

ET-Agent: Incentivizing Effective Tool-Integrated Reasoning Agent via Behavior Calibration

Yifei Chen, Guanting Dong, Zhicheng Dou[†]

Gaoling School of Artificial Intelligence, Renmin University of China
zhangboguodong@ruc.edu.cn, dou@ruc.edu.cn

Abstract

Large Language Models (LLMs) can extend their parameter knowledge limits by adopting the Tool-Integrated Reasoning (TIR) paradigm. However, existing LLM-based agent training framework often focuses on answers' accuracy, overlooking specific alignment for behavior patterns. Consequently, agent often exhibits ineffective actions during TIR tasks, such as redundant and insufficient tool calls. How to calibrate erroneous behavioral patterns when executing TIR tasks, thereby exploring effective trajectories, remains an open-ended problem. In this paper, we propose ET-Agent, a training framework for calibrating agent's tool-use behavior through two synergistic perspectives: Self-evolving Data Flywheel and Behavior Calibration Training. Specifically, we introduce a self-evolutionary data flywheel to generate enhanced data, used to fine-tune LLM to improve its exploration ability. Based on this, we implement an two-phases behavior-calibration training framework. It is designed to progressively calibrate erroneous behavioral patterns to optimal behaviors. Further in-depth experiments confirm the superiority of ET-Agent across multiple dimensions, including correctness, efficiency, reasoning conciseness, and tool execution accuracy. Our ET-Agent framework provides practical insights for research in the TIR field. Codes can be found in <https://github.com/asilverlight/ET-Agent>

1 Introduction

Recently, Large Language Models (LLMs) have demonstrated excellent performance in various complex reasoning tasks (Huang and Chang, 2023; Shao et al., 2024; Xie et al., 2025a; Li et al., 2024b; Li and Wang, 2025; Seed; Qiao et al., 2024). However, when faced with problems beyond their own knowledge and capabilities, LLMs often struggle to perform effectively (Min et al., 2024; Qin et al., 2024; Chen et al., 2023). As an emerging paradigm, Tool-Integrated Reasoning (TIR) has

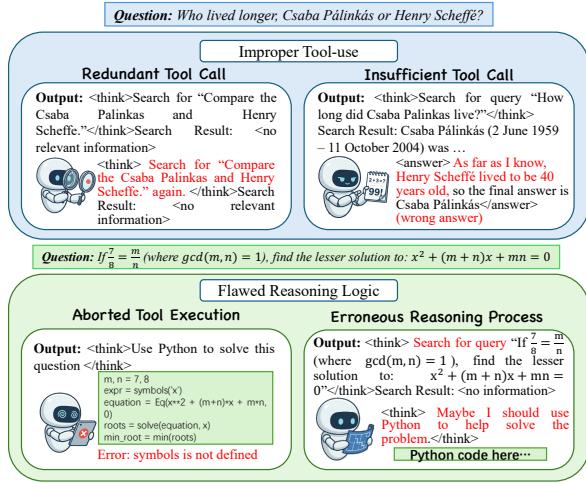


Figure 1: Illustration of two different error tool-use behavior patterns.

recently received increasing attention (Gou et al., 2024; Li et al., 2025c,h). LLM-based agents with TIR paradigm can achieve better performance than ordinary reasoning models in tasks such as mathematical reasoning, in-depth search, and scientific question-answering (Li et al., 2025h,e; Dong et al., 2025b).

Nevertheless, most current works only focus on the accuracy of TIR paradigm in downstream tasks. There is a lack of necessary attention to TIR's behavioral patterns correctness. For example, some works excessively interact with retrieval systems to achieve deeper and more accurate information acquisition (Gao et al., 2025b; Chen et al., 2025a; Wu et al., 2025a; Li et al., 2025d). But these methods undoubtedly introduce redundant tool calls, making it more difficult to practically apply these methods.

To optimize TIR behavioral patterns, several works elaborately design data construction processes, aiming to calibrate agents' behaviors to be more efficient (Su et al., 2025; Qian et al., 2025; Li et al., 2024a, 2025d; Gou et al., 2024). Another branch of efforts start with Reinforcement Learning

(RL) algorithms to stimulate LLM’s potential to effectively utilize tools (Huang et al., 2025; Wang et al., 2025b; Wu et al., 2025c; Wang et al., 2025a; Tao et al., 2025a). However, these works only focus on mitigating redundant tool calls, which is very limited. In real TIR scenarios, there are still other erroneous behavioral patterns that need to be calibrated. Tool-Light pioneers the investigation into broader issues such as tool call insufficiency and reasoning redundancy (Chen et al., 2025c). Nevertheless, Tool-Light employs the DPO algorithm. This binary comparison optimization can cause the model output to collapse into a very narrow action space, resulting in limited exploration. It is worth noting that TIR scenarios feature an extremely extensive tool-use action space, which can give rise to a multitude of potential trajectories (Lin et al., 2025; Shi et al., 2024; Naik et al., 2024; Yu et al., 2025a; Dong et al., 2025d). Relying solely on imitation learning and binary comparison optimization, it is difficult to fully calibrate the agent’s behavioral patterns. Therefore, how to utilize sufficient exploration within the TIR action space to effectively calibrate behavioral patterns via on-policy training remains an open challenge.

In this paper, we first conduct a comprehensive investigation into erroneous behavioral patterns within TIR tasks. By summarizing previous works and conducting quantitative analysis in preliminary experiments (He et al., 2025; Xiong et al., 2025; Sun et al., 2024; Yin et al., 2025), we categorize these error patterns into two primary aspects: *improper tool-use* and *flawed reasoning logic*, as shown in Figure 1. Motivated by this analysis, we propose ET-Agent, an iterative, behavior-calibration training framework optimizing agent’s TIR behavioral patterns at both the data and algorithmic levels. From data perspective, we introduce a **Self-evolving Data Flywheel**. We carefully design methods to augment existing trajectories. Training on these samples can broaden the tool-use action space coverage, enabling the agent to fully explore tool-use action space.

From an algorithmic perspective, we propose a **Behavior Calibration Training** framework with two phases. First, we implement Action Space Exploration Fine-tuning with the data from the flywheel, aiming to broaden the model’s exploration of the tool-use action space. Based on this phase, we implement an Iterative Behavior Calibration Reinforcement Learning phase. This process alternates between Group-wise Pareto Sampling and

Curriculum RL Training. The former identifies the most helpful data for training, while the latter gradually calibrates agent’s actions toward optimal and standardized behavioral patterns.

To verify the effectiveness of ET-Agent, we test multiple metrics on six challenging tasks. Overall, the results show that ET-Agent can achieve best performance in all dimensions, indicating the success of TIR behavior calibration. In summary, our contributions are as follows:

- We provide a comprehensive quantitative analysis of erroneous behavioral patterns in TIR. Inspired by this, we propose ET-Agent, a framework for optimizing TIR’s behavioral patterns.
- We introduce a self-evolving data flywheel, an iterative loop where the model continuously refines its previous trajectories. This mechanism effectively unfolds the model’s action space coverage beyond its initial exploration.
- Based on the flywheel, we present a behavior calibration training framework with two phases, aiming to calibrate the model’s exploration in tool-use action space to optimal trajectories.
- Extensive experiments demonstrate that ET-Agent substantially improves behavioral efficiency, reasoning conciseness, and execution success rates while maintaining high accuracy.

2 Related Work

Tool-Integrated Reasoning. Tool-Integrated Reasoning (TIR) empowers LLMs to solve complex tasks via external tools (Sui et al., 2025; Su et al., 2025; Fang et al., 2025; Li et al., 2025d, 2024c; Dong et al., 2025e). Existing works generally endow LLM-based agents with TIR ability through: (1) **Inference-time Scaling**, focusing on prompt engineering or test-time computation scaling methods (Li et al., 2025e; Wu et al., 2025b; Lu et al., 2025; Yao et al., 2023); and (2) **Training-based Method**, optimizing LLM policies using SFT, DPO, or RL methods (Li et al., 2024a, 2025c; Zhang et al., 2025; Li et al., 2025a; Zhao et al., 2026). For example, works like Search-o1 guide agents to complete TIR tasks by designing workflows (Li et al., 2025e), while works like ToRL fully enhance agents’ capabilities using RL methods (Li et al., 2025h). Currently, most works only focus on prioritizing accuracy, and only a few works pay attention

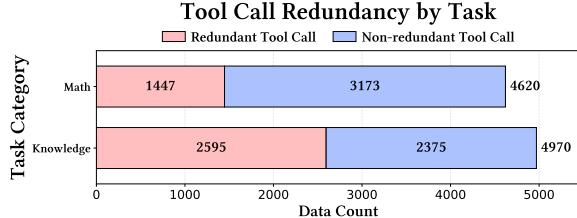


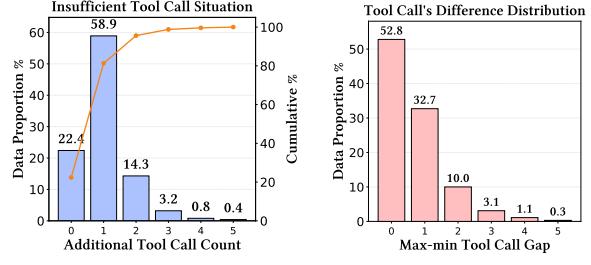
Figure 2: Statistics about redundant tool calls.

to behavioral patterns’ effectiveness such as the efficiency of TIR. In contrast, ET-Agent optimizes the behavioral patterns in TIR more comprehensively, using the on-policy training method to progressively calibrate TIR behavioral patterns.

Agentic Reinforcement Learning. Recently, agentic reinforcement learning (RL) method plays an important role in TIR tasks (Zhang et al., 2025; Li et al., 2025a). Early works focus on using RL methods to accomplish specific domain tasks (Li et al., 2025h; Jin et al., 2025a; Song et al., 2025; Li et al., 2025b). Building on this, some works make breakthroughs in more difficult TIR tasks or successfully expand the domain of TIR tasks (Li et al., 2024c; Chen et al., 2025d; Gao et al., 2025b; Chen et al., 2025a). For example, a series of works by Tongyi DeepResearch successfully improve agents’ performance in deep search tasks (Wu et al., 2025a; Tao et al., 2025a; Li et al., 2025d; Tao et al., 2025b). Works such as Tool-Star and SMART enhance the performance of TIR agents in tasks across a wider range of domains (Dong et al., 2025b; Qian et al., 2025; Wei et al., 2025; Huang et al., 2025; Qiao et al., 2025; Li et al., 2025f). Recently, some RL algorithms or frameworks specifically adapted for agents have emerged (Dong et al., 2025a; Feng et al., 2025; Hou et al., 2025; Li et al., 2025i). For instance, ARPO optimizes the branching strategy of GRPO from the perspective of information entropy, while VerlTool designs a training framework suitable for tool use (Dong et al., 2025d; Jiang et al., 2025). Inspired by these works, ET-Agent successfully uses RL methods to optimize agents’ tool-use behavioral patterns.

