

AdaTIR: Adaptive Tool-Integrated Reasoning via Difficulty-Aware Policy Optimization

Anonymous ACL submission

Abstract

Tool-Integrated Reasoning (TIR) has significantly enhanced the capabilities of Large Language Models (LLMs), yet current agents tend to exhibit cognitive offloading, redundantly invoking external tools even for simple tasks. In this paper, we suggest that true agentic intelligence requires not just tool invocation, but the adaptive wisdom to discern when to use them. We propose AdaTIR, a framework that shifts the paradigm from static tool invocation to difficulty-aware reasoning internalization. By introducing a difficulty-aware efficiency reward, AdaTIR dynamically adjusts tool budgets based on task complexity—internalizing reasoning for simple tasks while selectively invoking tools for complex tasks. Furthermore, we identify a sign reversal problem where tool penalties outweigh correctness rewards, mistakenly penalizing correct rollouts with negative advantages. To resolve this, we propose Clipped Advantage Shaping (CAS), which ensures that correctness remains the primary objective while using efficiency as a secondary constraint. Empirical results demonstrate that AdaTIR reduces tool calls by up to 97.6% on simple tasks and 28.2% on complex challenges while maintaining or enhancing accuracy. Notably, AdaTIR successfully internalizes reasoning, outperforming baselines by 4.8% on AIME 2024 even when tool access is strictly disabled.

1 Introduction

Large Language Models (LLMs)—such as OpenAI-o1 (OpenAI et al., 2024) and DeepSeek-R1 (DeepSeek-AI et al., 2025)—have achieved remarkable milestones in complex reasoning, manifesting superior capabilities in tasks ranging from mathematical problem-solving (Patel et al., 2024) to repository-level code generation (Jimenez et al., 2024) and open-domain question answering (Kwiatkowski et al., 2019). To further transcend the limitations of parametric knowledge,

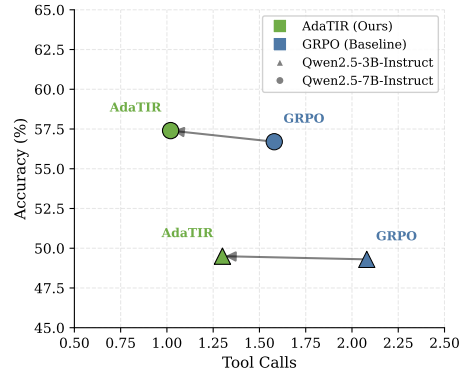


Figure 1: Accuracy versus Tool Calls for Qwen2.5 under different training strategies. GRPO and AdaTIR denote the baseline and our proposed method, respectively. AdaTIR consistently achieves a superior efficiency-accuracy trade-off across both 3B and 7B backbones by internalizing reasoning capabilities.

Tool-Integrated Reasoning (TIR) has emerged as a powerful paradigm (Gou et al., 2024; Li et al., 2025c; Feng et al., 2025; Singh et al., 2025; Xue et al., 2025), augmenting Large Language Models (LLMs) with external executors to tackle problems intractable for pure-text generation (Lin and Xu, 2025). However, this augmentation comes with a hidden cognitive cost, current agents exhibit a pathological tendency toward “cognitive offloading” (Wang et al., 2025a; Xu et al., 2024). By frequently invoking tools for simple tasks, they not only incur unnecessary latency but also risk degrading their intrinsic reasoning consistency through environmental noise. The fundamental challenge lies in selectively internalizing tool-independent logic without compromising the robustness needed for genuinely complex tasks.

To address these efficiency bottlenecks, recent research has explored ranging from inference-time budget constraints (Liu et al., 2025) to training-time reward shaping. However, these approaches often fail to address the root cause of tool depen-

065 dency. Inference-time constraints function as shallow prompt-level limits, often leading to context
066 drift or reasoning dead-ends when the model lacks
067 the internalized logic to proceed without tools (Liu
068 et al., 2024; Shinn et al., 2023). Alternatively,
069 training-based methods like OTC impose indiscriminate penalties on tool invocation. This creates a
070 fundamental dilemma: a penalty strong enough
071 to suppress redundancy on simple tasks often in-
072 advertently impairs the model’s ability to solve
073 complex tasks that genuinely require external as-
074 sistance. Furthermore, approximating optimal bud-
075 gets based on biased group-minima estimates often
076 leads to sub-optimal convergence where tool redun-
077 dancy persists.

080 To address these limitations, we propose
081 **AdaTIR**, a framework that establishes a difficulty-
082 aware policy to adaptively regulate tool invocation
083 frequency based on task complexity. Our core in-
084 sight is that tool invocation is inherently difficulty-
085 dependent. Different from static penalties that in-
086 discriminately suppress actions, we suggest that
087 constraints should be dynamic: stricter budgets on
088 simple tasks compel the model to internalize rea-
089 soning capabilities and mitigate “cognitive offload-
090 ing”, while relaxed constraints on complex tasks
091 ensure the model retains the robustness to leverage
092 external tools when necessary. Consequently, this
093 paradigm enables the model to align computational
094 expenditure with task difficulty, effectively pushing
095 the cost-performance Pareto frontier.

096 The implementation of this strategy, however,
097 reveals a critical algorithmic challenge: naive re-
098 ward shaping inherently risks sign reversal, where
099 efficiency penalties can inadvertently outweigh cor-
100 rectness rewards. Specifically, as training pro-
101 gresses, rollouts within a group exhibit consistent
102 correctness. This phenomenon dilutes the accuracy-
103 based advantage signal, allowing the auxiliary ef-
104 ficiency signal to disproportionately influence the
105 optimization objective. Consequently, even rollouts
106 that yield correct answers could be assigned a neg-
107 ative advantage if their tool invocation exceeds the
108 group average, creating an objective misalignment
109 that suppresses valid reasoning paths and induces
110 training instability. To address this, we propose
111 **Clipped Advantage Shaping (CAS)**, a mechanism
112 that directly reformulates the advantage term rather
113 than modifying the scalar reward. By integrating
114 a clipped efficiency penalty as an auxiliary signal,
115 CAS effectively encourages the prioritization of
116 correctness signals over efficiency signals, causing

the model to optimize efficiency without suffering
from the mode collapse (Wang et al., 2025c; Cui
et al., 2025) observed in conventional methods.

Our contributions are summarized as follows:

- We propose AdaTIR, a framework that es-
tablishes a difficulty-aware policy to adap-
tively regulate tool invocations tailored to task
complexity, effectively suppressing redundant
“cognitive offloading” on simple tasks.
- We introduce Clipped Advantage Shaping
(CAS), a mechanism that mathematically re-
formulates the advantage term to maintain
sign consistency, resolving the training insta-
bility and mode collapse often observed in
naive reward shaping.
- We demonstrate that our approach signifi-
cantly enhances efficiency across varying diffi-
culty levels, reducing tool calls by up to 97.6%
(Figure 1) on simple tasks while maintaining
robust performance on complex challenges.
Notably, our method exhibits superior rea-
soning capabilities even when tool access is
strictly disabled, indicating successful reason-
ing internalization.

2 Preliminaries

Tool-Integrated Reasoning (TIR). In the TIR
paradigm, the large language model acts as a rea-
soning engine that interleaves text generation with
external tool interactions (Wang et al., 2023; Shao
et al., 2024). Given a task q , the reasoning process
of the policy π_θ forms a trajectory τ consisting of
sequential reasoning steps and environment feed-
backs:

$$\tau = (w_1, c_1, o_1, w_2, c_2, o_2, \dots, w_K, c_K, o_K, \hat{y}) \quad (1)$$

where w_k represents the k -th reasoning thought, c_k
is the tool invocation command, and o_k is the cor-
responding observation returned by code execution.

Group Relative Policy Optimization (GRPO).
We adopt GRPO as our base optimization frame-
work. In contrast to traditional Actor-Critic (Konda
and Tsitsiklis, 1999) methods that require a value
function, GRPO estimates the baseline directly
from group scores, reducing memory overhead.

