TRAC: Teacher-Guided Token Reward with Adaptive Calibration for Robust Policy Optimization

Sitong Wu[1] **Haoru Tan**[2,✉] **Xichen Zhang**[3] **Bin Xia**[1] **Wenhu Zhang**[3]
Xiaojuan Qi[2] **Bei Yu**[1] **Jiaya Jia**[1,3,✉]

[1]The Chinese University of Hong Kong [2]The University of Hong Kong
[3]The Hong Kong University of Science and Technology

Abstract

Reinforcement Learning (RL) with sparse outcome rewards suffers from inefficient credit assignment in complex LLM reasoning tasks. While utilizing stronger LLMs as teachers to derive dense token-level supervision offers a cost-effective alternative to proprietary reward models, it relies on the flawed assumption that teachers are perfect oracles. In reality, teacher models exhibit capability limitations and uncertainty, producing noisy signals that make student policies susceptible to reward hacking. To address this, we propose Teacher Reward Adaptive Calibration (**TRAC**), a robust framework that filters noisy supervision by dynamically modulating teacher influence via a multi-granularity calibration mechanism. TRAC evaluates teacher reliability across three principled dimensions: problem-level expertise, trajectory-level discrimination, and token-level confidence. Furthermore, we integrate TRAC with Group Relative Policy Optimization (GRPO), formulating as **TRAC-GRPO**, which treats calibrated teacher-derived reward as an additive advantage reshaping term to ensure fair advantage estimation. Extensive experiments demonstrate that TRAC effectively mitigates teacher noise, significantly enhancing the reasoning capabilities and training stability of LLMs compared to standard baselines. The code will be available at: https://github.com/JIA-Lab-research/TRAC.

1 Introduction

The efficacy of Reinforcement Learning (RL) for enhancing Large Language Model (LLM) reasoning is increasingly bottlenecked by the inherent limitations of reward design. Current paradigms, such as GRPO (Guo et al., 2025) and DAPO (Yu et al., 2025), rely primarily on outcome rewards that uniformly propagate a single scalar value across all tokens based on final correctness. This coarse-grained supervision suffers from severe credit assignment noise. Consequently, token-level dense rewarding is indispensable for facilitating precise gradient estimation and ensuring a stable learning process.

In stark contrast, high-performing LLMs with strong reasoning capabilities are increasingly accessible to the community through both open-weight releases (*e.g.,* Qwen (Yang et al., 2025), LLaMA (Dubey et al., 2024)) or commercial APIs (*e.g.,* GPT (OpenAI, 2024a), Gemini (DeepMind, 2025), Claude (Anthropic, 2025)). Most dense reward methods depend on specialized reward models trained on massive annotated datasets. However, these models are rarely open-sourced or require prohibitively expensive annotation to replicate. To address this, recent studies have leveraged high-performance LLMs, such as Qwen (Yang et al., 2025) and GPT (OpenAI, 2024a), to construct fine-grained rewards. For instance, recent efforts (Lu et al., 2025) utilize the per-token log-probability difference between a stronger teacher model and the policy model as a granular reward signal. However, using teacher probabilities as rewards assumes the teacher is a perfect oracle, a premise that fails in complex reasoning where even advanced models suffer from hallucinations and uncertainty. Our empirical observations confirm this; even capable models like Qwen3-32B occasionally assign higher rewards to incorrect trajectories than to correct ones. This noisy supervision is particularly hazardous in RL, as models aggressively exploit reward signals, leading to reward hacking and overfitting to high-confidence errors. Thus, simply adopting a stronger teacher is insufficient; a robust mechanism to explicitly calibrate and filter these signals is indispensable for stable and effective policy optimization.

To address this challenge, we propose **T**eacher **R**eward **A**daptive **C**alibration (**TRAC**), a framework that robustly incorporates token-level teacher reward by dynamically modulating their influence via a multi-granularity calibration mecha-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 47869–47884
July 2-7, 2026

nism. Specifically, we formulate the weight of the token-level teacher's reward as the product of three calibration coefficients, designed to filter noise based on three principled dimensions of reliability: **(1) Problem-level Expertise** (λ^{expert}), which assesses the teacher's proficiency regarding the specific problem. **(2) Trajectory-level Discrimination** (λ^{disc}), which quantifies the teacher's ability to distinguish correct versus incorrect reasoning trajectories generated by the policy model (*i.e.,* higher average rewards to correct trajectories than incorrect ones). **(3) Token-level Confidence** (λ^{conf}), which reflects the teacher's certainty in assessing the quality of each token within the policy's trajectory. By integrating these factors multiplicatively, our method effectively suppresses the teacher's guidance in unreliable regions where the teacher is incompetent, inconsistent, or uncertain, while preserving valuable supervision with high confidence and reliability.