3 Preliminary Experiment

After summarizing the TIR behavioral patterns in previous works (He et al., 2025; Xiong et al., 2025; Sun et al., 2024), we categorize erroneous behavioral patterns into two types:



(a) Additional tool calls needed for modification. (b) Distribution of different trajectories’ tool calls.

Figure 3: Distribution of additional tool calls required to modify incorrect outputs, and differences in tool call across trajectories for identical questions.

- **Improper Tool-Use:** Incorrect tool triggering methods affect downstream tasks’ performance. This includes (1) **Redundant Tool Calls**, where tool invocations fail to deliver additional useful information for solving the problem; and (2) **Aborted Tool Execution**, wherein tool execution is terminated due to improperly formulated tool calls generated by the agent—such as null query or defective code.
- **Flawed Reasoning Logic:** Errors originating from defective cognitive processes. This includes (1) **Insufficient Tool Calls**, where the agent does not realize that additional tool calls are needed, leading to an incorrect result; and (2) **Erroneous Reasoning Process**, which involves logically irrelevant or erroneous steps derived from flawed planning.

Analysis of Improper Tool-Use. We first analyze redundant tool call situations. By employing GPT-4o (Hurst et al., 2024) to identify redundant tool calls in outputs, as shown in Figure 2, we observe a high prevalence of redundancy across both tasks, which highlights that uncalibrated agents struggle to call tools efficiently. We also analyze Aborted Tool Execution situations. As shown in Figure 4, the majority of trajectories involve only one time error, while a small number of samples have errors occurring more than once. It suggests that improving triggering tool call accuracy through training is highly feasible.

Analysis of Flawed Reasoning Logic. To investigate insufficient tool call situations, we inject hints at the end of incorrect trajectories to induce continued reasoning. As shown in Figure 3(a), a considerable number of trajectories can be modified after continued reasoning. It suggests that

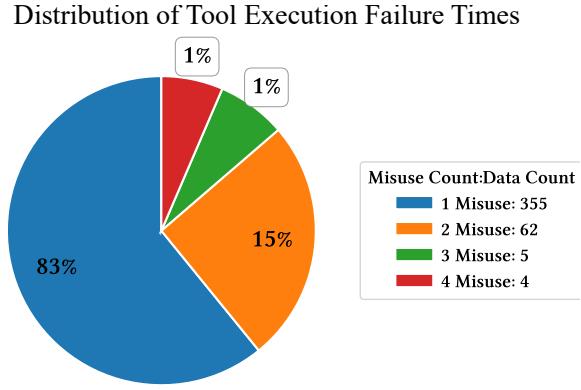


Figure 4: Statistics on Aborted Tool Execution Cases.

it is hard for the agent to invoke additional tools, resulting in the “premature closure” issue.

Analysis of Trajectories’ Diversity. Finally, we investigate the intrinsic complexity of the TIR action space by analyzing the diversity among correct solutions. Figure 3(b) reveals substantial disparities in tool call frequency even across correct outputs. This observation confirms that the action space for TIR tasks is extensive. Agents without calibration can hardly explore an effective reasoning trajectory.

Full dataset details and error case studies are in Appendix A.1 and F, and full experiment details can be found in Appendix C.1. Examples for each type of error can be found in Appendix F.

4 The Proposed Method: ET-Agent

We introduce the ET-Agent framework, aiming to calibrate LLM-based agent’s TIR behavioral correctness. As shown in Figure 5, ET-Agent comprises two primary aspects: **(1) Self-Evolving Data Flywheel** obtains rich and diverse training data through self-corrective refinement or expanded exploration to iteratively optimize trajectories. **(2) Behavior Calibration Training** conducts a two-phase training framework to calibrate TIR behavioral patterns.

4.1 Self-Evolving Data Flywheel

In order to more effectively calibrate the agent’s behavior to the optimal patterns, it is necessary for the agent to fully explore the action space. Inspired by LLM Self-Evolution and Test-Time Scaling methods (Gao et al., 2025a; Dong et al., 2025c; ?), we propose a Self-evolving Data Flywheel to enrich training dataset. Specifically, the agent is iteratively guided to either refine generated trajectories or explore diverse solution paths, merging valid samples

back into the pool. This mechanism can yield diverse, high-quality trajectories for fine-tuning.

Initialization. For each question q in the original data D_{source} , we use prompt I_1 and guide LLM (defined as M) in generating multiple outputs, establishing an initial action distribution. Based on answer correctness, we divide these outputs (defined as D_{rollout}^q) into *Correct Set* and *Incorrect Set*, serving as the foundation for the flywheel.

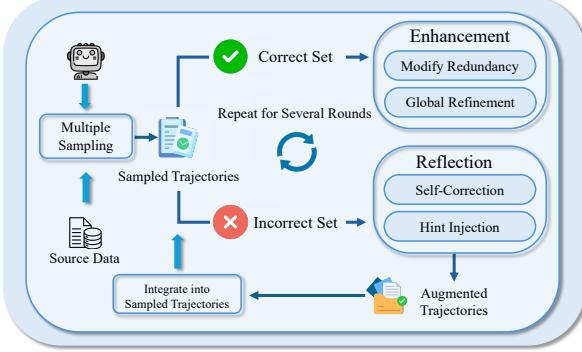
Correct Reasoning Enhancement. For the *Correct Set*, we design two enhancement strategies to remove redundant thinking steps or tool calls.

(1) Redundant Modification: According to recent studies (Liao et al., 2025; Chen et al., 2025c), we specifically refine the first redundant tool call. We define a “step” as the content generated by the LLM between the previous tool execution result (or initial query) and the subsequent tool call (or final answer). We use prompt I_2 to instruct M to identify the first step with redundancy, generate a refined step, and regenerate the subsequent path from it. **(2) Global Refinement:** Complementarily, we design prompt I_3 to instruct M to refine all thinking processes while preserving the logical consistency of the tool call sequence.

Incorrect Reasoning Reflection. For the *Incorrect Set*, we also design two enhancement strategies for modification. **(1) Self-Correction:** We design prompt I_4 and instruct M to find the first incorrect reasoning step, modify it, and continue reasoning. **(2) Hint Injection:** As described in Section 3, agents without behavior calibration may occur insufficient tool call issues. START guides the agent in reasoning with tool calls by inserting hints into the reasoning chain (Li et al., 2025c). Inspired by this, we design different hints for different tasks, and insert them after the identified error step or at the end of the whole trajectory. This operation aims to encourage M to call additional tools to assist in reasoning and derive alternative solutions.

Iterative Data Evolution. We reintegrate enhanced trajectories into D_{rollout}^q and repeat this process for R iterations, yielding the final dataset D_{aug} . It significantly broadens the action space coverage of initial dataset. It can guide M to learn comprehensive exploration strategies during the fine-tuning phase, which establishes a robust foundation for subsequent group-wise behavior calibration. The algorithm is outlined in Algorithm 1, and prompts and hints can be found in Appendix E.

Part I: Self-Evolving Data Flywheel



Part II: Behavior Calibration Training Framework

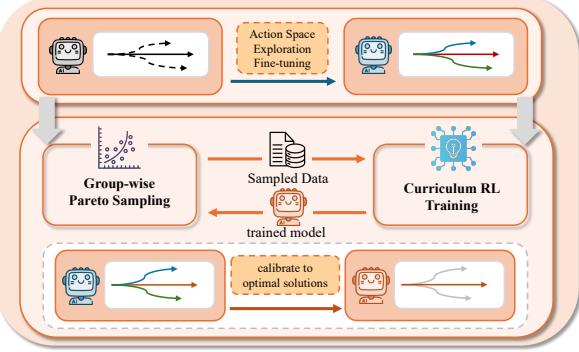


Figure 5: The overall process of ET-Agent. The left part represents Self-Evolving Data Flywheel, while the right part represents Behavior Calibration Training Framework.

Algorithm 1 Self-Evolving Data Flywheel

```

Input initial policy model  $M$ ; original training data  $D_{\text{source}}$ ;
Maximum evolution rounds  $R$ 
1: Initialize  $D_{\text{aug}} = \emptyset$ 
2: for  $q$  in  $D_{\text{source}}$  do
3:    $D_{\text{rollout}}^q = M(I_1, q)$ 
4:   for  $r = 1, \dots, R$  do
5:     Divide  $D_{\text{rollout}}^q$  into Correct Set, Incorrect Set
6:     Trajs1 = Enhancement(Correct Set)
7:     Trajs2 = Reflection(Incorrect Set)
8:      $D_{\text{rollout}}^q.\text{append}(\text{Trajs}^1, \text{Trajs}^2)$ 
9:   end for
10:   $D_{\text{aug}}.\text{append}(D_{\text{rollout}}^q)$ 
11: end for
Output  $D_{\text{aug}}$ 
```

4.2 Behavior Calibration Training Framework

In this section, we introduce a Behavior Calibration Training framework. First, we conduct an **Action Space Exploration Fine-tuning**. Then, we conduct **Iterative Behavior Calibration Reinforcement Learning**. It operates by alternating between Group-wise Pareto sampling and Curriculum RL Training. Through this training framework, the model’s tool call actions in the action space can be calibrated to optimal behavior patterns.

4.2.1 Action Space Exploration Fine-tuning

We first impose quality controls on D_{aug} , discarding trajectories with incorrect answers, formatting errors, or execution failures. Then we perform rejection sampling fine-tuning (RFT) on M (Yuan et al., 2023). The training target can be expressed as: $\max_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\sum_{t=1}^{|y|} \log P_{\theta}(y_t | x, y_{<t}) \right]$. RFT on flywheel-enhanced trajectories expands the agent’s exploration of the action space, while quality controls calibrate the agent’s behavior to eliminate Aborted Tool Execution and format er-

rors. Driven by expanded exploration, trained M exhibits pronounced group-wise trajectory diversity, providing the necessary variance for effective group-wise calibration in the subsequent RL phase.

4.2.2 Group-wise Pareto Sampling

Traditional group-based RL methods are susceptible to trajectory homogenization, leading to vanishing gradients (Xie et al., 2025b; DeepSeek-AI, 2025; Dong et al., 2025a). To overcome this, drawing on existing works (Deb et al., 2002; Rooker, 1991), we introduce a Pareto Sampling strategy. By strictly selecting samples from the Pareto frontier which preserves enough diversity, this approach ensures significant reward differentials within groups, preventing gradient decay and calibrating the agent toward high-quality trajectories.

Group-wise Dispersion Definition To quantify the distributional divergence, we sample K trajectories $\{\tau_i\}_{i=1}^K$ for q in D_{source} (with $K = 16$ to approximate the real distribution). To effectively calibrate the agent’s behavioral patterns while further improving outcome correctness, we construct two complementary sampling indices: **(1) Correctness Dispersion (S_{corr})** is defined as the standard deviation of trajectory correctness scores:

$$S_{\text{corr}}(x) = \sqrt{\frac{1}{K} \sum_{i=1}^K (r(\tau_i) - \bar{r})^2}, \text{ where } \bar{r} \text{ is the mean score. A higher } S_{\text{corr}} \text{ implies greater potential for optimization in problem-solving accuracy.}$$

(2) Behavioral Dispersion (S_{tool}) is defined as the standard deviation of tool invocation counts: $S_{\text{tool}}(x) = \sqrt{\frac{1}{K} \sum_{i=1}^K (l(\tau_i) - \bar{l})^2}$, where \bar{l} is the mean invocation count. High S_{tool} implies rich TIR signals for behavioral exploration.