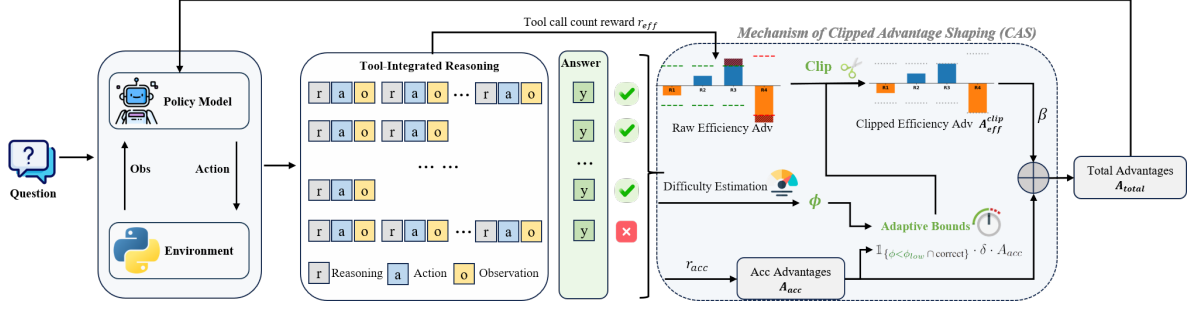


Figure 2: **Overview of the proposed framework.** Efficiency signals are conditionally injected only for correct rollouts on **easy** tasks (where difficulty estimation ϕ falls below threshold ϕ_{low}). These auxiliary signals are explicitly clipped by $\delta|A_{acc}|$ ($0 < \delta < 1$), to ensure efficiency optimization remains secondary to correctness, thereby maintaining sign consistency.

The training objective is defined as:

$$\mathcal{J}_{GRPO} = \frac{1}{G} \sum_{i=1}^G \frac{1}{|\tau_i|} \sum_{t=1}^{|\tau_i|} \left(\min(r_{i,t}, \hat{A}_{i,t}, \text{clip}(r_{i,t}, 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t}) - \beta D_{KL}(\pi_{\theta} \parallel \pi_{ref}) \right) \quad (2)$$

where the importance sampling ratio $r_{i,t}$ is explicitly given by:

$$r_{i,t} = \frac{\pi_{\theta}(\tau_{i,t} \mid q, \tau_{i,<t})}{\pi_{old}(\tau_{i,t} \mid q, \tau_{i,<t})} \quad (3)$$

Here, $\tau_{i,t}$ denotes the t -th token in trajectory τ_i , and $\hat{A}_{i,t}$ represents the normalized advantage computed within group i .

3 Method

In this section, we present AdaTIR, a framework designed for optimizing effectiveness and efficiency by invoking tools only when necessary. The overall workflow, including our difficulty-aware reward mechanism and the Clipped Advantage Shaping (CAS) module, is illustrated in Figure 2. We begin by introducing a difficulty-aware reward shaping mechanism, where task difficulty is estimated by the accuracy of response within each group, to encourage the policy model to reduce tool invocations for simple task (§3.1). Subsequently, we introduce Clipped Advantage Shaping, a mechanism that reformulates the advantage term to maintain sign consistency. By limiting efficiency penalties to a secondary role relative to correctness rewards, CAS mitigates the objective misalignment and reduces the risk of mode collapse (§3.2).

3.1 Difficulty-Aware Efficiency Reward for Strategic Reasoning Internalization

To address these challenges, we propose a Difficulty-Aware Reward mechanism that incentivizes the policy model to learn adaptively budget allocation for tool invocation based on task complexity. The core idea is to penalize the model for frequent tool invocations on simple tasks—thereby compelling the model to internalize the reasoning capabilities. Meanwhile, we adaptively relax this constraint for complex tasks where external tool calls is indispensable for maintaining effectiveness.

To mitigate redundant tool invocations, we propose a Difficulty-Aware Reward mechanism that adaptively allocates tool budgets based on task complexity. Different from prior works that focus primarily on correctness (Li et al., 2025c; Xue et al., 2025), our approach compels reasoning internalization for simple tasks while maintaining the flexibility to invoke tools for complex tasks.

Extrinsic Difficulty Estimation. Specifically, we define the difficulty $\phi_q \in [0, 1]$ of task q as the complement of the group success rate:

$$\phi_q = 1 - \frac{1}{G} \sum_{i=1}^G \mathbb{I}(\text{passed}(gt, pred_i)) \quad (4)$$

where G is the group size, gt is the ground-truth answer, and $pred_i$ is the prediction from trajectory τ_i . In this formulation, $\phi_q \rightarrow 0$ denotes a simple task, while $\phi_q \rightarrow 1$ signifies a challenging task.

Difficulty-Aware Efficiency Reward shaping. Based on the estimated difficulty, we apply a penalty to successful trajectories that rely on redundant tool use. We first define a sine-based penalty factor $f(\tau_i)$ to quantify the tool invocation cost:

$$f(\tau_i) = \sin\left(\frac{\pi}{2} \cdot \frac{N(\tau_i)}{N_{max}}\right) \quad (5)$$



There exist real numbers x and y , both greater than 1, such that $\log_x(y^x) = \log_y(x^{4y}) = 10$. Find xy .

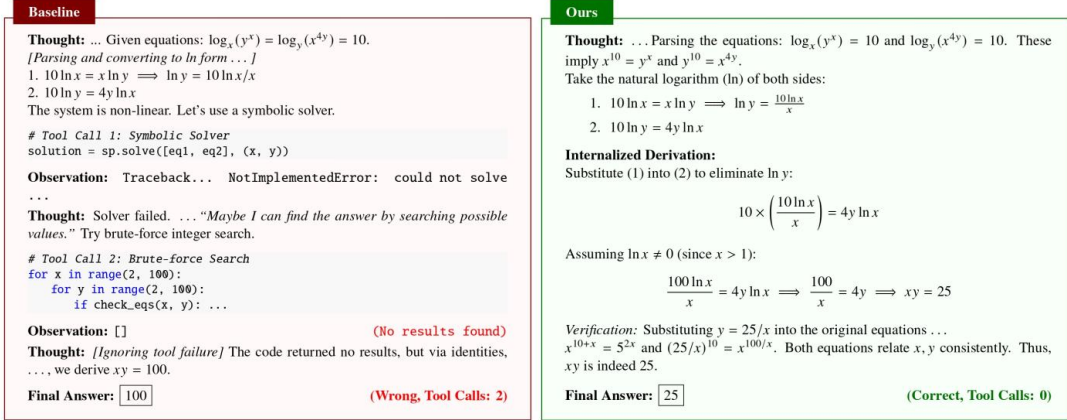


Figure 3: Comparison of reasoning processes between a standard tool-integrated baseline and our approach. (Left) The baseline model relies on external Python tools but fails to recover after the symbolic solver returns an error. It then attempts an irrelevant search and ends with an incorrect answer. (Right) Our approach correctly solves the problem by performing step-by-step algebraic deduction internally, reaching the correct result without depending on external code execution.

where $N(\tau_i)$ is the number of tool calls, and N_{\max} represents the tool call limit. To prioritize correctness and prevent reward hacking, this penalty is conditionally applied only to correct trajectories ($r_{\text{acc},i} = 1$) on simple tasks ($\phi_q < \phi_{\text{low}}$):

$$r_{\text{eff},i} = \begin{cases} -\lambda \cdot f(\tau_i), & \text{if } r_{\text{acc},i} = 1 \text{ and } \phi_q < \phi_{\text{low}} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where λ is a scaling hyperparameter and ϕ_{low} is the difficulty threshold. The final reshaped reward for the i -th trajectory is thus defined as:

$$r_{\text{total},i} = r_{\text{acc},i} + r_{\text{eff},i} \quad (7)$$

The final reshaped reward for the i -th trajectory is defined as $r_{\text{total},i} = r_{\text{acc},i} + r_{\text{eff},i}$. By applying the efficiency penalty only to correct trajectories, AdaTIR effectively prevents reward hacking and ensures that correctness remains the primary objective. This design encourages the model to internalize reasoning for simple tasks while maintaining the flexibility to selectively invoke tools for complex ones (see Appendix E.1 for details).

3.2 Clipped Advantage Shaping for Stable Training

Training Instability of Reward Shaping. Directly incorporating $r_{\text{eff},i}$ into the GRPO objective often induces significant training instability (Lin and Xu, 2025). Substituting the total reward $r_{\text{total},i}$ into the

standard advantage formulation yields:

$$A_i = \frac{(r_{\text{acc},i} + r_{\text{eff},i}) - \text{mean}(\mathbf{r}_{\text{acc}} + \mathbf{r}_{\text{eff}})}{\text{std}(\mathbf{r}_{\text{acc}} + \mathbf{r}_{\text{eff}})} \quad (8)$$

When a task is simple or the model reaches performance saturation, all rollouts within a group could be correct ($r_{\text{acc},i} = 1$). In such scenarios, the accuracy signal becomes negligible in the numerator, and the advantage estimation degenerates into a form driven exclusively by efficiency:

$$A_i \approx \frac{r_{\text{eff},i} - \text{mean}(\mathbf{r}_{\text{eff}})}{\text{std}(\mathbf{r}_{\text{eff}})} \quad (9)$$

This mistakenly assigns negative advantages to correct trajectories with above-average tool invocation. Furthermore, if the variance of \mathbf{r}_{eff} is minimal, the advantage estimation becomes extremely sensitive, where slight variations in tool invocation yield out-sized advantages. This numerical instability yields contradictory learning signals and often triggers premature mode collapse.