To evaluate the effectiveness of our TRAC framework, we incorporate it with GRPO, resulting in TRAC-GRPO. Extensive experiments demonstrate that TRAC-GRPO consistently outperforms GRPO, DAPO, and other dense rewarding methods. For example, on AIME24, TRAC-GRPO surpasses GRPO by +7.8 on DeepSeek-R1-Distill-Qwen-1.5B, and improves OpenMath-Nemotron-1.5B by +5.5. Moreover, TRAC-GRPO introduces only slight additional cost on GRPO (30s per step) and is nearly 2× faster than DAPO (210s *vs.* 371s).

2 Observation and Motivation

The core premise of utilizing a stronger LLM as a teacher is that it provides superior supervision compared to the student policy. However, in complex reasoning tasks, even state-of-the-art models are not infallible. To investigate the reliability of teacher-derived supervision, we conducted a preliminary empirical analysis using Qwen3-8B/32B as the teacher, and Qwen3-8B-Base as the policy model on the DeepScaleR-40K training dataset. Our observations reveal **three critical sources of noise** in the teacher's supervision signal (as Eq. (1)).

(1) Teacher has a clear capability limitation. We first evaluate the teacher's expertise (Pass@1 accuracy) across the training problems. As shown in Figure 1 (a), the teacher fails to solve a non-negligible portion of the problems. If the teacher itself lacks the capability to solve a specific problem (*i.e.,* falls outside its competence boundary), the supervision signals derived from them on these failure cases are likely flawed or misleading, which may propagate errors and hinder the learning of correct reasoning logic.

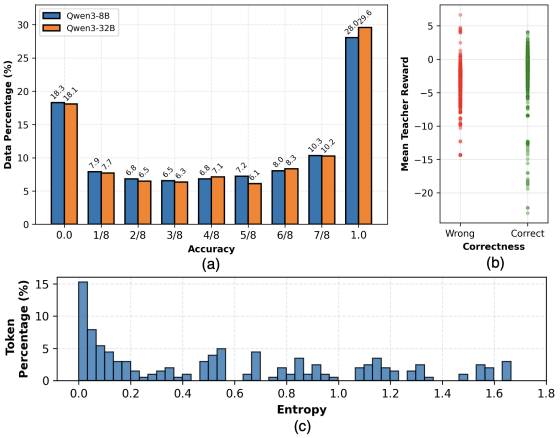

Figure 1: Empirical analysis of teacher reliability and potential noise sources. (a) Accuracy distribution of Qwen3-8B and 32B models on DeepScaleR-40K dataset. (b) Distribution of average teacher rewards assigned to correct versus incorrect student trajectories. (c) Histogram of teacher's token-wise entropy on a correct student trajectory.

(2) Misalignment of Teacher Reward with Correctness. Ideally, a reliable teacher should assign strictly higher rewards to correct reasoning paths than to incorrect ones. To verify this, we sampled trajectories from the student policy and categorized them into Correct and Wrong groups based on the final answer. We then plotted the teacher's average reward (log-probability difference) for each trajectory in Figure 1 (b). Surprisingly, we observe significant anomalies where the teacher assigns high rewards to logically flawed paths (false positives) while paradoxically under-rewarding valid trajectories (false negatives). This phenomenon indicates that the teacher's internal preference does not always align with objective correctness, which likely stems from either the teacher's inherent capability limitations or their failure to recognize alternative valid strategies for complex problems (*i.e.,* limited coverage of the diverse solution space).

(3) Intrinsic Token-wise Uncertainty as a Quality Evaluator. Even within a high-quality trajectory, the teacher's certainty is not uniform. Figure 1 (c) visualizes the token-level entropy of the teacher on a correct student trajectory. We can observe that while the teacher is confident in most tokens (low entropy), it exhibits high uncertainty (high entropy)

at certain tokens. High entropy implies that the teacher considers multiple tokens as plausible continuations or is perplexed by the current context. Treating the teacher's reward on these uncertain positions as gold labels introduces high-variance noise into the gradient estimation.

3 Methodology

In this section, we present our proposed Teacher Reward Adaptive Calibration (**TRAC**), a generalizable framework designed to derive robust, fine-grained token-level reward signals from a stronger teacher model. The specific formulation and calibration mechanisms of TRAC are detailed in Sec. 3.1. As a reward-centric enhancement, TRAC is inherently compatible with various policy optimization algorithms. In Sec. 3.2, we employ the widely adopted Group Relative Policy Optimization (GRPO) to exemplify the integration of TRAC with policy optimization methods, formulating the combined algorithm as **TRAC-GRPO**.

3.1 Teacher Reward Adaptive Calibration

TRAC addresses the inherent noise in teacher-derived rewards through a multi-granular adaptive calibration framework that selectively leverages high-quality guidance while mitigating unreliable signals. We first formalize the calculation of the raw token-level teacher reward signal in Sec. 3.1.1. Subsequently, in Sec. 3.1.2, we introduce our adaptive calibration mechanism, which dynamically modulates the raw teacher rewards based on three complementary dimensions of teacher reliability.