Sample Selection Strategy We define the dominance relationship between samples. A sample p dominates q ($p \succ q$) if p is non-inferior in all indices and superior in at least one. Based on this, we implement a two-stage selection strategy.

(1) Fast Non-dominated Sorting: We partition the dataset D_{source} into Pareto frontiers $\mathcal{F}_1, \mathcal{F}_2, \dots$, where \mathcal{F}_1 contains non-dominated optimal solutions, and \mathcal{F}_{i+1} are the suboptimal solutions only dominated by \mathcal{F}_i . We prioritize selecting samples from higher frontiers to maximize gradient contribution. **(2) Crowding Distance Truncation:** When adding frontier \mathcal{F}_i exceeds the target size N , we calculate the crowding distance in this frontier. We sort samples along S_{corr} and S_{tool} dimensions, and compute crowding distances based on adjacent intervals. We prioritize retaining samples with larger crowding distances to maximize diversity.¹

4.2.3 Curriculum RL Training

In this section, we present a Curriculum RL framework to calibrate TIR behaviors. We employ group-wise policy optimization with a reward mechanism. This can effectively calibrates behavior patterns while strictly preserving reasoning correctness.

Reward Mechanism Design. To maintain format, outcome, and thinking process accuracy while calibrating behavioral patterns, we employ a multi-objective group-wise reward mechanism. Specifically, we design four key components to target these distinct objectives: format reward, correctness reward, TIR behavior pattern reward, and thinking process reward.

- **Format and Correctness Rewards.** Following prior works (Dong et al., 2025b; Li et al., 2025h; Jin et al., 2025a), we define the correctness reward (R_{corr}), employing the F1 score for knowledge-intensive task, and a binary reward for mathematical reasoning task. We also impose a formatting constraint (R_{format}), with -1 reward for format-incorrect outputs.
- **Efficiency-driven Behavior Rewards.** We introduce rewards for tool call and reasoning length to penalize inefficient behaviors. For trajectory i , the tool call score f_k is as follows:

$$f_{\text{tool}} = \frac{1}{1 + e^{\sigma_{\text{tool}} \cdot (v_{i,\text{tool}} - \bar{v}_{\text{tool}})}}. \quad (1)$$

¹The specific sampling flow can be found in Appendix B.

The reasoning length score is as follows:

$$f_{\text{len}} = \frac{1}{1 + e^{\sigma_{\text{len}} \cdot (v_{i,\text{len}} - \bar{v}_{\text{len}})}}. \quad (2)$$

Among them, $v_{i,\text{tool}}$, \bar{v}_{tool} , σ_{tool} , $v_{i,\text{len}}$, \bar{v}_{len} , and σ_{len} represent the tool call count for trajectory i , the group average tool call count, the tool call hyperparameter, the reasoning length, the group average reasoning length, and the reasoning length hyperparameter, respectively. For group-wise reasoning lengths, they do not include tool execution results, and we perform Z-score standardization among them. This mechanism incentivizes efficiency: by assigning higher rewards to samples with superior indicators, it explicitly steers the agent to more concise reasoning paths.

Overall, the reward for trajectory i aggregates these components:

$$R_i = R_{\text{corr}}^i \cdot f_{\text{tool}}^i \cdot f_{\text{len}}^i + R_{\text{format}}^i. \quad (3)$$

By maximizing rewards for correct, efficient, and well-formatted samples, this mechanism compels calibration towards superior trajectories, promoting overall behavioral effectiveness during training.

ARPO Training. We employ Agentic Reinforced Policy Optimization (ARPO) for RL (Dong et al., 2025d). The objective function is:

$$\begin{aligned} \mathcal{J}_{\text{ARPO}}(\theta) &= \mathbb{E}_{[q \sim D, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | q)]} \\ &\left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min(r_t(\theta) A_i, \text{clip}(r_t(\theta), 1 \pm \epsilon) A_i) \right], \end{aligned} \quad (4)$$

where $r_t(\theta)$ is the probability ratio, and ϵ is the clipping hyperparameter. The advantage A_i for the i -th trajectory is derived from the normalized rewards within the group: $A_i = \frac{R_i - \bar{R}}{\text{std}(\bar{R})}$. Drawing on curriculum learning strategies (Chen et al., 2025c; Li et al., 2025g; Chen et al., 2025d), we alternate between Pareto sampling and ARPO training over multiple rounds. To prevent reward hacking, we progressively decrease σ_{tool} and σ_{len} in each round. This iterative process gradually guides the model’s exploration toward optimal trajectories, stabilizing both behavioral patterns and answer accuracy.

5 Experiments

5.1 Experimental Setup

Tool Design. According to existing works (Li et al., 2025e,h; Dong et al., 2025b; Chen et al.,

Table 1: Results on six tasks, with top two results highlighted in **bold** and underlined. and represent method with code compiler and search tool, respectively. For fairness, we retrain a WebSailor-7B model using Wikipedia-based data. Abbreviation: it (Instruct), 2Wiki (2WikiMultiHopQA), Bamb (Bamboogle), MSQ (MuSiQue).

Method	Mathematical-Reasoning Tasks						Knowledge-Intensive Tasks						Avg.	
	AIME24		AMC23		MATH500		2Wiki		Bamb		MSQ			
	LJ	Effi	LJ	Effi	LJ	Effi	F1	Effi	F1	Effi	F1	Effi	P	Effi
Direct Inference														
Qwen2.5-7B-it	10.0	-	30.0	-	57.2	-	18.9	-	24.5	-	8.5	-	24.9	-
Single-TIR Methods														
Search-o1	10.0	10.0	37.5	37.5	65.4	65.2	42.9	26.2	48.5	33.9	13.3	10.9	36.3	30.6
Search-R1	23.3	18.3	35.0	25.6	57.8	39.3	49.2	15.5	40.5	13.7	18.6	4.8	37.4	19.6
Research	10.0	5.8	52.5	21.7	64.0	31.2	<u>67.1</u>	25.5	58.4	24.8	27.8	9.3	46.6	19.7
WebThinker	16.7	16.7	45.0	45.0	66.0	65.7	60.8	24.8	46.3	23.4	18.6	9.1	42.2	30.8
WebSailor*	26.7	2.9	65.0	23.5	72.2	13.9	61.5	5.3	58.4	6.1	17.0	1.6	50.1	8.9
ToRL	<u>40.0</u>	<u>40.0</u>	<u>75.0</u>	<u>71.1</u>	84.6	83.8	35.1	<u>35.1</u>	14.2	14.2	5.8	5.8	42.5	41.6
DotaMath	16.7	16.7	50.0	50.0	74.6	72.5	16.3	16.3	18.9	18.9	7.6	7.6	30.7	30.3
START	23.3	23.3	67.5	62.1	78.4	74.8	15.6	15.6	19.8	19.8	5.7	5.7	35.1	33.6
Multi-TIR Methods														
Prompting	13.3	6.1	47.5	33.54	67.0	51.7	18.0	10.5	17.9	12.4	8.5	5.6	28.7	20.0
IKEA	10.0	6.1	42.5	32.0	63.6	44.2	34.7	9.7	35.8	11.2	14.8	3.8	33.6	17.8
SMART	3.3	3.3	17.5	15.8	39.4	33.0	21.7	21.1	40.7	40.6	12.3	11.8	22.5	20.9
AutoTIR	33.3	21.3	62.5	46.4	72.6	61.2	66.2	27.8	56.9	29.5	30.9	12.0	53.7	33.1
Tool-Star	33.3	25.6	60.0	53.8	80.4	70.0	65.8	29.0	55.4	34.2	24.5	<u>12.6</u>	52.3	37.5
Tool-Light	33.3	33.3	72.5	67.5	79.0	72.8	60.4	24.6	<u>58.7</u>	23.0	26.0	9.8	<u>55.0</u>	38.5
ET-Agent (Ours)	46.7	43.3	77.5	71.3	<u>81.6</u>	<u>76.1</u>	67.2	35.6	59.4	<u>34.4</u>	<u>28.0</u>	15.3	60.1	46.0

2025c), we design two tools to assist in reasoning tasks: **Search Engines** (including local search and web search) retrieve and summarize external knowledge, and **Code Compiler** executes LLM-generated code. These tools are sufficient for supporting tasks evaluated in this study.

Datasets We conduct evaluations on two types of reasoning tasks: (1) Mathematical-Reasoning tasks (including AIME24,² AMC23,³ MATH500 (Lightman et al., 2024)) and (2) Knowledge-Intensive tasks (including 2WikiMultiHopQA (Ho et al., 2020), Bamboogle (Press et al., 2023), and MuSiQue (Trivedi et al., 2022)).

Evaluation Metrics We measure ET-Agent’s performance on correctness and efficiency aspects. For correctness, we utilize LLM-as-Judge (Li et al., 2024d) (Qwen2.5-72B-Instruct (Yang et al., 2024) for Mathematical-Reasoning tasks) and F1 score (for Knowledge-Intensive tasks). Efficiency metric is calculated as the average correctness per tool call: $Effi = \frac{1}{N} \sum_{i=1}^N \frac{P_i}{T_i}$, where P_i is the correctness score and T_i is the tool call count for sample i .

²<https://huggingface.co/datasets/AI-MO/aimo-validation-aime>

³<https://huggingface.co/datasets/zwhe99/amc23>

Baselines We categorize baselines into three groups: (1) **Direct Reasoning**, utilizing Qwen2.5-7B-Instruct (Yang et al., 2024) without tools; (2) **Single-TIR methods**, including Search-o1, Search-R1, Research, WebThinker, WebSailor, ToRL, DotaMath, and START (Li et al., 2025e; Jin et al., 2025a; Li et al., 2025g,h,c; Chen et al., 2025b; Li et al., 2025d, 2024a); and (3) **Multi-TIR methods**, including prompt-based approaches, IKEA, SMART, AutoTIR, Tool-Star, and Tool-Light (Huang et al., 2025; Qian et al., 2025; Wei et al., 2025; Dong et al., 2025b; Chen et al., 2025c).

Implementation Details Our training is all based on the Wikipedia dataset. During testing, for Knowledge-Intensive tasks, we use local retrieval based on Wikipedia documents. For Mathematical-Reasoning Tasks, we use Google Search. More details can be found in Appendix A and C.

5.2 Main Results

Main results are shown in Table 1. In correctness and efficiency aspects, ET-Agent achieves the best performance in most tasks. Through further analysis, we can draw the following conclusions:

The Necessity of Training. Models without training struggle to arrive at correct answers. For instance, the average correctness of Search-o1 is

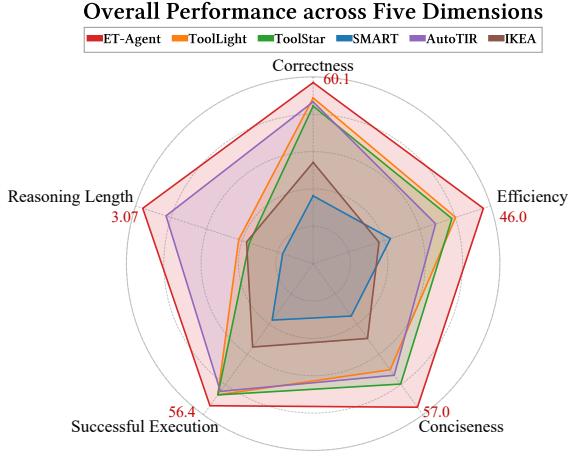


Figure 6: Performances across five metrics.

lower than that of Research by 10.3 points. These findings underscore the importance of training.