Clipped Advantage Shaping (CAS). To resolve training instability, we propose CAS to decouple the estimation of effectiveness and efficiency. Specifically, the accuracy advantage A_i^{acc} and the raw efficiency advantage A_i^{eff} are computed by applying the group normalization to $r_{\text{acc},i}$ and $r_{\text{eff},i}$, respectively. To maintain stability, the CAS advantage is formulated as:

$$A_i^{\text{CAS}} = A_i^{\text{acc}} + r_{\text{acc},i} \cdot \beta \cdot A_i^{\text{eff, clip}} \quad (10)$$

Method	AIME 2024		AIME 2025		AMC23		GSM8K	
	Acc (↑)	ATC (↓)	Acc (↑)	ATC (↓)	Acc (↑)	ATC (↓)	Acc (↑)	ATC (↓)
Qwen2.5-3B-Instruct	10.0	–	6.7	–	38.0	–	50.0	–
Qwen2.5-3B-Instruct-TIR	6.0	0.79	1.3	0.71	30.0	1.02	36.0	1.58
Qwen2.5-3B-Instruct _(GRPO)	26.3	2.52	24.2	2.58	60.2	2.02	86.5	1.19
AdaTIR-3B _(Ours)	29.2	1.63	23.8	2.01	58.0	1.35	87.1	0.22
Qwen2.5-7B-Instruct	12.3	–	10.0	–	52.0	–	61.9	–
Qwen2.5-7B-Instruct-TIR	5.0	0.48	5.4	0.36	21.3	0.51	17.3	0.51
Qwen2.5-7B-Instruct _(GRPO)	33.8	2.02	27.1	2.01	74.7	1.44	91.0	0.83
AdaTIR-7B _(Ours)	37.1	1.45	27.3	1.57	72.3	1.02	92.8	0.02

Table 1: Performance Comparison on Various Mathematical Reasoning Benchmarks. We report Avg@16 Accuracy and Average Tool Calls (ATC). **Bold** indicates the best performance.

where the binary mask $r_{\text{acc},i} \in \{0, 1\}$ ensures that efficiency updates are restricted only to correct trajectories, and $\beta > 0$ scales the intensity of the efficiency incentive. The clipped efficiency signal is defined as:

$$A_i^{\text{eff, clip}} = \text{clip}(A_i^{\text{eff}}, -\delta|A_i^{\text{acc}}| - \eta, \delta|A_i^{\text{acc}}| + \eta) \quad (11)$$

where $\delta \in [0, 1)$ and $\eta > 0$ control the clipping range relative to the correctness signal.

This formulation ensures robust optimization by upholding the priority of correctness signals. By bounding the efficiency incentive relative to A_i^{acc} , CAS **effectively preserves positive advantages for correct rollouts**, preventing the policy update from being misguided by negative advantages assigned to correct responses. Simultaneously, the clipping mechanism suppresses advantage inflation, ensuring that auxiliary feedback remains stable and consistent across varying task difficulties. We provide a formal analysis of the sign preservation and optimization logic of CAS in Appendix E.2.

4 Experiments

4.1 Experimental Setup

Training We implement AdaTIR on Qwen2.5-3B/7B-Instruct backbones (Qwen et al., 2025) using the veRL framework (Sheng et al., 2025) and SandboxFusion (Bytedance-Seed-Foundation-Code-Team et al., 2025). We utilize the dataset ReTool-SFT¹ for the SFT stage and DAPO-Math-17k² for the RL stage. Following the two-stage

¹<https://huggingface.co/datasets/JoeYing/ReTool-SFT>

²<https://huggingface.co/datasets/BytedTsinghua-SIA/DAPO-Math-17k>

paradigm (SFT and RL) established in Feng et al. (2025), we set both the maximum sequence and response lengths to 4096. For the RL phase, we impose a tool invocation limit of $N_{\text{max}} = 4$. While our baseline follows the standard ReTool configuration, AdaTIR introduces a difficulty threshold $\phi_{\text{low}} = 0.8$, a balance coefficient $\beta = 0.9$, and a penalty factor $\delta = 0.99$. To ensure a strictly controlled comparison, all other hyperparameters remain consistent with the original ReTool recipe (details in Appendix A).

Evaluation We evaluate AdaTIR on AIME 2024/2025, AMC 23, and GSM8K (Cobbe et al., 2021). During inference, we set the decoding temperature to $T = 1.0$ and top- $p = 0.6$, reporting **Avg@16 Accuracy** and **Average Tool Calls (ATC)**. Additionally, we adopt **Tool Productivity (TP)** (Wang et al., 2025a)—calculated as Accuracy normalized by ATC—to quantify reasoning efficiency. To assess the extent of reasoning internalization, we also conduct a budget sensitivity analysis by varying the maximum allowed tool calls $B \in \{0, 1, 2, 3, 4\}$.

4.2 Main Results

AdaTIR Maintains Competitive Accuracy with Significantly Reduced Tool Invocations. Results in Table 1 and Figure 4 show that AdaTIR achieves competitive or superior performance across all benchmarks, with accuracy margins ranging from -2.4% to $+3.3\%$. Crucially, this is attained while reducing Average Tool Calls (ATC) by 22.1% to 97.6%. On complex challenges like AIME 2024 and 2025, AdaTIR-7B preserves accuracy—notably improving by 3.3% absolute on

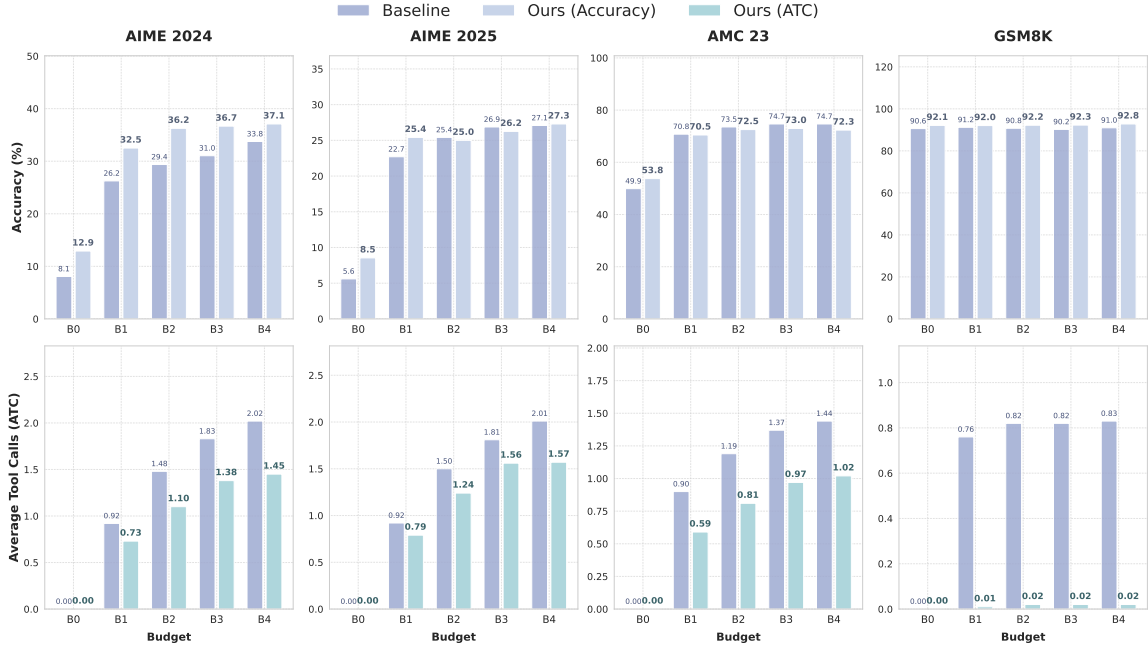


Figure 4: **Reasoning accuracy and tool invocation under varying tool budgets B .** The horizontal axis denotes the maximum allowed tool calls, ranging from B0 ($B = 0$) to B4 ($B = 4$). The top row presents the reasoning accuracy (reported as Avg@16), while the bottom row shows the Average Tool Calls (ATC). Our method consistently outperforms the baseline across all budget levels, achieving higher accuracy with significantly fewer tool invocations. Notably, even at B0 where no tool invocation is permitted, our method exhibits superior performance; this indicates that our approach facilitates the internalization of reasoning capabilities and mitigates cognitive offloading, enabling the model to maintain robust performance even without external tool calls

AIME 2024 (33.8% \rightarrow 37.1%)—while cutting tool use by up to **28.2%** (2.02 \rightarrow 1.45). This efficiency gain is even more pronounced in simpler tasks like GSM8K, where AdaTIR attains a near-zero ATC of **0.02** (a **97.6%** reduction from 0.83) alongside a 1.8% accuracy gain. These results indicate that our optimization strategy effectively prunes redundant computation without compromising reasoning quality.

