Reward Given by MatchTIR (OT): 0.1667

Turn 5:

<think> Okay, so the architect inspired by Hearst Castle is Lily Chen. Now, I need to find out </think>

<tool_call> "name": "person_locator", "arguments": {"name": "Lily Chen"} </tool_call>

<tool_response> Santa Barbara </tool_response>

[CORRECT]

Reward Given by MatchTIR (KM): 1.0

Reward Given by MatchTIR (OT): 0.1667

Turn 6:

<think> Okay, so Lily Chen is located in Santa Barbara. Now, the next step is to </think>

<tool_call> "name": "building_material_analyzer", "arguments": {"location": "Santa Barbara"} </tool_call>

<tool_response> "result": "Analyzing building materials in Santa Barbara for modern period. Building type: all types. Environmental impact not considered. Availability considered. Technological advancements not considered. Cultural influences not considered. Economic factors not considered. Common building material: Stone." </tool_response>

[CORRECT]

Reward Given by MatchTIR (KM): 1.0

Reward Given by MatchTIR (OT): 0.1667

Turn 7:

<think> The common building material in Santa Barbara, where the architect Lily Chen (inspired by Hearst Castle) is located, is stone. </think>

<answer> Stone. </answer>

[CORRECT]

Outcome Reward: 1.0

Table 9: A case study of turn-level reward given by MatchTIR (KM) and MatchTIR (OT).