Efficiency Neglection in Current Baselines.

Existing methods often struggle to balance accuracy with efficiency. For example, although ToolStar achieves high correctness (52.3), its efficiency is surpassed by ToRL and Tool-Light. Similarly, while AutoTIR attains the highest correctness on MSQ, its efficiency (12.0) is significantly inferior to that of ET-Agent (15.3). This discrepancy highlights that many current TIR approaches neglect the optimization of behavioral patterns.

Superior Performance of ET-Agent. On average, ET-Agent achieves the best results in both correctness and efficiency indices (60.1, 46.0). It significantly outperforms baseline methods on datasets such as AIME24, AMC23, and 2Wiki, while maintaining a leading position across others. These results validate that our ET-Agent pipeline effectively enhances reasoning accuracy while standardizing the model’s behavioral patterns.

5.3 Tool-use Behavioral Patterns Analysis

To comprehensively analyze the effectiveness of ET-Agent, we additionally define three metrics: (1) **Conciseness** measures the brevity of tool calls. It is defined as N_{nr}/N , where N_{nr} denotes the number of samples without redundant tool calls. (2) **Successful Execution** evaluates the reliability of tool calls. It is defined as $\frac{1}{N} \sum_{i=1}^N \frac{P_i}{\text{Wrong}_i}$, where Wrong_i represents the number of incorrect tool executions in sample i . (3) **Reasoning Length** quantifies the compactness of the thought process. It is calculated as $\frac{1}{N} \sum_{i=1}^N \frac{L_i}{P_i}$, where L_i is the token length of the reasoning chain (excluding tool execution results).

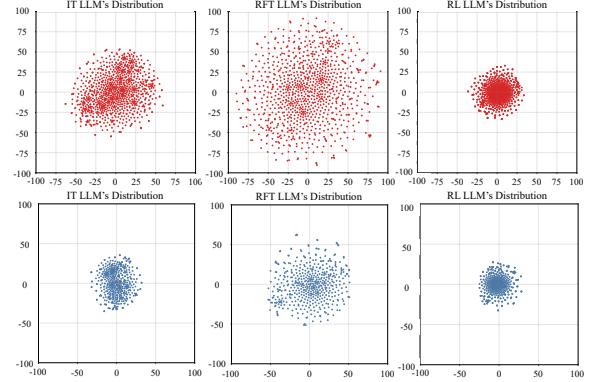


Figure 7: Distribution of action space exploration in different training stages. The more dispersed the distribution, the higher the diversity. **Top:** Qwen2.5-7B-it. **Bottom:** Llama3.1-8B-it.

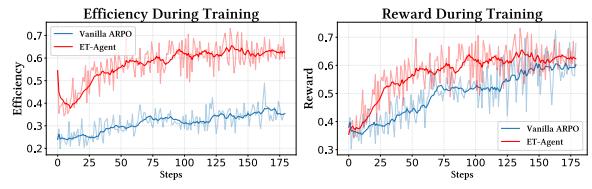


Figure 8: Comparison of efficiency and reward between vanilla ARPO and ET-Agent during the RL phase (only the final RL phase of ET-Agent is chosen).

Figure 6 illustrates the performance across these dimensions. ET-Agent achieves state-of-the-art results on all five metrics, significantly outperforming mainstream Multi-TIR baselines. This comprehensive superiority validates that our pipeline effectively standardizes the agent’s behavioral patterns while ensuring robust reasoning.⁴

5.4 Quantitative Analysis

Action Space Distribution Analysis. We analyze the agent’s output distributions across different training stages. We sample ten trajectories per query across three stages: the base Instruct LLM, the RFT LLM, and the RL LLM. All outputs are encoded using the e5 model (Wang et al., 2022) and projected for visualization via t-SNE (van der Maaten, 2013), as shown in Figure 7. The transition from the IT to RFT exhibits a more dispersed distribution, signifying expanded exploration of the action space. Subsequently, the RL phase yields a more compact distribution, indicating that the LLM converges this broad exploration into precise, optimal trajectories.

⁴Full evaluation results can be found in Appendix D.

Table 2: Results of ablation experiments on different components, with top two results highlighted in **bold** and underlined. Abbreviation: MR (Mathematical-Reasoning Tasks), KI (Knowledge-Intensive Tasks).

Method	MR		KI		Avg.	
	LJ	Effi	LJ	Effi	LJ	Effi
Qwen2.5-7B-it	42.6	30.4	14.8	9.5	28.7	20.0
SFT	58.0	-	40.6	-	49.3	-
w/o Flywheel	<u>54.8</u>	-	<u>39.7</u>	-	<u>47.3</u>	-
ET-Agent	68.6	63.6	51.5	<u>28.4</u>	60.1	46.0
w/o Pareto	59.3	54.4	48.8	25.8	54.1	40.1
w/o Reward	<u>66.6</u>	59.4	<u>50.5</u>	27.3	<u>58.5</u>	43.4
w/o σ Decrease	62.2	<u>59.5</u>	33.9	32.4	48.1	46.0

Training Dynamic Analysis. We analyze the training dynamics by tracking efficiency and reward changes. We compare ET-Agent with a ‘Vanilla ARPO’ baseline, which runs ARPO solely for result correctness. As shown in Figure 8, ET-Agent outperforms that baseline in both efficiency and reward metrics, thereby validating the effectiveness of our behavior pattern reward design.

5.5 Ablation Study

To fully explore the role of each component within the ET-Agent framework, we conduct ablation experiments, specifically focusing on the Self-evolving Data Flywheel, Group-wise Pareto sampling, and the reward mechanism design. As shown in Table 2: (1) Eliminating any component leads to performance degradation across both correctness and efficiency, validating the rationality of ET-Agent design. (2) Data quality is critical, as both the Flywheel and Pareto Sampling significantly outperform their respective baselines. (3) The reward design is essential; removing the group-wise reward mechanism lowers efficiency, while fixing σ causes severe correctness degradation (indicative of reward hacking). This confirms that all components in our reward mechanism are indispensable for robust optimization.

6 Conclusion

In this paper, we introduce ET-Agent, a framework designed to comprehensively calibrate TIR behaviors. We begin by analyzing common erroneous behavioral patterns in TIR models. Building on this analysis, we propose a pipeline that first enhances action space exploration and subsequently calibrates this exploration toward optimal trajectories. Empirical results demonstrate that ET-Agent

exhibits effective behavioral patterns while ensuring inference correctness. We believe this work establishes a robust foundation for future research into TIR behaviors calibration.

Limitations

Due to resource constraints, our experiments are restricted to local Wikipedia retrieval, and it is relatively difficult to apply the ET-Agent framework to larger models. In the future, we intend to scale the ET-Agent framework to larger architectures and integrate live web search capabilities, thereby broadening the scope of TIR behavioral exploration.

References

- Guoxin Chen, Zile Qiao, Xuanzhong Chen, Donglei Yu, Haotian Xu, Wayne Xin Zhao, Ruihua Song, Wenbiao Yin, Huifeng Yin, Liwen Zhang, Kuan Li, Minpeng Liao, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025a. Iterresearch: Rethinking long-horizon agents via markovian state reconstruction. *Preprint*, arXiv:2511.07327.
- Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025b. Research: Learning to reason with search for llms via reinforcement learning. *CoRR*, abs/2503.19470.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Trans. Mach. Learn. Res.*, 2023.
- Yifei Chen, Guanting Dong, and Zhicheng Dou. 2025c. Toward effective tool-integrated reasoning via self-evolved preference learning. *CoRR*, abs/2509.23285.
- Yongchao Chen, Yueying Liu, Junwei Zhou, Yilun Hao, Jingquan Wang, Yang Zhang, and Chuchu Fan. 2025d. R1-code-interpreter: Training llms to reason with code via supervised and reinforcement learning. *CoRR*, abs/2505.21668.
- Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
- Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 6(2):182–197.
- DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

- Guanting Dong, Licheng Bao, Zhongyuan Wang, Kangzhi Zhao, Xiaoxi Li, Jiajie Jin, Jinghan Yang, Hangyu Mao, Fuzheng Zhang, Kun Gai, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. 2025a. [Agentic entropy-balanced policy optimization](#). *Preprint*, arXiv:2510.14545.
- Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. 2025b. [Toolstar: Empowering llm-brained multi-tool reasoner via reinforcement learning](#). *Preprint*, arXiv:2505.16410.
- Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. 2025c. [Self-play with execution feedback: Improving instruction-following capabilities of large language models](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazen Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. 2025d. [Agentic reinforced policy optimization](#). *Preprint*, arXiv:2507.19849.
- Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Ji-Rong Wen, and Zhicheng Dou. 2025e. [Understand what LLM needs: Dual preference alignment for retrieval-augmented generation](#). In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pages 4206–4225. ACM.
- Runnan Fang, Shihao Cai, Baixuan Li, Jialong Wu, Guangyu Li, Wenbiao Yin, Xinyu Wang, Xiaobin Wang, Liangcai Su, Zhen Zhang, Shibin Wu, Zhengwei Tao, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. [Towards general agentic intelligence via environment scaling](#). *CoRR*, abs/2509.13311.
- Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025. [Group-in-group policy optimization for LLM agent training](#). *CoRR*, abs/2505.10978.
- Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahuo Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, and 8 others. 2025a. [A survey of self-evolving agents: On path to artificial super intelligence](#). *CoRR*, abs/2507.21046.
- Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. 2025b. [Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous RL](#). *CoRR*, abs/2508.07976.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2024. [Tora: A tool-integrated reasoning agent for mathematical problem solving](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.
- Pengfei He, Zhenwei Dai, Bing He, Hui Liu, Xianfeng Tang, Hanqing Lu, Juanhui Li, Jiayuan Ding, Subhabrata Mukherjee, Suhang Wang, Yue Xing, Jiliang Tang, and Benoît Dumoulin. 2025. [Traject-bench:a trajectory-aware benchmark for evaluating agentic tool use](#). *CoRR*, abs/2510.04550.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps](#). In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625.
- Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. 2025. [Treerl: LLM reinforcement learning with on-policy tree search](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 12355–12369.
- Jie Huang and Kevin Chen-Chuan Chang. 2023. [Towards reasoning in large language models: A survey](#). In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1049–1065.
- Ziyang Huang, Xiaowei Yuan, Yiming Ju, Jun Zhao, and Kang Liu. 2025. [Reinforced internal-external knowledge synergistic reasoning for efficient adaptive search agent](#). *CoRR*, abs/2505.07596.
- Hugging Face. 2025. [Open r1: A fully open reproduction of deepseek-r1](#).
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. [Gpt-4o system card](#). *CoRR*, abs/2410.21276.
- Dongfu Jiang, Yi Lu, Zhuofeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou, Chao Du, Tianyu Pang, and Wenhui Chen. 2025. [Verltool: Towards holistic agentic reinforcement learning with tool use](#). *CoRR*, abs/2509.01055.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025a. [Search-r1: Training llms to reason and leverage search engines with reinforcement learning](#). *CoRR*, abs/2503.09516.
- Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. 2025b. [Flashrag: A modular toolkit for efficient retrieval-augmented generation research](#). In *Companion Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, October 1-5, 2025*.