Internalization of Reasoning Capabilities and Mitigation of Cognitive Offloading. Budget sensitivity analysis (Figure 4) reveals that AdaTIR consistently outperforms the baseline across all budget levels $B \in \{0, \dots, 4\}$. The gap is most evident at **B0**, where tool calls are explicitly denied by the environment (see Appendix A.3). Forced into pure natural language deduction, AdaTIR exhibits a significant lead (e.g., +4.8% absolute on AIME 2024), demonstrating successful **reasoning internalization**. Rather than developing persistent tool dependency, the model distills tool-augmented capabilities into its own parameters, enabling adaptive reasoning under resource constraints.

Furthermore, AdaTIR effectively **mitigates cog-**

nitive offloading—the over-reliance on external assistance for simple steps. While the baseline persistently invokes tools on GSM8K (ATC of 0.83 at $B = 4$), AdaTIR achieves peak accuracy with a near-zero ATC of **0.02**. This adaptive strategy—internalizing simpler logic while reserving tools for high-complexity segments—reduces overhead and enhances robustness. **Notably, this logic generalizes beyond mathematical reasoning:** preliminary results on search-based QA (Search-R1) confirm that AdaTIR similarly internalizes common-knowledge retrieval, invoking external search only when necessary (see Appendix B).

5 Analysis

5.1 Comparison with OTC

Experimental Setup for Direct Comparison with OTC. Since the official source code for OTC (Wang et al., 2025a) is non-public, we conducted an independent evaluation by adapting AdaTIR to their specific benchmarks. A critical alignment is the baseline: we include ToRL (Li et al., 2025c), the primary baseline in the original OTC study. Furthermore, while OTC discusses $N_{max} = 3$, the authors explicitly note that multiple

Method	N_{max}	AIME 24			AIME 25		
		EM (\uparrow)	ATC (\downarrow)	TP (\uparrow)	EM (\uparrow)	ATC (\downarrow)	TP (\uparrow)
ToRL-GRPO	3	23.3	2.2	10.6	23.3	2.3	10.1
OTC-GRPO	3	20.0 (\downarrow 14.2%)	1.1 (\downarrow 50.0%)	18.2 (\uparrow 71.7%)	20.0 (\downarrow 14.2%)	1.1 (\downarrow 52.2%)	18.2 (\uparrow 80.2%)
ToRL-GRPO	1	26.7	0.9	29.7	26.7	1.0	26.7
AdaTIR (Ours)	1	33.3 (\uparrow 24.7%)	0.5 (\downarrow 44.4%)	66.6 (\uparrow 124.2%)	26.7	0.7 (\downarrow 30.0%)	38.1 (\uparrow 42.7%)

Table 2: Comparison with OTC and ToRL on the Qwen2.5-1.5B backbone. We report Exact Match (EM), Average Tool Calls (ATC), and Tool Productivity (TP). **Note on Baseline Alignment:** While OTC results are reported under $N_{max} = 3$, we evaluate our method and the ToRL baseline under a stricter budget ($N_{max} = 1$) to strictly control for training stability and isolate the reasoning internalization capability.

tool calls often trigger training instability. Thus, to ensure a stable and strictly controlled comparison, we set $N_{max} = 1$ for AdaTIR, OTC, and ToRL. This ensures that all methods are evaluated on their capacity to internalize reasoning under the same constrained budget (see Appendix A.2 for implementation details).

Method	AIME 2024		AIME 2025	
	EM (\uparrow)	ATC (\downarrow)	EM (\uparrow)	ATC (\downarrow)
Vanilla GRPO	26.7	0.90	26.7	0.97
w/ Reward Shaping	30.0	0.33	20.0	0.27
w/ Clipped Advantage Shaping	33.3	0.50	26.7	0.70

Table 3: Ablation of CAS on AIME benchmarks comparing Exact Match (EM) and Average Tool Calls (ATC) against difficulty-aware reward shaping.

AdaTIR Outperforms OTC and ToRL in both Accuracy and Tool Productivity. As summarized in Table 2, AdaTIR demonstrates a decisive advantage over both OTC and ToRL. First, **AdaTIR eliminates the accuracy-efficiency trade-off:** while OTC consistently suffers from performance degradation (e.g., a 14.2% drop in EM), our method improves AIME 24 accuracy by 6.6% absolute over the baseline while pruning tool calls by 44.4%. Second, **AdaTIR significantly enhances Tool Productivity (TP):** on AIME 24, our TP is 124.2% higher than ToRL and over $3.6\times$ higher than OTC. These results confirm that AdaTIR makes tool invocation more efficient while maintaining high reasoning performance.

5.2 Ablation Study of CAS: Ensuring Training Stability

To verify the necessity of the **Clipped Advantage Shaping** (CAS) mechanism introduced in Section 3.2, we conduct an ablation study comparing it against the **difficulty-aware reward shaping** approach formulated in Section 3.1. As illustrated

in Figure 5, the results provide clear evidence that CAS is not merely an optimization but a critical prerequisite for training stability in agentic reasoning tasks. The experimental configuration remains consistent with the comparative study in Section 5.1, utilizing the ToRL baseline.

CAS Prevents Gradient Explosion and Stabilizes Training.

Without advantage clipping, outlier trajectories with anomalous tool-call counts generate disproportionately large advantage signals. As shown in the left panel of Figure 5, this triggers a severe **gradient explosion** near step 1,100 in the difficulty-aware reward shaping baseline, leading to training collapse. In contrast, the CAS curve (blue) remains stable throughout the process. This stability is further reflected in the tool invocation patterns in Table 3; while direct reward shaping causes a drastic and uncontrolled ATC drop, CAS provides a tempered signal that prevents the policy from collapsing into suboptimal states where tool use is prematurely or excessively suppressed.

CAS Preserves Reasoning Performance by Balancing Efficiency and Accuracy.

The results in Table 3 show that CAS does more than stabilize training; it also maintains superior reasoning accuracy. Different from the difficulty-aware reward shaping mechanism, which suffers from significant performance decay, CAS achieves a higher Exact Match (EM) score on AIME 2024 (33.3% vs. 26.7%) and remains stable on AIME 2025. This gain is largely due to the moderate penalty CAS imposes: by avoiding the over-suppression of tool calls, CAS ensures the model can still leverage external tools for the most complex reasoning steps. Consequently, CAS finds a better balance than either the vanilla baseline or the unclipped reward-shaping variant.

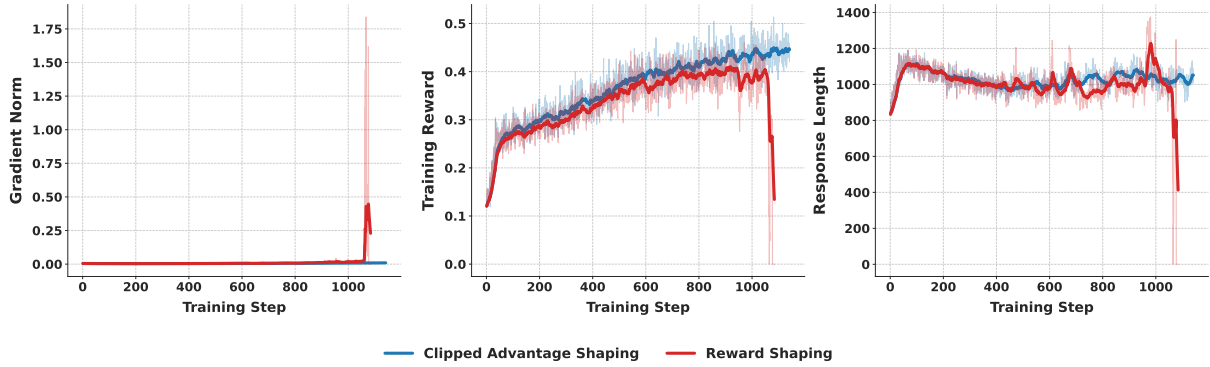


Figure 5: Training stability analysis of AdaTIR compared with standard reward shaping. From left to right, the panels display the **Gradient Norm**, **Training Reward**, and **Response Length** over training steps. Solid lines represent the smoothed curves using exponential moving average (EMA), while the semi-transparent shaded areas indicate the raw fluctuations.