3.1.1 Teacher Reward Calculation

For a trajectory $y_{i,j} = (y_{i,j,1}, \ldots, y_{i,j,L_{i,j}})$ sampled from the policy $\pi_{\theta_{\text{old}}}$ for problem p_i, we derive dense supervisory signals from a stronger, fixed teacher model $\pi_{\theta_{\text{teacher}}}$. Instead of using the raw teacher likelihood, we formulate the token-level reward $R^{\text{teacher}}_{i,j,t}$ as the log-probability difference between the teacher and the policy model:

$$R^{\text{teacher}}_{i,j,t} = \log \pi_{\theta_{\text{teacher}}}(y_{i,j,t} \mid y_{i,j,<t}, p_i) - \log \pi_{\theta_{\text{old}}}(y_{i,j,t} \mid y_{i,j,<t}, p_i), \quad (1)$$

where $y_{i,j,<t}$ denotes the sequence of tokens preceding position t.

Rationalization. This formulation employs the teacher model effectively as a fine-grained quality evaluator rather than a static imitation target. Functioning as a contrastive evaluation mechanism, it measures the informational divergence between expert judgment and the student's current belief:

- Positive Reward ($R^{\text{teacher}}_{i,j,t} > 0$): This condition ($\pi_{\theta_{\text{teacher}}} > \pi_{\theta_{\text{old}}}$) implies the teacher assigns higher validity to a token than the student does. It highlights a critical blind spot, a correct reasoning step, or insight that the student underestimated. The positive reward incentivizes the policy to increase its probability mass on these expert-preferred tokens.
- Negative Reward ($R^{\text{teacher}}_{i,j,t} < 0$): Conversely, this condition ($\pi_{\theta_{\text{teacher}}} < \pi_{\theta_{\text{old}}}$) signals that the student is overconfident in a token that the teacher deems unlikely, such as a hallucination or logical flaw. The negative reward serves as a penalty, discouraging the policy from reinforcing such suboptimal paths.

Unlike binary outcome rewards, these continuous teacher rewards provide token-wise feedback, distinguishing between reasonable intermediate tokens and invalid ones even before the final answer is reached. Crucially, this fine-grained supervision is derived directly from the teacher, eliminating the need for training a separate reward model.

3.1.2 Adaptive Calibration Mechanism

Although teacher-derived rewards offer valuable dense supervision, they are inherently prone to noise stemming from the teacher's capability boundaries, potential hallucinations, and reasoning uncertainty. Our empirical observations (in Sec. 2) indicate that naively utilizing these raw teacher rewards without reliability assessment leads to unstable policy updates and suboptimal convergence. To address this challenge, we introduce an adaptive calibration mechanism that dynamically modulates the raw teacher reward $R^{\text{teacher}}_{i,j,t}$ through a multiplicative coefficient:

$$R^{\text{trac}}_{i,j,t} = \lambda_{i,j,t} \cdot R^{\text{teacher}}_{i,j,t}, \quad (2)$$

where $\lambda_{i,j,t} \in [0, 1]$ serves as a reliability coefficient, which continuously adjusts the magnitude of the raw teacher reward signal $R^{\text{teacher}}_{i,j,t}$ (defined in Eq. (1)), preserving the full strength of high-confidence supervision ($\lambda_{i,j,t} \to 1$) while smoothly attenuating unreliable or ambiguous guidance ($\lambda_{i,j,t} \to 0$).

Our calibration coefficient $\lambda_{i,j,t}$ factorizes across three complementary dimensions, each capturing a

distinct aspect of teacher reliability:

$$\lambda_{i,j,t} = \underbrace{\lambda_i^{\text{expert}}}_{\text{problem-wise}} \cdot \underbrace{\lambda_{i,j}^{\text{disc}}}_{\text{trajectory-wise}} \cdot \underbrace{\lambda_{i,j,t}^{\text{conf}}}_{\text{token-wise}}, \quad (3)$$

where $\lambda_i^{\text{expert}}$ quantifies the teacher's expertise on the specific problem (as Eq. (5)). $\lambda_{i,j}^{\text{disc}}$ measures the teacher's discriminative capability in distinguishing between correct and incorrect trajectories sampled by the policy model (defined in Eq. (7)). $\lambda_{i,j,t}^{\text{conf}}$ reflects the teacher's certainty when evaluating the quality of an individual token (see Eq. (9)). Crucially, this multiplicative design effectively acts as a logical AND gate, ensuring that the supervision signal is attenuated if the teacher fails to demonstrate reliability in any single dimension.

In the following, we detail the motivation, rationale, and specific formulation for each coefficient.

Problem-wise Expertise Coefficient ($\lambda_i^{\text{expert}}$). This coefficient acts as a reliability modulator to attenuate the influence of teacher reward on the problems that are not firmly within the teacher's capability boundary. Supervision derived from problems where the teacher themselves is incompetent is likely to be misleading and should be down-weighted.