Australia, 28 April 2025 - 2 May 2025, pages 737–740.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. **Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611.

Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Hufeng Yin, Jialong Wu, Jingren Zhou, Kuan Li, Liangcai Su, Litu Ou, Liwen Zhang, Pengjun Xie, Rui Ye, Wenbiao Yin, Xinmiao Yu, Xinyu Wang, Xixi Wu, and 36 others. 2025a. **Tongyi deepresearch technical report**. *CoRR*, abs/2510.24701.

Chengpeng Li, Guanting Dong, Mingfeng Xue, Ru Peng, Xiang Wang, and Dayiheng Liu. 2024a. **Dotamath: Decomposition of thought with code assistance and self-correction for mathematical reasoning**. *CoRR*, abs/2407.04078.

Chengpeng Li, Zhengyang Tang, Ziniu Li, Mingfeng Xue, Keqin Bao, Tian Ding, Ruoyu Sun, Benyou Wang, Xiang Wang, Junyang Lin, and Dayiheng Liu. 2025b. **Cort: Code-integrated reasoning within thinking**. *CoRR*, abs/2506.09820.

Chengpeng Li, Mingfeng Xue, Zhenru Zhang, Jiaxi Yang, Beichen Zhang, Xiang Wang, Bowen Yu, Binyuan Hui, Junyang Lin, and Dayiheng Liu. 2025c. **START: self-taught reasoner with tools**. *CoRR*, abs/2503.04625.

Chengpeng Li, Zheng Yuan, Hongyi Yuan, Guanting Dong, Keming Lu, Jiancan Wu, Chuanqi Tan, Xiang Wang, and Chang Zhou. 2024b. **Mugglemath: Assessing the impact of query and response augmentation on math reasoning**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10230–10258.

Chengshu Li, Jacky Liang, Andy Zeng, Xinyun Chen, Karol Hausman, Dorsa Sadigh, Sergey Levine, Li Fei-Fei, Fei Xia, and Brian Ichter. 2024c. **Chain of code: Reasoning with a language model-augmented code emulator**. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024d. **From generation to judgment: Opportunities and challenges of llm-as-a-judge**. *CoRR*, abs/2411.16594.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume

Lample, and Stanislas Polu. 2024. **Numina-math**. [<https://github.com/project-numina/aimo-progress-prize>] (https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf).

Kuan Li, Zhongwang Zhang, Hufeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025d. **Websailor: Navigating super-human reasoning for web agent**. *CoRR*, abs/2507.02592.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025e. **Search-o1: Agentic search-enhanced large reasoning models**. *CoRR*, abs/2501.05366.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Guanting Dong, Jiajie Jin, Yinuo Wang, Hao Wang, Yutao Zhu, Ji-Rong Wen, Yuan Lu, and Zhicheng Dou. 2025f. **Deepagent: A general reasoning agent with scalable toolsets**. *CoRR*, abs/2510.21618.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025g. **Webthinker: Empowering large reasoning models with deep research capability**. *CoRR*, abs/2504.21776.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025h. **Torl: Scaling tool-integrated RL**. *CoRR*, abs/2503.23383.

Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tian-shun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, Zheng Zhang, Wei Shen, Qian Liu, Chenghua Lin, Jian Yang, Ge Zhang, and Wenhao Huang. 2025i. **Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling**. *CoRR*, abs/2508.17445.

Zongjie Li and Shuai Wang. 2025. **Reasoning as a resource: Optimizing fast and slow thinking in code generation models**. *CoRR*, abs/2506.09396.

Baohao Liao, Xinyi Chen, Sara Rajaei, Yuhui Xu, Christian Herold, Anders Søgaard, Maarten de Rijke, and Christof Monz. 2025. **Lost at the beginning of reasoning**. *CoRR*, abs/2506.22058.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Jiaye Lin, Yifu Guo, Yuzhen Han, Sen Hu, Ziyi Ni, Licheng Wang, Mingguang Chen, Hongzhang Liu, Ronghao Chen, Yangfan He, Dixin Jiang, Binxing Jiao, Chen Hu, and Huacan Wang. 2025. **Se-agent: Self-evolution trajectory optimization in multi-step reasoning with llm-based agents**. *CoRR*, abs/2508.02085.

- Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, and James Zou. 2025. Octotools: An agentic framework with extensible tools for complex reasoning. *CoRR*, abs/2502.11271.
- Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2024. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *CoRR*, abs/2412.09413.
- Ranjita Naik, Varun Chandrasekaran, Mert Yuksekgonul, Hamid Palangi, and Besmira Nushi. 2024. Diversity of thought improves reasoning abilities of llms. *Preprint*, arXiv:2310.07088.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5687–5711.
- Cheng Qian, Emre Can Acikgoz, Hongru Wang, Xiusi Chen, Avirup Sil, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2025. SMART: self-aware agent for tool overuse mitigation. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 4604–4621.
- Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, MiaoXuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.
- Runqi Qiao, Qiuna Tan, Minghan Yang, Guanting Dong, Peiqing Yang, Shiqiang Lang, Enhui Wan, Xiaowan Wang, Yida Xu, Lan Yang, Chong Sun, Chen Li, and Honggang Zhang. 2025. V-thinker: Interactive thinking with images. *CoRR*, abs/2511.04460.
- Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. 2024. O1 replication journey: A strategic progress report - part 1. *CoRR*, abs/2410.18982.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506.
- Terry Rooker. 1991. Review of genetic algorithms in search, optimization, and machine learning. *AI Mag.*, 12(1):102–103.
- Bytedance Seed. Seed1.8 model card: Towards generalized real-world agency.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/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, EuroSys 2025, Rotterdam, The Netherlands, 30 March 2025 - 3 April 2025*, pages 1279–1297.
- Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A thorough examination of decoding methods in the era of llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8601–8629.
- Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *CoRR*, abs/2503.05592.
- Liangcai Su, Zhen Zhang, Guangyu Li, Zhuo Chen, Chenxi Wang, Maojia Song, Xinyu Wang, Kuan Li, Jialong Wu, Xuanzhong Chen, Zile Qiao, Zhongwang Zhang, Hufeng Yin, Shihao Cai, Runnan Fang, Zhengwei Tao, Wenbiao Yin, Chenxiong Qian, Yong Jiang, and 3 others. 2025. Scaling agents via continual pre-training. *CoRR*, abs/2509.13310.
- Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *Trans. Mach. Learn. Res.*, 2025.
- Jimin Sun, So Yeon Min, Yingshan Chang, and Yonatan Bisk. 2024. Tools fail: Detecting silent errors in faulty tools. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 14272–14289.
- Zhengwei Tao, Haiyang Shen, Baixuan Li, Wenbiao Yin, Jialong Wu, Kuan Li, Zhongwang Zhang, Hufeng Yin, Rui Ye, Liwen Zhang, Xinyu Wang, Pengjun Xie, Jingren Zhou, and Yong Jiang. 2025a. Webbleaper: Empowering efficiency and efficacy in webagent via enabling info-rich seeking. *CoRR*, abs/2510.24697.
- Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025b. Webshaper: Agentically data synthesizing via information-seeking formalization. *CoRR*, abs/2507.15061.

- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. *Musique: Multi-hop questions via single-hop question composition*. *Trans. Assoc. Comput. Linguistics*, 10:539–554.
- Laurens van der Maaten. 2013. *Barnes-hut-sne*. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Conference Track Proceedings*.
- Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. 2025a. *Acting less is reasoning more! teaching model to act efficiently*. *Preprint*, arXiv:2504.14870.
- Hongru Wang, Boyang Xue, Baohang Zhou, Tianhua Zhang, Cunxiang Wang, Huimin Wang, Guanhua Chen, and Kam-Fai Wong. 2025b. *Self-dc: When to reason and when to act? self divide-and-conquer for compositional unknown questions*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 6510–6525.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. *Text embeddings by weakly-supervised contrastive pre-training*. *CoRR*, abs/2212.03533.
- Yifan Wei, Xiaoyan Yu, Yixuan Weng, Tengfei Pan, Angsheng Li, and Li Du. 2025. *Autotir: Autonomous tools integrated reasoning via reinforcement learning*. *CoRR*, abs/2507.21836.
- Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025a. *Webdancer: Towards autonomous information seeking agency*. *CoRR*, abs/2505.22648.
- Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. 2025b. *Agentic reasoning: A streamlined framework for enhancing LLM reasoning with agentic tools*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 28489–28503.
- Peilin Wu, Mian Zhang, Xinlu Zhang, Xinya Du, and Zhiyu Zoey Chen. 2025c. *Search wisely: Mitigating sub-optimal agentic searches by reducing uncertainty*. *CoRR*, abs/2505.17281.
- Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025a. *Logic-rl: Unleashing LLM reasoning with rule-based reinforcement learning*. *CoRR*, abs/2502.14768.
- Xuan Xie, Xuan Wang, and Wenjie Wang. 2025b. *Dagro: Rectifying gradient conflict in reasoning via distinctiveness-aware group relative policy optimization*. *Preprint*, arXiv:2512.06337.
- Qian Xiong, Yuekai Huang, Ziyou Jiang, Zhiyuan Chang, Yu Zheng, Tianhao Li, and Mingyang Li. 2025. *Butterfly effects in toolchains: A comprehensive analysis of failed parameter filling in LLM tool-agent systems*. *CoRR*, abs/2507.15296.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. *Qwen2.5 technical report*. *CoRR*, abs/2412.15115.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. *Hotpotqa: A dataset for diverse, explainable multi-hop question answering*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. *React: Synergizing reasoning and acting in language models*. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.
- Chenlong Yin, Zeyang Sha, Shiwen Cui, and Changhua Meng. 2025. *The reasoning trap: How enhancing LLM reasoning amplifies tool hallucination*. *CoRR*, abs/2510.22977.
- Fei Xu Yu, Gina C. Adam, Nathaniel D. Bastian, and Tian Lan. 2025a. *Optimizing prompt sequences using monte carlo tree search for llm-based optimization*. *CoRR*, abs/2508.05995.
- Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025b. *DAPO: an open-source LLM reinforcement learning system at scale*. *CoRR*, abs/2503.14476.
- Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. 2023. *Scaling relationship on learning mathematical reasoning with large language models*. *CoRR*, abs/2308.01825.
- Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, Yifan Zhou, Yang Chen, Chen Zhang, Yutao Fan, Zihu Wang, Songtao Huang, Yue Liao, Hongru Wang, Mengyue Yang, and 6 others. 2025. *The landscape of agentic reinforcement learning for llms: A survey*. *CoRR*, abs/2509.02547.