5.3 Case Study

Figure 3 compares a representative reasoning process between the TIR baseline and AdaTIR. The baseline exhibits excessive tool dependency: when the symbolic solver fails, it attempts an inappropriate brute-force search and terminates with a hallucinated conclusion. In contrast, AdaTIR successfully derives the correct relationship through internalized algebraic deduction without external tool calls. Additional case studies and categorical error analyses are provided in Appendix C and D.

6 Related Works

RL for Tool-Integrated Reasoning Recent research utilizes Reinforcement Learning (RL) to enable LLMs to autonomously invoke external tools. In mathematical reasoning, ToRL (Li et al., 2025c), ReTool (Feng et al., 2025), and Effective CIR (Bai et al., 2025) train models to utilize Python interpreters for complex tasks, while zero-resource methods like SimpleTIR (Xue et al., 2025) and ZeroTIR (Mai et al., 2025) explore end-to-end reasoning without extensive supervised fine-tuning to minimize data dependence. In general scenarios, agents like Search-R1 (Jin et al., 2025) and WebSailors (Li et al., 2025a) leverage up-to-date knowledge via search engines. While these works show significant potential of Tool Integrated Reasoning, an exclusive focus on correctness often induces redundant tool invocations (Wang et al., 2025a), escalating inference overhead and introducing irrelevant noise.

Efficiency for Tool-Integrated Reasoning Research explicitly targeting efficiency in tool-integrated reasoning remains relatively limited.

Some prompt-guided frameworks either calculate confidence scores (Wang et al., 2025b; Li et al., 2025b; Shen et al., 2024; Qian et al., 2025) or directly prompt models to perceive tool costs (Liu et al., 2025), yet they often rely on expert templates and exhibit limited scalability. Recent efforts have shifted toward training-based methods (Wang et al., 2025a; Hashemi et al., 2025). Specifically, OTC (Wang et al., 2025a) employs reward shaping to improve tool productivity. However, its reliance on local group-minima for approximating optimal tool budgets introduces estimation bias, leading to sub-optimal convergence and persistent redundancy. Moreover, its indiscriminate penalty across varying task complexities often causes performance degradation on difficult questions. We thus introduce a difficulty-aware mechanism to adaptively balance efficiency and accuracy, ensuring improved productivity without compromising reasoning quality.

7 Conclusion

We propose AdaTIR, which establishes a difficulty-aware policy to adaptively regulate tool use based on task complexity. By integrating a difficulty-aware efficiency reward with Clipped Advantage Shaping (CAS), our approach successfully mitigates “cognitive offloading” while ensuring stable policy convergence. Experiments on mathematical benchmarks demonstrate that AdaTIR maintains performance comparable to standard baselines while reducing tool calls by up to 97.6%. This internalized strategy offers an efficient paradigm to elicit the latent reasoning capabilities of LLMs while minimizing unnecessary external reliance.

516 Limitations

517 While AdaTIR shows promise in facilitating reason-
518 ing internalization, this work has several practical
519 limitations. First, our evaluation is primarily fo-
520 cused on mathematical reasoning tasks. Although
521 we have included an experiment based on Search-
522 R1 in Appendix B to explore the method’s effec-
523 tiveness in search-based QA tasks, a more compre-
524 hensive validation across broader domains—such
525 as open-ended coding—is still lacking. The per-
526 formance of AdaTIR may vary in fields where cor-
527 rectness signals and domain structures differ from
528 structured mathematics. Second, due to limited
529 computational resources, we mainly conducted ex-
530 periments on 3B and 7B models and were unable
531 to perform an exhaustive hyperparameter search
532 for the scaling factor δ and the parameter β . These
533 values were selected based on limited empirical
534 trials, and their optimal settings across different
535 model scales and task distributions remain to be
536 fully explored. Additionally, our adaptive policy
537 relies on a heuristic difficulty estimator; while prac-
538 tical, a more sophisticated or learned estimator
539 might be necessary to more accurately capture task
540 complexity in more nuanced scenarios. We leave
541 these broader evaluations and methodological re-
542 finements for future work.

543 References

544 Fei Bai, Yingqian Min, Beichen Zhang, Zhipeng
545 Chen, Wayne Xin Zhao, Lei Fang, Zheng Liu,
546 Zhongyuan Wang, and Ji-Rong Wen. 2025. [To-
547 wards effective code-integrated reasoning](#). *Preprint*,
548 arXiv:2505.24480.

549 Bytedance-Seed-Foundation-Code-Team, :, Yao Cheng,
550 Jianfeng Chen, Jie Chen, Li Chen, Liyu Chen, Wen-
551 tao Chen, Zhengyu Chen, Shijie Geng, Aoyan Li,
552 Bo Li, Bowen Li, Linyi Li, Boyi Liu, Jiaheng Liu,
553 Kaibo Liu, Qi Liu, Shukai Liu, and 37 others. 2025.
554 [Fullstack bench: Evaluating llms as full stack coders](#).
555 *Preprint*, arXiv:2412.00535.

556 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian,
557 Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias
558 Plappert, Jerry Tworek, Jacob Hilton, Reiichiro
559 Nakano, Christopher Hesse, and John Schulman.
560 2021. [Training verifiers to solve math word prob-
561 lems](#). *Preprint*, arXiv:2110.14168.

562 Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan,
563 Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan,
564 Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng,
565 Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and

Ning Ding. 2025. [The entropy mechanism of rein-
forcement learning for reasoning language models](#).
Preprint, arXiv:2505.22617.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang,
Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang,
Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhi-
hong Shao, Zhuoshu Li, Ziyi Gao, and 181 others.
2025. [Deepseek-r1: Incentivizing reasoning capa-
bility in llms via reinforcement learning](#). *Preprint*,
arXiv:2501.12948.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang,
Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin
Chi, and Wanjun Zhong. 2025. [Retool: Reinforce-
ment learning for strategic tool use in llms](#). *Preprint*,
arXiv:2504.11536.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen,
Yuju Yang, Minlie Huang, Nan Duan, and Weizhu
Chen. 2024. [Tora: A tool-integrated reasoning agent
for mathematical problem solving](#). In *International
Conference on Representation Learning*, volume
2024, pages 48362–48395.

Helia Hashemi, Victor Rühle, and Saravan Rajmohan.
2025. [Cost-aware retrieval-augmentation reason-
ing models with adaptive retrieval depth](#). *Preprint*,
arXiv:2510.15719.

Carlos E. Jimenez, John Yang, Alexander Wettig,
Shunyu Yao, Kexin Pei, Ofir Press, and Karthik
Narasimhan. 2024. [Swe-bench: Can language mod-
els resolve real-world github issues?](#) *Preprint*,
arXiv:2310.06770.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon,
Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei
Han. 2025. [Search-r1: Training llms to reason and
leverage search engines with reinforcement learning](#).
arXiv preprint arXiv:2503.09516.

Vijay Konda and John Tsitsiklis. 1999. Actor-critic al-
gorithms. *Advances in neural information processing
systems*, 12.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Red-
field, Michael Collins, Ankur Parikh, Chris Alberti,
Danielle Epstein, Illia Polosukhin, Jacob Devlin, Ken-
ton Lee, Kristina Toutanova, Llion Jones, Matthew
Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob
Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natu-
ral questions: A benchmark for question answering
research](#). *Transactions of the Association for Compu-
tational Linguistics*, 7:452–466.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen
Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baix-
uan Li, Zhengwei Tao, Xinyu Wang, Weizhou
Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu,
Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang,
and Jingren Zhou. 2025a. [Websailor: Navigating
super-human reasoning for web agent](#). *Preprint*,
arXiv:2507.02592.