Ideally, a teacher's expertise on a problem can be quantified by its accuracy (*e.g.,* Pass@1). However, reliable accuracy estimation typically requires sampling multiple trajectories (*e.g.,* 8 times), due to the inherent stochasticity of the decoding process. It is computationally prohibitive to perform such extensive sampling for every problem in the training set, particularly for deep reasoning models with long chain-of-thought trajectories.

To circumvent this challenge, we propose an efficient proxy that requires only a single forward pass: leveraging the teacher's entropy on the ground-truth solution as a surrogate for its accuracy. Specifically, for problem p_i, we feed its ground-truth solution y_i^* into the teacher model and compute the average entropy E_i^* of its prediction distribution at all tokens, formulated as follows:

$$E_i^* = -\frac{1}{|y_i^*|}\sum_{t=1}^{|y_i^*|}\sum_{v\in\mathcal{V}} p_{i,t}(v)\log p_{i,t}(v), \quad (4)$$
$$p_{i,t}(v) = \pi_{\theta_{\text{teacher}}}(v \mid y_{i,<t}^*, p_i),$$

where $\mathcal{V}$ denotes the complete vocabulary of the teacher model, and the inner summation iterates over every possible token $v \in \mathcal{V}$ to compute the entropy of the predictive distribution at position t.

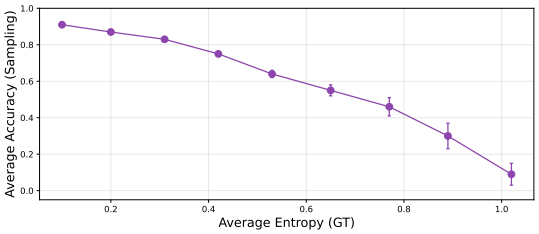

Figure 2: Relationship between accuracy and entropy.

Intuitively, E_i^* quantifies the teacher's endorsement of the ground-truth solution. Low entropy implies the teacher confidently predicts the ground truth, indicating high expertise, while high entropy suggests the teacher is perplexed by the correct logic. As shown in Figure 2, a clear negative correlation can be observed between E_i^* and the actual accuracy, validating E_i^* as a reliable proxy for model accuracy. Since the entropy E_i^* is a continuous scalar with a varying range, we map it to the normalized reliability weight $\lambda_i^{\text{expert}} \in [0, 1]$ using a linear rectification function. We introduce two hyperparameters, $\varepsilon_{\text{min}}^*$ and $\varepsilon_{\text{max}}^*$, to define the boundaries of teacher reliability:

$$\lambda_i^{\text{expert}} = \begin{cases} 1 & \text{if } E_i^* \le \varepsilon_{\text{min}}^*, \\ \frac{\varepsilon_{\text{max}}^* - E_i^*}{\varepsilon_{\text{max}}^* - \varepsilon_{\text{min}}^*} & \text{if } \varepsilon_{\text{min}}^* < E_i^* < \varepsilon_{\text{max}}^*, \\ 0 & \text{if } E_i^* \ge \varepsilon_{\text{max}}^*. \end{cases} \quad (5)$$

where the entropy E_i^* is defined in Eq. (4). Under this mapping, problems where the teacher is highly confident (entropy below $\varepsilon_{\text{min}}^*$) retain full influence, while those inducing high perplexity (entropy above $\varepsilon_{\text{max}}^*$) are zeroed out. The linear transition between these thresholds ensures a smooth decay of trust as uncertainty increases.

It is worth noting that if computational resources permit, $\lambda_i^{\text{expert}}$ can alternatively be derived from the teacher's actual sampling accuracy. However, our ablation study demonstrates that the proposed entropy-based proxy retains 97% of the performance achieved by the expensive accuracy-based method, while reducing the computational cost.

Trajectory-wise Discrimination Coefficient ($\lambda_{i,j}^{\text{disc}}$). Although the expertise coefficient $\lambda_i^{\text{expert}}$ addresses the problems where the teacher is incompetent, it does not guarantee that the teacher can successfully distinguish between the specific correct and incorrect responses generated by the policy model. To address this, we introduce a trajectory-wise coefficient that validates the relative ranking consistency of the teacher's feedback, *i.e.,*, verifying whether the teacher assigns a higher

average reward to correct trajectories compared to incorrect ones.

Specifically, for a problem p_i, we first partition the sampled trajectory set $\mathcal{Y}_i$ into a set of correct trajectories $\mathcal{Y}_i^+$ and a set of incorrect trajectories $\mathcal{Y}_i^-$. Then, for each group, we calculate the teacher's average log-probability on all trajectories belonging to this group:

$$P_i^+ = \frac{1}{|\mathcal{Y}_i^+|} \sum_{y \in \mathcal{Y}_i^+} P(y), \quad P_i^- = \frac{1}{|\mathcal{Y}_i^-|} \sum_{y \in \mathcal{Y}_i^-} P(y),$$
$$P(y) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_{\theta_{\text{teacher}}}(y_t \mid y_{<t}, p_i). \tag{6}$$

Intuitively, a reliable teacher should exhibit a higher average log-probability in correct trajectories compared to incorrect ones.