Muyang Zhao, Qi Qi, and Hao Sun. 2026. [Roi-reasoning: Rational optimization for inference via pre-computation meta-cognition](#). *Preprint*, arXiv:2601.03822.

A Details about Datasets and Baselines

In this section, we introduce all the datasets and baselines used in the experiment.

A.1 Datasets

A.1.1 Training Datasets

Our training data is composed as shown in Table 3.

AIME is derived from the American Invitational Mathematics Examination. The questions usually consist of 30 problems each year, covering Algebra, Geometry, Number Theory, Combinatorics, and Probability. It has become an important standard for measuring the reasoning level of models.

NuminaMATH ([LI et al., 2024](#)) is developed by the Numina team and Hugging Face, aiming to significantly enhance LLMs’ reasoning ability in solving mathematical problems. It contains approximately 860,000 pairs of mathematical problems and their detailed solutions, covering various difficulty levels from high school exercises to top international Olympic competitions.

OpenR1-Math-220k ([Hugging Face, 2025](#)) is a large-scale mathematical reasoning dataset released by the Open R1 team of Hugging Face in early 2025. It contains approximately 220k pieces of high-quality mathematical reasoning data. These data are directly generated locally using DeepSeek-R1 on 512 H100 graphics cards, covering various difficulty levels from basic education to competition level.

DAPO ([Yu et al., 2025b](#)) is jointly launched by ByteDance and Tsinghua University’s AIR Institute, aiming to solve the training challenges of LLMs in Long-CoT. DAPO-Math-17k is the core open-source dataset of this project, specifically designed to enhance the mathematical reasoning ability of models.

2WikiMultiHopQA ([Ho et al., 2020](#)) is a benchmark dataset used to evaluate the multi-hop reasoning capabilities of LLMs. It is proposed in 2020 at the COLING conference, aiming to address issues present in mainstream datasets at the time (such as HotpotQA), including “incomplete logical chains” and “solving problems through shortcuts without requiring multi-hop reasoning”.

Table 3: The composition of training data after sampling. For AIME, we use questions from 2023 and earlier. Abbreviation: MR (Mathematical-Reasoning Tasks), KI (Knowledge-Intensive Tasks), DS (Deepsearch Tasks).

Dataset	Task	# Counts
AIME	MR	0.9k
Numina-CoT (LI et al., 2024)	MR	5k
Numina-TIR (LI et al., 2024)	MR	5k
OpenR1-Math-220k (Hugging Face, 2025)	MR	10k
DAPO-Math-17k (Yu et al., 2025b)	MR	5k
2WikiMultiHopQA (Ho et al., 2020)	KI	5k
HotpotQA (Yang et al., 2018)	KI	5k
MuSiQue (Trivedi et al., 2022)	KI	5k
TriviaQA (Joshi et al., 2017)	KI	5k
ASearcherLRM35k (Gao et al., 2025b)	DS	10k
WebShaper (Tao et al., 2025b)	DS	0.5k

HotpotQA ([Yang et al., 2018](#)) is a highly renowned large-scale multi-hop question-answering dataset in the field of Natural Language Processing (NLP). The original design intent is to advance machines’ capabilities in human-like reasoning chains. Unlike traditional question-answering datasets, HotpotQA requires models to integrate information from multiple different documents and perform reasoning to arrive at an answer.

MuSiQue ([Trivedi et al., 2022](#)) is currently a highly challenging multi-hop question-answering benchmark in the field of Natural Language Processing. Instead of directly asking crowdsourcing personnel to write complex questions, it starts from simple “single-hop questions” and constructs complex 2-4 hop questions through rigorous logical combinations. It contains approximately 25,000 questions, covering logical depths of 2 to 4 hops.

TriviaQA ([Joshi et al., 2017](#)) is a highly classic and challenging large-scale open-domain question answering and reading comprehension dataset in the field of Natural Language Processing, primarily consisting of “question-answer-evidence document” triples. The original design intent of TriviaQA was to provide an evaluation benchmark that is closer to the real world and more difficult to solve through simple keyword matching.

ASearcherLRM35k ([Gao et al., 2025b](#)) is a dataset autonomously generated by a Data Synthesis Agent. The core value of this dataset lies in addressing the pain points of data scarcity and overly short interaction turns for search agents during training. Through a fully automated synthesis

workflow, it constructs a high-quality task library that supports models in performing “long-range search reasoning”.

WebShaper (Tao et al., 2025b) is a formal-driven information-seeking data synthesis framework and dataset released by Alibaba Tongyi Lab in 2025. WebShaper shifts toward a “formal-driven” approach, utilizing mathematical set theory to systematically construct complex search and reasoning tasks.

A.1.2 Evaluation Datasets

AMC23 is an important benchmark in the field of artificial intelligence used to evaluate the competition-level mathematical reasoning capabilities of LLMs. It primarily consists of real exam questions from the 2023 American Mathematics Competitions (AMC 10 and AMC 12). It can effectively test whether a model truly possesses reasoning abilities or is merely “memorizing” old problems from its training set.

MATH500 (Lightman et al., 2024) is currently one of the core benchmarks for measuring the competition-level mathematical reasoning capabilities of LLMs. Its emergence was primarily to address the issue of excessive evaluation costs associated with full-scale datasets, while ensuring that the evaluation problems possess sufficient discriminative power. It covers seven mathematical subfields, including Algebra, Geometry, and Number Theory.

Bamboogle (Press et al., 2023) is a challenging dataset specifically designed to evaluate the multi-hop reasoning and Retrieval-Augmented Generation (RAG) capabilities of LLMs. It is used to test whether search agents can piece together the final answer through multiple, multi-dimensional searches.

A.2 Baselines

Direct Inference We use the Qwen2.5-7B-Instruct (Yang et al., 2024) model to directly perform reasoning on the task set, and the prompt is shown in Figure E.2.

Search-o1 (Li et al., 2025e) is a framework designed to enhance the search reasoning capabilities of LLMs. Its core concept is to combine agentic search with slow thinking reasoning processes to solve the knowledge deficit or hallucination problems. The system triggers a search only when the

model generates specific search instructions (such as query terms with special identifiers) within its thinking process.

Search-R1 (Jin et al., 2025a) is a novel model training framework that deeply integrates Reinforcement Learning with search engines. Search-R1 enables the model to learn like a human researcher: thinking while deciding whether a search is necessary, and refining its thought process based on search results through multiple iterations. This RL-based training approach allows the model to autonomously evolve highly efficient search strategies.

Research (Chen et al., 2025b) The core idea of research is to enable LLMs to autonomously learn, through reinforcement learning, when to search the web and how to utilize the search results, thereby solving complex multi-hop reasoning problems. By providing the model with search tools, it spontaneously learns search strategies through trial and error.

WebThinker (Li et al., 2025g) empowers large reasoning models with the ability to think, search, and write simultaneously like a human researcher. It breaks static limitations, granting the model the power to autonomously explore the internet. WebThinker performs excellently across multiple highly challenging benchmarks.

WebSailor (Li et al., 2025d) is an advanced AI Agent training method proposed by Alibaba Tongyi Lab in 2025. Its core goal is to enable open-source large models to possess reasoning capabilities comparable to closed-source systems when handling complex web information retrieval tasks. Experimental results indicate that WebSailor achieves a qualitative leap compared to previous open-source agents when faced with extremely difficult tasks.

ToRL (Li et al., 2025h) is a framework allowing large models to autonomously learn how to use tools through trial and error, rather than imitating templates provided by humans. During the reasoning process, the model generates code, the system immediately invokes an interpreter to run it, and feeds the results back to the model. In high-difficulty mathematics competitions such as AIME24, the performance of the ToRL-7B model can surpass that of large-scale models without tools.

DotaMath (Li et al., 2024a) is a novel training and inference framework specifically designed to enhance the mathematical reasoning capabilities of LLMs. By introducing methods such as decomposition of thought, code assistance, and self-correction, DotaMath aims to address the issue of traditional LLMs being prone to errors when processing complex mathematics (such as the MATH dataset).

START (Li et al., 2025c) is a brand-new training framework for LLMs proposed by the Alibaba research team in March 2025. This method constructs training data to fine-tune models through prompt-guided reasoning and Hint Rejection Sampling Fine-Tuning. The START model, trained based on the QwQ-32B base model, has delivered astonishing performance on multiple hardcore leaderboards.

IKEA (Huang et al., 2025) is a brand-new method aimed at improving the efficiency of AI Search Agents. This method is designed to address the current issues where AI agents either over-rely on external searches or ignore their own internal knowledge when processing tasks. The IKEA method exhibits advantages such as high efficiency, greater accuracy, and strong generalization when handling knowledge-intensive tasks.

SMART (Qian et al., 2025) is a research framework designed to address the tool-overuse problem that occurs when LLM agents invoke external tools. The SMART method is inspired by human Metacognition, or “thinking about thinking”. It aims to calibrate the agent’s self-awareness, enabling it to determine when to use its own knowledge to think and when to use tools to search.

AutoTIR (Wei et al., 2025) is a novel post-training framework for LLMs. The core purpose of this method is to allow the model to autonomously decide when to use a tool and which tool to use during the reasoning process, rather than following predefined, rigid templates. AutoTIR achieves an excellent balance between core language capabilities and tool integration capabilities.

Tool-Star (Dong et al., 2025b) is a Reinforcement Learning based framework designed to enhance the ability of LLMs to use multiple external tools autonomously, efficiently, and collaboratively during complex reasoning processes. Through two-stage training and systematic data synthesis, Tool-

Star aims to solve problems such as the difficulty of multi-tool collaboration and low tool-calling efficiency.

B Group-wise Pareto Sampling Flow

In this section, we introduce the algorithmic process of Group-wise Pareto Sampling, as shown in Algorithm 2 and 3.

Algorithm 2 Group-wise Pareto Sampling

Input Dataset D_{source} , target selection size N , objectives S_{corr} and S_{tool}

```

1: for  $x$  in  $D_{\text{source}}$  do
2:   Calculate dispersion metrics  $S_{\text{corr}}(x)$  and  $S_{\text{tool}}(x)$ 
   based on  $K$  sampled trajectories
3: end for
4: Partition  $D_{\text{source}}$  into hierarchical Pareto frontiers
    $\{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_k\}$  according to relation  $\succ$ 
5: Initialize  $\mathcal{S} \leftarrow \emptyset$ 
6: for  $\mathcal{F}$  in  $\{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_k\}$  do
7:   if  $|\mathcal{S}| + |\mathcal{F}| \leq N$  then
8:      $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{F}$ 
9:   else
10:    Distances  $\leftarrow \text{ComputeCrowdingDistance}(\mathcal{F})$ 
11:    Sort  $\mathcal{F}$  by Distances in descending order
12:     $\mathcal{S} \leftarrow \mathcal{S} \cup \text{Top}(N - |\mathcal{S}|)$  samples from  $\mathcal{F}$ 
13:   end if
14: end for
Output  $\mathcal{S}$ 

```

Algorithm 3 ComputeCrowdingDistance

Input \mathcal{F}

```

1: Initialize Distance  $D$ 
2: for  $j$  in range( $D$ ) do
3:   Initialize  $D_j \leftarrow 0$ 
4: end for
5: for objective  $m$  in  $\{S_{\text{corr}}, S_{\text{tool}}\}$  do
6:   Sort  $\mathcal{F}$  by objective  $m$  in ascending order
7:   for  $j$  in range( $\mathcal{F}$ ) do
8:      $D_j \leftarrow D_j + \frac{m(j+1) - m(j-1)}{\text{Range}(m)}$ 
9:   end for
10: end for
Output  $D$ 

```

C Implementation Details

C.1 Preliminary Experiments Details

For preliminary experiments, mathematical task includes AIME, Numina-Math, and OpenR1-Math-220k; knowledge-intensive task includes 2Wiki-MultihopQA, HotpotQA, MuSiQue, and TriviaQA. For insufficient tool call analysis experiment, approximately 20.8% of failures can be corrected after continued reasoning. Hints used for injection can be found in Figure 9. For the analysis experiment on the tool call number differences, we sample 16 trajectories for each query.