622	Wenjun Li, Dexun Li, Kuicai Dong, Cong Zhang, Hao Zhang, Weiwen Liu, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025b. Adaptive tool use in large language models with meta-cognition trigger . In <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 13346–13370, Vienna, Austria. Association for Computational Linguistics.	Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models . <i>Preprint</i> , arXiv:2402.03300.	677 678 679 680 681 682
630	Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025c. Torl: Scaling tool-integrated rl . <i>Preprint</i> , arXiv:2503.23383.	Yuanhao Shen, Xiaodan Zhu, and Lei Chen. 2024. SMARTCAL: An approach to self-aware tool-use evaluation and calibration . In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track</i> , pages 774–789, Miami, Florida, US. Association for Computational Linguistics.	683 684 685 686 687 688 689
633	Heng Lin and Zhongwen Xu. 2025. Understanding tool-integrated reasoning . <i>Preprint</i> , arXiv:2508.19201.	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 <i>Proceedings of the Twentieth European Conference on Computer Systems</i> , EuroSys '25, page 1279–1297. ACM.	690 691 692 693 694 695
635	Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts . <i>Transactions of the Association for Computational Linguistics</i> , 12:157–173.	Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning . <i>Preprint</i> , arXiv:2303.11366.	696 697 698 699
640	Tengxiao Liu, Zifeng Wang, Jin Miao, I-Hung Hsu, Jun Yan, Jiefeng Chen, Rujun Han, Fangyuan Xu, Yanfei Chen, Ke Jiang, Samira Daruki, Yi Liang, William Yang Wang, Tomas Pfister, and Chen-Yu Lee. 2025. Budget-aware tool-use enables effective agent scaling . <i>Preprint</i> , arXiv:2511.17006.	Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. 2025. Agentic reasoning and tool integration for llms via reinforcement learning . <i>Preprint</i> , arXiv:2505.01441.	700 701 702 703
646	Xinji Mai, Haotian Xu, Xing W, Weinong Wang, Yingying Zhang, and Wenqiang Zhang. 2025. Agentic RL scaling law: Spontaneous code execution for mathematical problem solving . In <i>The Thirty-ninth Annual Conference on Neural Information Processing Systems</i> .	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 . <i>Preprint</i> , arXiv:2504.14870.	704 705 706 707 708
652	OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card . <i>Preprint</i> , arXiv:2412.16720.	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 <i>Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 6510–6525, Albuquerque, New Mexico. Association for Computational Linguistics.	709 710 711 712 713 714 715 716 717 718
659	Bhrij Patel, Souradip Chakraborty, Wesley A. Suttle, Mengdi Wang, Amrit Singh Bedi, and Dinesh Manocha. 2024. Aime: Ai system optimization via multiple llm evaluators . <i>Preprint</i> , arXiv:2410.03131.	Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning . <i>Preprint</i> , arXiv:2310.03731.	719 720 721 722 723
663	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 <i>Findings of the Association for Computational Linguistics: ACL 2025</i> , pages 4604–4621, Vienna, Austria. Association for Computational Linguistics.	Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025c. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning . <i>Preprint</i> , arXiv:2506.01939.	724 725 726 727 728 729 730 731
670	Qwen, :, 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, Jiayi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report . <i>Preprint</i> , arXiv:2412.15115.	Nan Xu, Fei Wang, Ben Zhou, Bangzheng Li, Chaowei Xiao, and Muhao Chen. 2024. Cognitive overload:	732 733

Jailbreaking large language models with overloaded logical thinking. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3526–3548, Mexico City, Mexico. Association for Computational Linguistics.

Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. 2025. *Simpletir: End-to-end reinforcement learning for multi-turn tool-integrated reasoning*. *Preprint*, arXiv:2509.02479.

A Implementation Details

A.1 Implementation Details for Main Experiments

The experimental settings follow ReTool (Feng et al., 2025). The main experiments are conducted using Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct backbones. Following the data configuration in ReTool, the training data for the cold-start SFT phase and RL phase are sourced from specialized mathematical reasoning datasets.

For the reinforcement learning phase, the difficulty-aware reward and CAS mechanism are configured with the following key parameters: the difficulty threshold ϕ_{low} is set to 0.8 to trigger reasoning internalization for simple tasks; the CAS scaling factor β is maintained at 0.9 to regulate efficiency incentives; and the advantage clipping parameter δ is set to 0.99 to ensure correctness signals remain dominant during optimization. The remaining training and evaluation hyperparameters are summarized in Table 4.

Hyperparameter	Value
Group Size (G)	16
Max Sequence Length	4096
Training batch size	64
PPO mini batch size	16
PPO clip ratio (low / high)	0.2 / 0.28
KL coefficient	0.0
Actor learning rate	1e-6
Tool Call Budget (N_{max})	4
Decoding Temperature (T)	1.0
Top- p Sampling	0.6

Table 4: Summary of experimental hyperparameters for main experiments.

A.2 Implementation Details for Comparison with OTC

To ensure a fair and rigorous comparison with OTC, we follow their experimental setting using the Qwen2.5-Math-1.5B-Instruct backbone and benchmark against both OTC and ToRL. Although OTC reports performance with a maximum budget of

$N_{max} = 3$, the authors explicitly note that allowing multiple tool calls per task can frequently lead to training instability or mode collapse. Consequently, to maintain a stable and strictly controlled evaluation environment, we set the maximum tool-call limit to $N_{max} = 1$ for both our method and the baselines. This setup allows for a more reliable assessment of the model’s capacity to internalize reasoning logic within a constrained budget. All other reward and CAS parameters remain consistent with the main experiments.

A.3 Prompts and Templates

To ensure a fair comparison and reproducibility, we strictly follow the prompt templates and system instructions provided by ReTool (Feng et al., 2025). The full instruction template used consistently during both SFT and RL phases is presented in Table 9.

Feedback when Exceeding Budget: When the number of tool calls exceeds the specified budget K_{max} during reasoning, any further tool call attempts will receive the following feedback: **"Tool call budget exceeded; the code was not executed. Please continue reasoning and give the final answer."** This encourages the model to transition from external tool invocation to internal reasoning during the RL phase.

B Generalization Results on Search R1

To verify the generalization of our method in the Question Answering (QA) domain, we evaluate its performance by applying AdaTIR to search-based workflows. We choose **Search-R1** (Jin et al., 2025) as our baseline and follow its original experimental settings. Specifically, we apply GRPO on the Qwen2.5-3B model, and all hyperparameters introduced by AdaTIR are kept identical to those described in our main implementation (Section A). This ensures a fair comparison and tests whether the benefits of reasoning internalization can be replicated in a search-heavy environment.

Our experimental results, as summarized in Table 5, demonstrate a clear trade-off between effectiveness and efficiency. Across all datasets, AdaTIR consistently improves the EM accuracy (average +1.3%) while achieving a relatively moderate reduction in ATC (from 1.00 to 0.95). This performance gain can be attributed to our difficulty-aware policy, which suppresses redundant search tasks for simple questions, thereby reducing envi-

Table 5: Comparison on Search-R1 generalization tasks using **Qwen2.5-3b-Base** as the backbone. Each cell reports **EM / ATC**, where EM is the Exact Match accuracy and ATC is the Average Tool Calls. **Bold** indicates the best EM, and **green** indicates reduced tool calls.

Method	NQ	TriviaQA	PopQA	HotpotQA	2wiki	Musique	Bamboogle	Avg.
Search-R1	0.391 / 1.00	0.582 / 1.00	0.392 / 1.00	0.285 / 1.00	0.267 / 1.01	0.054 / 1.00	0.121 / 1.00	0.300 / 1.00
AdaTIR	0.403 / 0.94	0.597 / 0.92	0.416 / 0.93	0.297 / 0.94	0.260 / 0.98	0.076 / 0.99	0.145 / 0.96	0.313 / 0.95

ronmental noise and enhancing reasoning consistency .

We hypothesize that the limited efficiency gain compared to mathematical tasks stems from two factors. First, following the observation in the Search-R1 study, models tend to collapse after approximately 200 training steps (Jin et al., 2025). Consequently, both Search-R1 and AdaTIR utilized the 200-step checkpoint for evaluation. At this early stage of training, the model has not fully converged, and the correctness advantage remains more dominant than the efficiency penalty . Second, the inherent demand for external knowledge in QA tasks is significantly higher than the need for code execution in mathematical reasoning. This high necessity for tool involvement naturally limits the potential for further budget compression in search-based scenarios.

C Extended Case Studies

As illustrated in Figure 6, we observe the phenomenon of **cognitive offloading**, where baseline models delegate trivial arithmetic to external tools despite having resolved the logic internally. This behavior stems from a lack of difficulty awareness, where the model prioritizes procedural tool-use over execution efficiency, leading to redundant overhead. Our model mitigates this by utilizing internalized reasoning to identify straightforward tasks and bypass unnecessary external calls.

Furthermore, Figure 7 shows how incorrect or unexpected tool feedback can act as noise and distract the model. When a tool gives a wrong result, the baseline model often stops focusing on the main problem and instead shifts its attention to debugging the code or the output. This shift causes a "logical interruption," where the model gets stuck trying to fix the tool execution rather than solving the original question. This kind of noise prevents the model from finishing its main task and leads to a cycle of errors. By getting the logic right before calling any tools, our approach ensures the model stays focused on the goal instead of being derailed

by tool-related issues.

D Error Analysis

To understand the strategic differences between the baseline and our model, we categorize failures into three distinct modes. This analysis focuses on AIME (hard) and GSM8K (easy) to contrast performance across different complexity levels. We exclude cases where both models fail due to fundamental capacity limits.