For a specific trajectory $y_{i,j}$, we define $\lambda_{i,j}^{\text{disc}}$ as a binary consistency gate:

$$\lambda_{i,j}^{\text{disc}} = \begin{cases} \mathbb{I}\left(P(y_{i,j}) > P_i^-\right) & \text{if } y_{i,j} \in \mathcal{Y}_i^+, \\ \mathbb{I}\left(P(y_{i,j}) < P_i^+\right) & \text{if } y_{i,j} \in \mathcal{Y}_i^-, \end{cases} \tag{7}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, which equals 1 if the enclosed condition holds and 0 otherwise. This mechanism functions as a hard binary gate based on log-probability scores. For a correct trajectory ($y_{i,j} \in \mathcal{Y}_i^+$), we require its score to exceed the average score of incorrect trajectories (P_i^-); otherwise, the teacher fails to recognize its superiority, and the teacher reward signal is discarded. Conversely, for an incorrect trajectory, its score must be lower than the average score of correct samples (P_i^+). This ensures that we only utilize supervision when the teacher has a reasonable preference for the policy's correct and incorrect trajectories. Note that if either $\mathcal{Y}_i^+$ or $\mathcal{Y}_i^-$ is empty, this discrimination consistency check is skipped, and $\lambda_{i,j}^{\text{disc}}$ defaults to 1.

Token-wise Confidence Coefficient ($\lambda_{i,j,t}^{\text{conf}}$). Even when the teacher demonstrates overall expertise on a problem and maintains discriminative consistency across trajectories, their confidence for trajectory quality evaluation is not static but fluctuates dynamically across different tokens of a single trajectory. This token-level uncertainty necessitates fine-grained calibration to prevent unreliable rewards from propagating through the policy update process. Therefore, we introduce a fine-grained coefficient to modulate supervision at the token level.

Specifically, for the t-th token $y_{i,j,t}$ in trajectory $y_{i,j}$, we directly measure the teacher's quality evaluation uncertainty for it using the entropy at this position t:

$$E_{i,j,t} = -\sum_{v \in \mathcal{V}} p_{i,j,t}(v) \log p_{i,j,t}(v),$$
$$p_{i,j,t}(v) = \pi_{\theta_{\text{teacher}}}(v \mid y_{i,j,<t}, p_i), \tag{8}$$

where $\mathcal{V}$ is the teacher vocabulary set, and $y_{i,j,<t}$ denotes preceding tokens at position t. High $E_{i,j,t}$ indicates significant uncertainty about the appropriate t-th token, making its reward signal less reliable, thus it should be down-weighted. Conversely, a low $E_{i,j,t}$ implies the teacher has a definite preference for the quality of token $y_{i,j,t}$. This entropy-based approach provides a theoretically sound uncertainty measure of the raw teacher rewards $R_{i,j,t}^{\text{teacher}}$.

Similar to the normalization in expertise coefficient, we map this continuous entropy to a weight $\lambda_{i,j,t}^{\text{conf}} \in [0, 1]$ using a linear rectification function with two hyper-parameters, $\varepsilon_{\text{min}}^{\text{conf}}$ and $\varepsilon_{\text{max}}^{\text{conf}}$:

$$\lambda_{i,j,t}^{\text{conf}} = \begin{cases} 1 & \text{if } E_{i,j,t} \le \varepsilon_{\text{min}}^{\text{conf}}, \\ \frac{\varepsilon_{\text{max}}^{\text{conf}} - E_{i,j,t}}{\varepsilon_{\text{max}}^{\text{conf}} - \varepsilon_{\text{min}}^{\text{conf}}} & \text{if } \varepsilon_{\text{min}}^{\text{conf}} < E_{i,j,t} < \varepsilon_{\text{max}}^{\text{conf}}, \\ 0 & \text{if } E_{i,j,t} \ge \varepsilon_{\text{max}}^{\text{conf}}. \end{cases} \tag{9}$$

This mechanism ensures that the student model receives strong guidance only when the teacher is confident in its reward. When the teacher is confused or ambiguous, the coefficient gradually drops to 0, effectively shielding the policy from noisy or uncertain teacher-derived reward.

3.2 Policy Optimization with TRAC

A key advantage of TRAC is its versatility: it operates strictly at the reward design level, making it orthogonal to the underlying advantage estimation or objective function. Consequently, TRAC is seamlessly compatible with various policy optimization algorithms. In this work, we integrate our TRAC with the widely adopted GRPO algorithm (Guo et al., 2025), termed TRAC-GRPO, to evaluate its effectiveness in enhancing the reasoning ability of LLMs.