Method	AIME24	AMC23	M500	2Wiki	Bamb	MSQ	Avg.
IKEA	<u>33.3</u>	30.0	24.4	30.0	28.8	31.5	29.7
SMART	6.7	20.0	33.6	19.0	33.6	12.0	20.8
AutoTIR	23.3	50.0	60.4	47.5	<u>53.6</u>	<u>31.0</u>	44.3
Tool-Star	30.0	55.0	68.2	64.5	47.2	22.0	<u>47.8</u>
Tool-Light	32.5	72.5	<u>75.8</u>	32.5	25.6	13.5	42.1
ET-Agent	53.3	<u>65.0</u>	84.6	<u>62.0</u>	58.4	18.5	57.0

Table 4: Results of **Conciseness** metric, with top two results highlighted in **bold** and underlined. Abbreviation: M500 (MATH500).

C.2 Training Details

Details in Action Space Exploration After data sampling, we select a total of 10.5k data for RFT. Our base model uses Qwen2.5-7B-Instruct (Yang et al., 2024), with a learning rate of 5e-6 and a batch size set to 32. We train for a total of three epochs. During training, we use DeepSpeed ZeRO-3 (Rasley et al., 2020) and FlashAttention2 (Dao, 2024).

Details in RL Training We use the VERL framework (Sheng et al., 2025) and train based on the ARPO algorithm (Dong et al., 2025d). We conduct three rounds of curriculum learning. The batch size is 64, we train for 3 epochs, the maximum length of the output text is 4096 tokens, and the learning rate is 1e-6. The initial values of both types of sigma are set to 0.1. All experiments are completed using 4 NVIDIA A800 GPUs. During training, we refer to FlashRAG (Jin et al., 2025b) and use Wikipedia documents for local retrieval. For each question, we retrieve 4 documents, and each document has a maximum length of 1200 characters.

C.3 Evaluation Details

For each test dataset, we select the first 500 samples for testing. For Mathematical-Reasoning Tasks, we use Google Search. For Knowledge-Intensive Tasks, we use local retrieval consistent with the training settings. We set the maximum call number for both the code compiler and the search engine to 6 times each. The maximum output sequence length for each round of reasoning is set to 4096 tokens.

D Complete Results of Behavioral Pattern Indicators

In this section, we present the complete results on all benchmarks shown in Figure 6. As depicted in Table 4, 5, and 6, ET-Agent outperforms most

Method	AIME24	AMC23	M500	2Wiki	Bamb	MSQ	Avg.
IKEA	10.0	42.5	63.2	33.8	35.0	14.2	33.1
SMART	3.3	17.5	39.1	21.6	40.8	12.1	22.4
AutoTIR	26.4	56.5	69.0	<u>64.6</u>	56.9	30.7	50.7
Tool-Star	30.0	<u>57.5</u>	<u>79.4</u>	65.8	<u>55.3</u>	24.4	<u>52.1</u>
Tool-Light	<u>33.3</u>	70.0	75.6	60.2	48.1	<u>25.5</u>	52.1
ET-Agent	43.3	70.0	84.5	63.2	54.3	23.3	56.4

Table 5: Results of **Successful Execution** metric, with top two results highlighted in **bold** and underlined. Abbreviation: M500 (MATH500).

Method	AIME24	AMC23	M500	2Wiki	Bamb	MSQ	Avg.
IKEA	0.17	0.75	2.19	1.49	2.10	0.47	1.20
SMART	0.13	0.33	1.07	1.20	0.98	0.25	0.66
AutoTIR	1.00	2.32	3.60	<u>3.61</u>	<u>3.85</u>	<u>1.49</u>	<u>2.65</u>
Tool-Star	0.42	0.76	1.56	2.24	1.34	0.57	1.15
Tool-Light	0.47	1.19	1.86	1.96	1.80	0.78	1.34
ET-Agent	<u>0.59</u>	1.23	<u>2.84</u>	6.34	5.62	1.81	3.07

Table 6: Results of **Reasoning Length** metric, with top two results highlighted in **bold** and underlined. Abbreviation: M500 (MATH500).

baseline methods on all metrics, and stably surpasses them in terms of average performance. This highlights the effectiveness of ET-Agent’s training process.

Dataset Type	Hints
KI & DS Tasks' Hints	Wait a minute, I might need to rethink this problem and then initiate a new query.
	Hold on, I think I should reconsider this question and generate a new query.
	But wait, I think I need to rethink how to solve it and then try a new search.
	Wait, maybe I should rethink this question and conduct a new search.
	But wait, maybe I should reconsider my approach and run a new search.
	Wait, I think rethinking the problem first and launching another query would help.
MR Task' Hints	Wait, perhaps I should use Python tools and rethink this problem.
	Wait, I think I should turn to Python tools and reassess the problem.
	Wait a moment, I might need to use Python tools and rethink this problem.
	But wait, I'll probably need a Python tool to help me reconsider this question.
	Hold on, I may need to use a Python tool to reconsider this question.
	Wait, I think I should revisit this problem and use a Python tool to help.

Figure 9: All hints used in Incorrect Trajectories Reflection stage.

E Prompt Templates

E.1 Prompt for Redundancy Judgment

Prompt for Redundancy Judgment

You are an expert analyst of tool-integrated reasoning tasks. Given a question, its golden answer, and a reasoning trajectory, your task is to examine the following content and determine whether the trajectory contains any redundant or ineffective tool calls.

[Question]
{question}
[Golden Answer]
{answer}
[Trajectory]
{trajectory}

Definitions:

- Redundant tool call: A call whose result is repetitive of previous content, unused in later reasoning, or does not contribute to the final answer.
- Ineffective tool call: A call that returns an irrelevant or unhelpful result providing no meaningful contribution to the reasoning process.

Your tasks:

1. Determine whether any redundant or ineffective tool calls exist in the trajectory.
2. If such calls exist, output only “<answer>yes</answer>”.
3. If none exist, output only “<answer>no</answer>”.

Output format:

Only output “yes” or “no”, enclosed within <answer> and </answer> tags, and do not output anything else.

E.2 Prompt for TIR Trajectory Sampling

Prompt for TIR Trajectory Sampling

You are a helpful assistant that can solve the given question step by step with the help of the wikipedia search tool and python interpreter tool. Given a question, you need to first think about the reasoning process in the mind and then provide the answer. During thinking, you can invoke the wikipedia search tool to search and python interpreter tool to calculate the math problem for fact information about specific topics if needed. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags respectively, and the search query and result are enclosed within <search> </search> and <result> </result> tags respectively. After receiving the search or python result, you should continue your reasoning process begin with <think>. For example, <think> This is the reasoning process. </think> <search> search query here </search> <result> search result here </result> <think> This is the reasoning process. </think> <python> python code here </python> <result> python interpreter result here </result> <think> This is the reasoning process. </think> <answer> The final answer is

answer here

</answer>. In the last part of the answer, the final exact answer is enclosed within \boxed{ } with latex format.^a

^aUnless otherwise specified, we use this prompt for reasoning during evaluation.

E.3 Prompt for First Step Optimization

Prompt for First Step Optimization

Role:
You are an expert in analyzing tool-integrated reasoning trajectory.

Task Description:
Your goal is to analyze a given trajectory to identify and locate the first occurrence with any redundant reasoning or redundant tool usage. If any redundancy is identified, you should find the earliest step that exists redundancy, and provide an analysis and a simplified, more efficient version for that step.

Redundancy Definitions:
1. Redundant Reasoning: The thought process includes unnecessary, overly complex, verbose, or repetitive content.
2. Redundant Tool Call: A tool is called when the necessary information is already sufficient, or the tool call's content is highly repetitive of a prior one seeking no new information.

Input Data:
Question: {question}
Trajectory:
{trajectory}

Output Format:
Your output must be one of the following two formats, with no additional explanations or text.

Scenario 1: The trajectory does not have any redundant reasoning or redundant tool usage. Only output "Step: no", and do not output anything else.

Scenario 2: The trajectory exists any redundant reasoning or redundant tool usage. Your output must be exactly three lines. The first line is the earliest step number with redundancy, the second line is the analysis of the redundancy in that step, and the third line is a simplified, more efficient version of that specific step. In this scenario, the output format is as follows:
Step: [an integer step number]
Analysis: [A concise analysis of the redundancy in that step]
Corrected Step: [the full, simplified step string]

For example:
In scenario 1, the output is "Step: no".

In scenario 2, the output format is as follows:
Step: [an integer step number]
Analysis: [A concise analysis of the redundancy in that step]
Corrected Step: <think> This is the reasoning process. </think> <search> search query here </search> <result> search result here </result> <think> This is the reasoning process. </think> <python> python code here </python> <result> python interpreter result here </result> <think> This is the reasoning process. </think> <answer> The final answer is
answer here
</answer>.

E.4 Prompt for Global Refinement

Prompt for Global Refinement

You are an expert in trajectory refinement. You will be given a question and a corresponding trajectory that uses a fixed format with tags such as "<think>", "<search>", "<result>", "<python>", and "<answer>".

Your task is to **refine ONLY the text inside "<think>...</think>" tags**, while keeping the trajectory's structure, tag order, and all other content unchanged.

Requirements:

1. **Do NOT add, remove, reorder, or rename any tags.**
The sequence of tags must remain exactly the same as in the original trajectory. For example, if the format is:
<think>...</think> <search>...</search> <result>...</result> <think>...</think> <answer>...</answer>
then the refined version must follow the exact same pattern.

2. **Only modify the content inside "<think>" tags.**

- * Remove redundant thoughts or irrelevant reasoning.
- * Eliminate obvious mistakes or unnecessary repetition.
- * Preserve the core reasoning and logical flow.
- * Keep the refined version concise, clear, and logically consistent.

3. **Do NOT alter anything inside "<search>", "<result>", "<python>", or "<answer>" tags.**
Their content, formatting, and text must remain identical.

4. **Keep the formatting and tags unchanged.**

Input

[Question]
{question}

[Original Trajectory]
{trajectory}

Output

Provide the **refined trajectory**, keeping the same structure and identical content outside the "<think>" tags, with only streamlined and corrected content inside each "<think>" section.