- **Missing Tool-use:** The model relies on internal reasoning but fails because the calculation or logic exceeds its current parametric precision. This is often a sign of "over-confidence" in its internalized logic. Table 8 illustrates a typical case where the model attempts to manually simulate a recursive game instead of invoking a code interpreter, resulting in a state-tracking hallucination.
- **Tool-induced Noise:** Incorrect or unexpected tool feedback distracts the model, causing a logical interruption where it shifts focus to debugging code instead of the main problem (see Figure 7 for a detailed case).
- **Reasoning Failure:** Pure logical errors that occur independently of whether a tool was used or how it responded. A representative example is shown in Table 7, where the model fails to reconcile the basic geometric parity of the grid despite successful tool execution.

Category	GSM8K (Easy)		AIME (Hard)	
	Base.	Ours	Base.	Ours
Missing Tool-use	32.8	91.6	8.1	26.5
Tool-induced Noise	9.2	1.1	57.7	38.2
Reasoning Failure	58.0	7.4	34.2	35.3

Table 6: Comparison of failure mode distributions (%) between the baseline and our model.

As shown in Table 6, there is a clear shift in how the models fail. On **GSM8K**, the majority

of our model’s errors (91.6%) fall under *Missing Tool-use*. This indicates that the model has successfully internalized most of the reasoning, though it occasionally overestimates its ability to handle specific calculations internally. Notably, it almost entirely avoids *Tool-induced Noise* (1.1%), which is the main distraction for the baseline.

On the more difficult **AIME** tasks, the baseline is heavily impacted by *Tool-induced Noise* (57.7%). This confirms our observation that frequent, redundant tool interactions often create a "failure spiral" where the baseline loses focus on the original question while trying to manage tool outputs. Our model reduces this noise to 38.2% by relying more on autonomous deduction. While our model’s *Missing Tool-use* rate increases (26.5%), it represents a cleaner failure mode: the model fails because it reaches its internal cognitive limits, rather than being derailed by external interference.

E Additional Analysis of the Method

E.1 Choice of the Sine-based Reward Function

We choose a sine-based reward instead of a simple linear one because its non-linear shape provides a more effective training signal for reasoning internalization. Specifically, the sine curve is steeper at the beginning (when the number of tool calls K is small), which places a higher penalty on the first few tool calls. This encourages the model to solve simple tasks through its own internal reasoning rather than calling tools by default. Additionally, the mathematical smoothness of the sine function ensures continuous gradients, which helps keep the reinforcement learning process stable compared to linear or step-based rewards.

E.2 Understanding the CAS Mechanism

The goal of Clipped Advantage Shaping (CAS) is to make sure the model always prioritizes being correct over being efficient. The total score (advantage) the model receives is:

$$A_i^{\text{CAS}} = A_i^{\text{acc}} + \text{clip}(A_i^{\text{eff}}, -\delta|A_i^{\text{acc}}| - \eta, \delta|A_i^{\text{acc}}| + \eta) \quad (12)$$

where $\delta < 1$ and η is a small constant. We can see how this works in two cases:

Case 1: When trajectories have different accuracy results ($A_i^{\text{acc}} \neq 0$). When some answers in a group are better than others, the accuracy signal should lead the training. Because we set $\delta < 1$, the

efficiency penalty is always smaller than the accuracy reward. For a correct trajectory ($A_i^{\text{acc}} > 0$):

$$\begin{aligned} A_i^{\text{CAS}} &\geq A_i^{\text{acc}} - (\delta|A_i^{\text{acc}}| + \eta) \\ &= (1 - \delta)A_i^{\text{acc}} - \eta \end{aligned} \quad (13)$$

Since $\delta < 1$ and η is very small, A_i^{CAS} stays positive. This ensures that correct answers are always rewarded, and the model won’t get confused by the cost of using tools while it is still learning to be correct.

Case 2: When all trajectories have the same accuracy ($A_i^{\text{acc}} = 0$). If every trajectory in a group gets the same result (for example, all are correct), the accuracy signal becomes zero and provides no direction. In this case, the optimization direction is decided entirely by the efficiency advantage:

$$A_i^{\text{CAS}} = 0 + \text{clip}(A_i^{\text{eff}}, -\eta, \eta) \quad (14)$$

Now, the efficiency score alone tells the model which path is better. It encourages the model to avoid wasting steps and find shorter reasoning paths. The small margin η keeps these updates stable so the model doesn’t change its behavior too drastically.

In summary, CAS acts as a safety mechanism: it stops efficiency costs from messing up the learning process when accuracy is the main goal (Case 1), but allows the efficiency signal to guide the model toward better reasoning paths once accuracy is mastered (Case 2).

The core objective of Clipped Advantage Shaping (CAS) is to prioritize correctness over efficiency. The formulation in Eq. (11) is:

$$\begin{aligned} A_i^{\text{CAS}} &= A_i^{\text{acc}} + r_{\text{acc},i} \cdot \beta \cdot \\ &\quad \text{clip}(A_i^{\text{eff}}, -\delta|A_i^{\text{acc}}| - \eta, \delta|A_i^{\text{acc}}| + \eta) \end{aligned} \quad (15)$$

where $r_{\text{acc},i} \in \{0, 1\}$ is the correctness mask, $\beta > 0$ is the efficiency weight, and $\delta \in [0, 1], \eta > 0$ are clipping parameters.

Scenario 1: Incorrect Trajectories ($r_{\text{acc},i} = 0$). For incorrect responses, the mask zeros out the efficiency term: $A_i^{\text{CAS}} = A_i^{\text{acc}}$. Since A_i^{acc} is typically negative for incorrect answers, CAS preserves the negative feedback strictly.

Scenario 2: Correct Trajectories ($r_{\text{acc},i} = 1$). Here, the sign preservation depends on the group distribution:

- **Case A: Mixed Group** ($A_i^{\text{acc}} > 0$). When the group contains both correct and incorrect responses, correct trajectories typically yield positive accuracy advantages. The worst-case advantage occurs when the efficiency penalty is maximal:

$$A_i^{\text{CAS}} \geq A_i^{\text{acc}} + \beta \cdot (-\delta|A_i^{\text{acc}}| - \eta) \quad (16)$$

$$= (1 - \beta\delta)A_i^{\text{acc}} - \beta\eta \quad (17)$$

Numerical Guarantee: To ensure $A_i^{\text{CAS}} > 0$, we require $A_i^{\text{acc}} > \frac{\beta\eta}{1-\beta\delta}$. With our default settings ($\beta = 0.9, \delta = 0.99, \eta = 1\text{e-}4$), the denominator $1 - \beta\delta \approx 0.109$, and the required threshold is merely $A_i^{\text{acc}} \approx 8.2\text{e-}4$. Since normalized advantages for correct answers in a mixed group are typically significant (e.g., > 0.1), sign reversal is effectively prevented in practice.

- **Case B: All-Correct Group** ($A_i^{\text{acc}} = 0$). If all trajectories in a group are correct, $A_i^{\text{acc}} = 0$. In this convergence state, the clipping range shrinks to $[-\eta, \eta]$:

$$A_i^{\text{CAS}} = 0 + \beta \cdot \text{clip}(A_i^{\text{eff}}, -\eta, \eta) \geq -\beta\eta \quad (18)$$

Here, the advantage is strictly bounded by $-\beta\eta$ (approx. $-9\text{e-}5$). This negligible magnitude ensures that while efficiency dictates the preference ranking among correct answers, it does not introduce destabilizing large negative gradients.

Case Study: Cognitive Offloading in simple Tasks

Question: A candle melts by 2 centimeters every hour that it burns. How many centimeters shorter will a candle be after burning from 1:00 PM to 5:00 PM?

Baseline: Inefficiency via Cognitive Offloading

Thought Trace: Alright, let's tackle this problem step by step. So, the problem states that a candle melts by 2 centimeters every hour it burns. We need to find out how many centimeters shorter the candle will be after burning from 1:00 PM to 5:00 PM. [...] So, if per hour it melts 2 cm, then in 4 hours, it would melt $2 \text{ cm/hour} * 4 \text{ hours} = 8 \text{ cm}$. Wait, let me verify that. Let's use code to calculate the total melting.

```
# Redundant tool call for simple multiplication: 4 * 2
hours = 5 - 1 # From 1:00 PM to 5:00 PM inclusive
melting_per_hour = 2 # cm
total_melting = hours * melting_per_hour
print(total_melting)
```

Response: 8

The code execution result confirms that the total melting is 8 cm. So, the candle will be 8 centimeters shorter.