For the problem p_i sampled from the training dataset $\mathcal{D}$, the policy model $\pi_{\theta_{\text{old}}}$ generates a group of G trajectories $\mathcal{Y}_i = \{y_{i,1}, \ldots, y_{i,G}\}$. Each trajectory $y_{i,j} = (y_{i,j,1}, \ldots, y_{i,j,L_{i,j}})$ consists of a sequence of tokens with length $L_{i,j}$.

Rewarding. We employ two distinct sources of reward signals. First, each trajectory $y_{i,j}$ receives a sparse outcome reward $R_{i,j}^{\text{outcome}}$ based solely on final answer correctness. Second, each token $y_{i,j,t}$ has an instantaneous reward $R_{i,j,t}^{\text{trac}}$ derived from the

teacher. Note that $R^{\text{trac}}_{i,j,t}$ is dynamically modulated by our calibration coefficients and may drop to zero if the teacher is deemed unreliable (as Eq. (1)).

Reshaped Advantage Estimation. To incorporate our TRAC reward, we employ a reshaped advantage estimation strategy. The core insight is to treat the calibrated teacher-derived reward $R^{\text{trac}}_{i,j,t}$ as an additive shaping term on the original group-relative advantage solely based on outcome reward.

Specifically, we first compute the outcome advantage $A^{\text{outcome}}_{i,j}$ by normalizing the sparse outcome rewards within the trajectory group $\mathcal{Y}_i$ for problem p_i. Then, we incorporate the fine-grained teacher-derived reward signal $R^{\text{trac}}_{i,j,t}$ as an additive term, resulting in the final advantage of each token $y_{i,j,t}$. This process can be formulated as follows:

$$A^{\text{outcome}}_{i,j} = \frac{R^{\text{outcome}}_{i,j} - \texttt{mean}(\{R^{\text{outcome}}_{i,j}\}_{j=1}^{G})}{\texttt{std}(\{R^{\text{outcome}}_{i,j}\}_{j=1}^{G})}, \tag{10}$$

$$A^{\text{reshaped}}_{i,j,t} = A^{\text{outcome}}_{i,j} + \omega \cdot R^{\text{trac}}_{i,j,t}, \tag{11}$$

where ω is a hyper-parameter balancing the two terms (default $\omega = 1$). $\texttt{mean}(\cdot)$ and $\texttt{std}(\cdot)$ denote the operation to calculate mean value and standard deviation, respectively. This additive reshape formulation ensures that the relative ranking of trajectories is anchored by the correctness, while the teacher rewards provide dense guidance without skewing the normalization statistics. We discuss and compare other designs for advantage estimation in the ablation study in (Appendix E.4).

Objective Function. Following GRPO (Guo et al., 2025), our TRAC-GRPO optimizes the policy model π_θ by maximizing a clipped surrogate objective that leverages our reshaped advantage estimates $A^{\text{reshaped}}_{i,j,t}$ in Eq. (11). The training objective is formulated as:

$$\mathcal{J}(\theta) = \mathbb{E}_{p_i \sim \mathcal{D}} \Bigg[\frac{1}{G} \sum_{j=1}^{G} \frac{1}{L_{i,j}} \sum_{t=1}^{L_{i,j}} \min \Big(\rho_{i,j,t}(\theta) A^{\text{reshaped}}_{i,j,t}, \texttt{clip}\left(\rho_{i,j,t}(\theta), 1-\epsilon, 1+\epsilon\right) A^{\text{reshaped}}_{i,j,t} \Big) \Bigg], \tag{12}$$

where $\mathcal{D}$ denotes the training dataset and $L_{i,j}$ is the length of trajectory $y_{i,j}$. The importance sampling ratio $\rho_{i,j,t}(\theta) = \frac{\pi_\theta(y_{i,j,t}|y_{i,j,<t},p_i)}{\pi_{\theta_{\text{old}}}(y_{i,j,t}|y_{i,j,<t},p_i)}$ corrects for distributional shift between the old policy $\pi_{\theta_{\text{old}}}$ that generated trajectories and the current policy π_θ being optimized. The $\texttt{clip}(\cdot, 1-\epsilon, 1+\epsilon)$ operation stabilizes training by constraining policy update magnitudes, with clipping threshold ϵ.

4 Experiments

All the experiment settings and implementation details are provided in Appendix D.

4.1 Main Results

We conduct extensive experiments across diverse LLMs to comprehensively evaluate the effectiveness of our TRAC-GRPO. Table 1 and Table 2 report the results on Base LLMs and SFT-finetuned LLMs, respectively. Besides, we apply further RL to the models that have already undergone RL, with results reported in Figure 5 and Figure 4.

Consistent Performance Superiority. Across all experimental settings, TRAC-GRPO consistently outperforms both GRPO (Shao et al., 2024b) and DAPO (Yu et al., 2025). For instance, on Qwen3-8B-Base (Table 1), TRAC-GRPO achieves a substantial improvement of +3.3 and +2.8 average accuracy compared to GRPO and DAPO. On DeepSeek-R1-Distill-Qwen-1.5B, TRAC-GRPO surpasses GPRO and DAPO by +7.4 and +5.4, respectively. These consistent gains validate that our calibrated teacher rewards provide robust and high-quality supervision, universally benefiting models regardless of their size or initialization state.