Your output:

E.5 Prompt for Self Correction

E.5.1 Prompt for First Flaw Identification

Prompt for First Flaw Identification

Role

You are an expert in analyzing tool-integrated reasoning trajectories.

Task Description

You will be given a reasoning trajectory that was produced by you in a previous turn. Your goal is to examine this trajectory step by step and identify the **first** step that contains any reasoning flaw or improper tool call. Once located, provide your analysis and practical suggestions for that step.

Flaw Categories

1. Flawed Reasoning:

- Logical Error: The thought process is incorrect, irrelevant, or does not contribute to solving the goal.
- Overthinking: The reasoning is unnecessarily long, repetitive, or contains redundant content.

2. Improper Tool Call:

- Redundant Call: A tool is called even though the available information is already sufficient, or the call unnecessarily repeats prior content.
- Inappropriate Call: The selected tool is unsuitable for the task, or the content passed to the tool is inappropriate.

Input

Question:
{question}

Trajectory:
{trajectory}

Output Requirements

Your response must contain **exactly two lines**, with no additional explanation, formatting, or text:

Line 1:

Step: [the integer index of the earliest flawed step]

Line 2:

Analysis: [a concise analysis and practical suggestions of the flaw in that step]

Among them, your analysis and practical suggestions should be written in a **single paragraph of plain text**. Do not use any headings, bullet points, numbering, or formatting. Just give a concise summary of the main flaws you found in the trajectory and how it could be improved, in continuous natural language.

For example:

Step: [an integer step number]

Analysis: [A paragraph with analysis and practical suggestions of the flaw]

Your output:

E.5.2 Prompt for First Flaw Modification

Prompt for First Flaw Modification

You are an expert in refining and completing tool-integrated reasoning trajectories.

Task Description:

You will be given:

1. A question
2. An incomplete trajectory
3. An analysis specifically about the **last step** in the trajectory

Your goal is to generate a **new, corrected final step** to replace the current last step based on the provided analysis. Only generate a new final step, do not generate any other content.

Input Data:

Question: {question}

Trajectory (incomplete):

{trajectory}

Analysis of the last step:

{analysis}

Output Format:

Your output must be exactly one line containing only the modified final step, in full.

For example:

your reasoning process. <search> search query here </search> <result> search result here </result> your reasoning process. <python> python code here </python> <result> python interpreter result here </result> your reasoning process. <answer> The final answer is

answer here

</answer>.

Your modified step:

E.6 Hints for Trajectory Reflection

For mathematical-reasoning tasks and knowledge-intensive tasks, we design different hints respectively. They can be found in Table 9.

F Case Study of Error Behavior Patterns

We provide a representative example for each of the four error patterns. They can illustrate the specific failure modes of the TIR model, as detailed in Table 7, 8, 9 and 10. **Red** contents represent the erroneous places.

Table 7: An example of Redundant Tool Call situation.

Example #1 from MuSiQue Dataset

Question:

What river does the system containing Umchabezi River serve as the mouth of?

Golden Answer:

Limpopo River

TIR Stratety

<think> To find out which river system the Umchabezi River serves as the mouth of, I'll need to search for information on this river. </think>

<search> Wikipedia search query: Umchabezi River mouth </search>

<result> The Umchabezi River is a tributary of the Mzingwane River in Beitbridge District and Gwanda District, Zimbabwe. The main dam on the river is Makado Dam, which supplies water for commercial irrigation. </result>

<search> Search for Mzingwane River mouth </search>

<result> The Mzingwane River, formerly known Umzingwane River or Umzingwani River is a major left-bank tributary of the Limpopo River in Zimbabwe. It rises near Fort Usher, Matobo District, south of Bulawayo and flows into the Limpopo River near Beitbridge, downstream of the mouth of the Shashe River and upstream of the mouth of the Bubye River. The Mzingwane River is an ephemeral river with flow generally restricted to the months when rain takes place (November to March), with most flow recorded between December and February, except where it has been modified by dam operations. </result>

<think> According to the search results, the river system that the Umchabezi River serves as a tributary of is the Limpopo River. Let's verify this with Python code. </think>

<python>

```
print('Limpopo River')
```

</python>

<result> Limpopo River </result>

<answer> Therefore, the final answer is Limpopo River. </answer>

Table 8: An example of Aborted Tool Execution situation.

Example #2 from MATH500 Dataset
<p>Question: Let $T = \frac{7}{8}$. The number T can be expressed as a reduced fraction $\frac{m}{n}$, where m and n are positive integers whose greatest common divisor is 1. The equation has two distinct real solutions. Compute the lesser of these two solutions.</p> <p>Golden Answer: -8</p>
<p>TIR Strategy</p> <p><think> We are given: $T = \frac{7}{8}$ - We express T as a reduced fraction $\frac{m}{n}$, so $m = 7, n = 8$ - Then we consider the quadratic equation: $x^2 + (m+n)x + mn = 0$ Substitute $m = 7, n = 8$: $m+n = 7+8 = 15$ $mn = 7 \cdot 8 = 56$ So the equation becomes: $x^2 + 15x + 56 = 0$ We are told this equation has two distinct real solutions, which it does because the discriminant is positive. The discriminant D of $ax^2 + bx + c = 0$ is: $D = b^2 - 4ac$ Here, $a = 1, b = 15, c = 56$: $D = 15^2 - 4(1)(56) = 225 - 224 = 1 > 0$ Yes — two distinct real roots. Now compute the roots using the quadratic formula: $x = \frac{-b \pm \sqrt{D}}{2a} = \frac{-15 \pm \sqrt{1}}{2} = \frac{-15 \pm 1}{2}$ Thus, - One root: $\frac{-15+1}{2} = \frac{-14}{2} = -7$- Other root: $\frac{-15-1}{2} = \frac{-16}{2} = -8$ These are equal? Wait! That's not right! Wait — let's double-check: $\frac{-15 \pm 1}{2}$ That gives: $(-15+1)/2 = -14/2 = -7$- $(-15-1)/2 = -16/2 = -8$ So the two roots are -7 and -8. So they are distinct. But now — what is the lesser of the two solutions? Clearly, $-8 < -7$, so the lesser solution is -8. But wait — let me confirm everything again carefully. Given: $T = \frac{7}{8}$- So $m = 7, n = 8$- Form the equation: $x^2 + (m+n)x + mn = 0$- Which is $x^2 + 15x + 56 = 0$- Discriminant: $15^2 - 4(1)(56) = 225 - 224 = 1$- Roots: $\frac{-15 \pm 1}{2}$- So roots: $\frac{-14}{2} = -7, \frac{-16}{2} = -8$- Yes, both valid; $-8 < -7$ Therefore, the lesser solution is -8. Let's use Python to check this answer. </think><python></p> <pre>from sympy import symbols, Eq, solve m, n = 7, 8 expr = symbols('x') equation = Eq(x**2 + (m+n)*x + m*n, 0) roots = solve(equation, x) min_root = min(roots) min_root.evalf()</pre> <p></python><result> Traceback (most recent call last): File <string>, line 4, in <module> NameError: name 'x' is not defined <think> Oops — I forgot to define x. Let me fix. </think> <python></p> <pre>from sympy import symbols, Eq, solve x = symbols('x') equation = Eq(x**2 + (7+8)*x + 7*8, 0) roots = solve(equation, x) min_root = min(roots) print(f"{min_root.evalf()=}")</pre> <p></python> <result> min_root.evalf()=-8.000000000000000 </result> <think> Hence, the lesser solution is:</think><answer> -8. </answer></p>

Table 9: An example of Insufficient Tool Call situation.

Example #3 from 2WikiMultiHopQA Dataset
Question: Who is Sancho Ramírez's maternal grandfather?
Golden Answer: Bernard-Roger, Count of Bigorre
TIR Strategy
<think> To determine who Sancho Ramírez's maternal grandfather was, we would need to know some biographical details about him, specifically focusing on his mother's side of the family. However, without access to specific historical records or more context about which Sancho Ramírez is being referred to, it's challenging to provide an accurate answer directly. Let's assume we're looking for information based on a notable figure named Sancho Ramírez. I will proceed by searching for general information on famous figures with this name. </think> <search> Sancho Ramírez biography </search> <result> Sancho Ramírez was an illegitimate son of King Ramiro I of Aragon and Amua, the firstborn and brother of his namesake who would inherit the throne and reign as Sancho Ramírez. Even though he could not inherit the throne because his father had legitimate issue, he was named count at an early age and was a prominent member of the curia regis first appearing in a charter dated 1049, suspected of being false, as a witness to a donation made by his father to the Monastery of San Victorin. In this charter, he confirms as Šancius Ranimiri regis filius primogenitus (Sancho, the firstborn son of King Ramiro) followed by his brother and namesake, Sancho, who confirms as Šancius Ranimiri regis filius prolis Ermisendis regine (Sancho son of King Ramiro and Queen Ermesinda). </result> <think> The provided summary doesn't contain explicit information about Sancho Ramírez's maternal grandfather. Further targeted searches may be required to find this particular detail. Based on the available data, there isn't enough evidence to identify Sancho Ramírez's maternal grandfather clearly. </think>
<answer> [Insufficient information] </answer>

Table 10: An example of Erroneous Reasoning Process situation.

Example #4 from MATH500 Dataset
Question: What is the smallest positive perfect cube that can be written as the sum of three consecutive integers?
Golden Answer: 27
TIR Strategy
<think> To find the smallest positive perfect cube that can be written as the sum of three consecutive integers, let's start by expressing the sum of three consecutive integers: $n - 1$, n , and $n + 1$ which equals $3n$. So we're looking for the smallest positive perfect cube that can be expressed as $3n$. A perfect cube is a number of the form k^3 where k is an integer. We need $k^3 = 3n$, implying that k^3 must be divisible by 3. Let's check the smallest values for k such that k^3 is a multiple of 3. </think> <search> smallest perfect cubes that are multiples of 3 </search> <result> Perfect magic cube In mathematics, a perfect magic cube is a magic cube in which not only the columns, rows, pillars, and main space diagonals, but also the cross section diagonals sum up to the cube's magic constant. Perfect magic cubes of order one are trivial; cubes of orders two to four can be proven not to exist, and cubes of orders five and six were first discovered by Walter Trump and Christian Boyer on November 13 and September 1, 2003, respectively. A perfect magic cube of order seven was given by A. H. Frost in 1866, and on March 11, 1875, an article was published in the Cincinnati Commercial newspaper on the discovery of a perfect magic cube of order 8 by Gustavus Franklin. </result>
<think> It seems that the smallest value of k cannot be found through searching. Let's try using Python. </think> <python>

```
# Check the smallest k that makes k^3 a multiple of 3
for k in range(1, 10):
    if (k**3) % 3 == 0:
        smallest_cube_multiple_of_3 = k**3
        break

smallest_cube_multiple_of_3
```

</python>
<result> 27 </result>
<think> From our calculations and understanding, the smallest positive perfect cube that can be written as the sum of three consecutive integers is 27. Since $27 = 3^3$ and it fits the requirement as 3×9 , it is indeed the correct answer. </think>
<answer> The smallest positive perfect cube that can be written as the sum of three consecutive integers is [27]. </answer>