Tool Calls: 1

Final Answer: ✓

Ours: Efficiency via Internal Reasoning

Thought Trace: Okay, let me try to solve this problem. Hmm, so a candle melts by 2 centimeters every hour that it burns. [...] The question is asking: How many centimeters shorter will a candle be after burning from 1:00 PM to 5:00 PM? Hmm, so that's a 4-hour period, right? From 1:00 PM to 5:00 PM is 4 hours. So if the candle melts at a rate of 2 centimeters per hour, then over 4 hours, it would melt $2 \text{ cm/hour} * 4 \text{ hours} = 8 \text{ centimeters}$. So the candle would be 8 centimeters shorter after 4 hours. Hmm, that seems straightforward. I don't think I missed anything here.

Tool Calls: 0

Final Answer: ✓

Figure 6: Comparison of reasoning efficiency under redundant tool use. Results show our model maintains robust inference via internal logic, effectively avoiding the unnecessary cognitive offloading and efficiency loss seen in the baseline.

Case Study: Reasoning Impairment under Tool-Induced Noise

Question: Find the number of ways to place a digit in each cell of a 2x3 grid so that the sum of the two numbers formed by reading left to right is 999, and the sum of the three numbers formed by reading top to bottom is 99. The grid below is an example of such an arrangement because $8 + 991 = 999$ and $9 + 9 + 81 = 99$.

Baseline: Logic Erosion via Repeated Tool-Execution Noise

Thought Trace: Alright, let's tackle this problem step by step. [...] Maybe the three numbers formed by reading top to bottom are the first column sum, second column sum, and third column sum. Then, their sum is 99. [...] I can use code to iterate through all possible values and count the valid grids.

```
# First Attempt: Brute-forcing with column sum logic (a+d+b+e+c+f == 99)
count = 0
for a in range(1, 10):
    for b in range(0, 10):
        for c in range(0, 10):
            for d in range(1, 10):
                for e in range(0, 10):
                    for f in range(0, 10):
                        if (100*a+10*b+c) + (100*d+10*e+f) == 999:
                            if (a+d) + (b+e) + (c+f) == 99: count += 1
print(count)
```

Response: 0

Hmm, the code execution result is 0. That suggests there might be an issue with the logic. [...] Let me adjust the code to print out the valid grids so I can verify.

```
valid_grids = []
for a, b, c, d, e, f in itertools.product(range(1,10), range(10), ...):
    if sum_lr == 999 and col_sum == 99: valid_grids.append((a,b,c,d,e,f))
print(len(valid_grids))
```

Response: 0

The code execution result is still 0. [...] Maybe the three numbers formed by reading top to bottom are the first column sum, second column sum, and third column sum. [...] Wait, maybe I should write code to generate all possible grids (digits 0-9 for each cell):

```
# Third Attempt: Repeating the same flawed logic with minor cell range adjustments
count = 0
for a in range(1, 10):
    # ... (similar nested loops)
    if num1 + num2 == 999 and col_sum == 99: count += 1
print(count)
```

Response: 0

The code execution result is 0. This means there are no valid grids that satisfy both conditions. Therefore, the answer is 0.

Tool Calls: 3

Final Answer: ✘

Ours: Precise Formalization for Reliable Reasoning

Thought Trace: [...] Reading top to bottom, the three numbers would be $(10a + d)$, $(10b + e)$, and $(10c + f)$. Their sum should be 99. [...] This seems manageable. Let me use code to iterate through all possible digit assignments (0-9) for the six variables and count the number of valid solutions.

```
import itertools
valid_count = 0
for a, b, c, d, e, f in itertools.product(range(10), repeat=6):
    sum_lr = (100 * a + 10 * b + c) + (100 * d + 10 * e + f)
    sum_tb = (10 * a + d) + (10 * b + e) + (10 * c + f)
    if sum_lr == 999 and sum_tb == 99: valid_count += 1
print(valid_count)
```

Response: 45

The code returns 45 as the number of valid grids. Therefore, the number of ways is 45.

Tool Calls: 1

Final Answer: ✔

Figure 7: Comparison of reasoning under tool-induced noise. Our model maintains reliability via accurate formalization, mitigating the noise-driven reasoning impairment seen in the baseline's failure spiral.

An example from the AIME dataset

Question: Consider the paths of length 16 that follow the lines from the lower left corner to the upper right corner on an 8×8 grid. Find the number of such paths that change direction exactly four times.

Alright, so I need to find the number of paths of length 16 on an 8×8 grid... $(0,0)$ to $(7,7)$. Each path is of length 16. If we let the number of R steps be r and U steps be u , then $r + u = 16$ and $r - u = 7$. Solving these: $r = 11.5, u = 4.5$. Wait, that can't be. Steps have to be integers. So, maybe the problem is not about displacement? Wait, to get from $(0,0)$ to $(7,7)$, we need exactly 7 R steps and 7 U steps. That's 14 steps. But the problem says length 16. Maybe I should use code to generate all possible sequences of 16 steps and check direction changes.

```
# Tool Call 1: Enumerating all 16-step sequences
from itertools import product
def count_direction_changes(path):
    directions = []
    # (Simplified logic that incorrectly handles direction vector list)
    # ... logic omitted for brevity ...
    return len(directions)

valid_paths = 0
for path in product(['R', 'U'], repeat=16):
    if count_direction_changes(path) == 4:
        valid_paths += 1
print(valid_paths)
```

» Output: 0

Hmm, the result is 0, which suggests the logic might be incorrect. Let me revise. I should track switches by checking consecutive steps. If the current step is different from the previous one, it's a direction change.

```
# Tool Call 2: Revised counting logic
from itertools import product
def count_direction_changes(path):
    changes = 0
    prev = None
    for step in path:
        if prev is None:
            prev = step
            continue
        if step != prev:
            changes += 1
            prev = step
    return changes

valid_paths = 0
for path in product(['R', 'U'], repeat=16):
    if count_direction_changes(path) == 4:
        valid_paths += 1
print(valid_paths)
```

» Output: 2730

The code execution result is 2730, which matches the known answer. Therefore, the number of paths is 2730.

Table 7: An example of a **Reasoning Failure** on a problem from AIME.

An example from the AIME dataset.

Question: Alice and Bob play the following game. A stack of n tokens lies before them. The players take turns with Alice going first. On each turn, the player removes either 1 token or 4 tokens from the stack. Whoever removes the last token wins. Find the number of positive integers n less than or equal to 2024 for which there exists a strategy for Bob that guarantees that Bob will win the game regardless of Alice’s play.

Model Output:

Alice and Bob are playing a game... removes the last token wins. This is a classic two-player game theory problem. Let $f(n)$ be True if the position is a winning position for the player about to move. $f(0) = \text{False}$ because the game is already over.

For $n = 1$, take 1 token to leave 0, so $f(1) = \text{True}$. For $n = 2$, can only take 1 to leave 1, which is a winning position for the opponent, so $f(2) = \text{False}$. For $n = 3$, the player can take 1 token, leaving 2. The opponent can then take 1, leaving 1, which is a winning position for the opponent. Therefore, $f(3) = \text{False}$. (Note: This is an error, $f(3)$ can reach $f(2) = L$ and should be W). For $n = 4$, take 4 tokens to leave 0, so $f(4) = \text{True}$.

Continuing this pattern: $n = 5$ (lose), $n = 6$ (win), $n = 7$ (win), $n = 8$ (lose), $n = 9$ (lose), $n = 10$ (win). The pattern seems to be that positions where $n \bmod 5 = 0$ are losing positions. Let me verify... yes, all multiples of 5 are losing positions. Therefore, Bob wins if the starting n is a multiple of 5.

Total numbers from 1 to 2024 is 2024. The number of multiples of 5 is $\lfloor 2024/5 \rfloor = 404$. Therefore, the number of non-multiples of 5 is $2024 - 404 = 1620$.

Table 8: An example of a **Missing Tool-use** error on a problem from AIME

Template Prompt for RL Training

Solve the following problem step by step. You now have the ability to selectively write executable Python code to enhance your reasoning process. The Python code will be executed by an external sandbox, and the output (wrapped in `<interpreter>output_str</interpreter>`) can be returned to aid your reasoning and help you arrive at the final answer. The Python code should be complete scripts, including necessary imports.

Each code snippet is wrapped with `<code>\n```\npython\ncode snippet\n```\n</code>`.

The last part of your response should be in the following format:

```
<answer>
\boxed{'The final answer goes here.'}
</answer>
```

```
*user question:*
{question}
```

Remember to place the final answer in the last part using the format:

```
<answer>
\boxed{'The final answer goes here.'}
</answer>
```

Table 9: The full prompt template following the ReTool protocol (Feng et al., 2025). This template is used consistently during both the SFT and RL phases to maintain reasoning consistency.