Accelerating Convergence. Beyond performance comparison, TRAC-GRPO significantly accelerates the training convergence. As illustrated in Figure 3(a), our method reaches the peak training accuracy of GRPO (originally achieved at step ∼275) by merely step ∼80, representing a ***3.4× speedup***. Furthermore, TRAC-GRPO does not merely converge faster but also converges to a strictly higher asymptotic performance level than both baselines. A similar trend is observed in the test accuracy curves shown in Figure 3(c), confirming that the reliable dense reward allows the policy to master reasoning patterns more efficiently.

Broadening Capability Boundaries. Beyond the average accuracy, we analyze a more challenging metric: the "Zero-Success Rate", defined as the proportion of training problems where the model fails to generate a single correct solution among all sampled rollouts. As Figure 3(b), TRAC-GRPO leads to a lower zero-success rate compared to GRPO and DAPO. This indicates that our method effectively broadens the model's capability boundaries, converting more previously unsolvable problems (where the model had zero probability of success) into solvable ones, rather than simply increasing the confidence in already mastered problems.

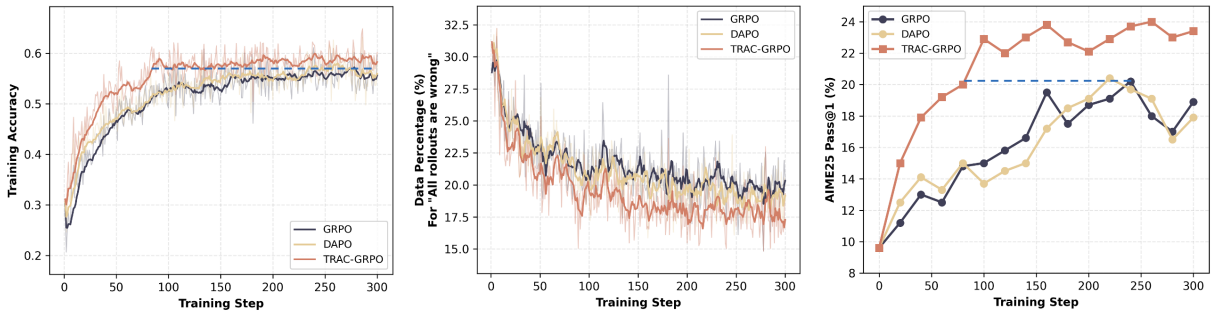

Figure 3: Comparison of different methods on (a) training accuracy, (b) the percentage of training data where all on-policy sampled rollouts are incorrect, and (c) test accuracy on the AIME25 benchmark. These results correspond to the Qwen3-8B-Base experiments in Table 1.

Model	AIME24	AIME25	AMC23	MATH-500	MinervaMath	OlympiadBench	Avg.
Qwen2.5-Math-7B-Base	2.1	2.9	11.9	19.5	7.5	6.1	8.3
+ GRPO	28.1	16.8	70.5	82.0	44.4	43.8	47.6
+ DAPO	30.0	19.6	72.5	82.2	43.8	43.4	48.6
+ TRAC-GRPO (Ours)	**30.8**	**22.4**	**75.5**	**83.6**	**45.6**	**44.7**	**50.4**
Qwen3-1.7B-Base	6.2	2.0	32.8	56.4	25.0	22.5	24.2
+ GRPO	12.8	7.9	42.5	68.4	34.6	32.1	33.1
+ DAPO	15.0	11.7	45.7	68.5	34.3	32.6	34.6
+ TRAC-GRPO (Ours)	**17.2**	**12.8**	**50.2**	**69.8**	**35.7**	**33.8**	**36.6**
Qwen3-4B-Base	8.7	2.1	41.2	64.5	36.1	34.6	31.2
+ GRPO	23.3	20.8	67.8	85.8	48.1	50.5	49.4
+ DAPO	24.1	21.3	71.2	86.0	47.8	50.9	50.2
+ TRAC-GRPO (Ours)	**25.4**	**23.8**	**74.8**	**87.0**	**49.3**	**52.6**	**52.2**
Qwen3-8B-Base	12.9	9.6	52.8	73.2	39.8	39.7	38.0
+ GRPO	25.6	20.2	69.7	85.8	49.5	50.7	50.3
+ DAPO	26.0	20.4	71.0	86.4	50.0	51.2	50.8
+ TRAC-GRPO (Ours)	**29.5**	**24.0**	**75.5**	**88.4**	**50.8**	**53.5**	**53.6**

Table 1: Comparison of different methods under Zero-RL (*i.e.,* directly perform RL on base LLMs after pretraining).

Model	AIME24	AIME25	AMC23	MATH-500	MinervaMath	OlympiadBench	Avg.
DeepSeek-R1-Distill-Qwen-1.5B	28.9	22.8	62.9	83.9	26.5	43.3	44.7
+ GRPO	30.2	23.5	67.0	84.1	27.3	45.1	46.2
+ DAPO	31.6	24.0	72.5	85.3	30.0	46.0	48.2
+ TRAC-GRPO (Ours)	**38.0**	**28.4**	**75.8**	**87.9**	**41.6**	**49.7**	**53.6**
OpenMath-Nemotron-1.5B	61.6	49.5	90.0	92.4	27.3	66.8	64.6
+ GRPO	63.3	52.9	89.1	92.3	26.4	67.2	65.2
+ DAPO	62.5	52.5	90.1	92.9	27.5	67.4	65.5
+ TRAC-GRPO (Ours)	**67.1**	**55.8**	**90.6**	**93.3**	**28.5**	**68.0**	**67.2**

Table 2: Comparison of different RL methods on SFT instruction-tuned LLMs. The experiments on DeepSeek-R1-Distill-Qwen-1.5B and OpenMath-Nemotron-1.5B are trained on DeepScaleR-40K and OpenMathReasoning dataset, respectively.

Stabilizing Training and Preventing Collapse. As Figure 4, under GRPO and DAPO, the model's performance on AIME24 exhibits a continuous downward trend, while AIME25 performance fluctuates briefly before deteriorating, indicating irreversible model degradation. In stark contrast, TRAC-GRPO maintains a stable upward trend throughout the training. This demonstrates that our calibrated fine-grained guidance by TRAC serves as a critical stabilizer, preventing training collapse and ensuring consistent improvement.

Amplifying Data Utilization. We explore whether fine-grained supervision can unlock further value from data that appears saturated under sparse supervision. Specifically, we continuously train DeepScaleR-1.5B-Preview model on DeepScaleR-40K, the exact dataset on which it was originally converged using GRPO. Results in Figure 5 reveal that standard GRPO struggles to yield further improvements, suggesting that sparse outcome signals are insufficient to capture the remaining subtle patterns in the data. Conversely, TRAC-GRPO revitalizes the learning process, achieving substantial gains (+10.8 on Minerva, +11.4 on AMC23, and +3.5 on AIME24). This underscores a critical insight: the limitation lies not only in the data itself, but in the reward signal design. While outcome rewards only capture binary correctness, TRAC's

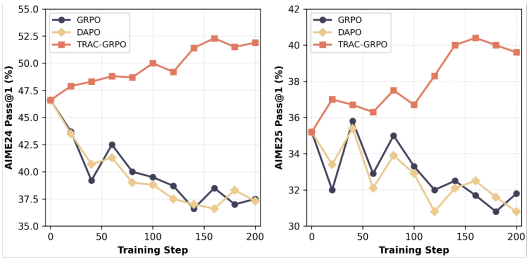

Figure 4: Comparison of training dynamics on AIME24 and AIME25 benchmarks using Qwen3-1.7B trained on DeepScaleR-40K. TRAC-GRPO demonstrates robust performance improvements, whereas GRPO and DAPO exhibit performance regression throughout the training process.

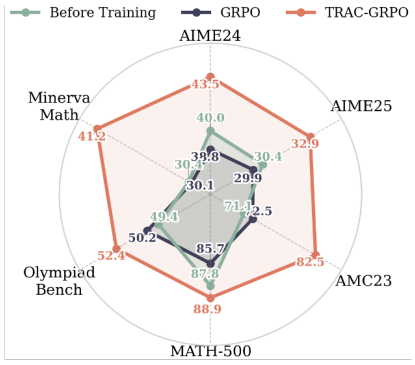

Figure 5: Performance comparison when continuously training DeepScaleR-1.5B-Preview on its original, GRPO-converged source dataset (DeepScaleR-40K).

dense feedback mechanism allows the model to re-evaluate and refine its intermediate reasoning steps, effectively mining high-value supervision from the same set of training trajectories.

Out-of-Domain (OOD) Comparison. We evaluate the models on two representative OOD benchmarks: GPQA (graduate-level science reasoning) and RiddleSense (logical puzzle reasoning). As shown in Table 3, TRAC-GRPO achieves a clear gain of +2.8 on GPQA and +1.2 on RiddleSense compared with GRPO, which demonstrates the superior transferability of TRAC-GRPO. This indicates that the superior in-domain capabilities acquired via TRAC are not a result of overfitting to the training distribution, but rather represent transferable reasoning skills that generalize effectively to other domains.

4.2 Speed Analysis

We evaluate the computational efficiency of our framework. Compared to the vanilla GRPO baseline, TRAC-GRPO incurs only a marginal computational overhead (185 *vs* 210 seconds per step). This increase is minimal because TRAC requires only a single inference pass from the teacher model to compute rewards.

Method	In-Domain Avg.	Out-Domain GPQA	Out-Domain RiddleSense
-	38.0	35.5	66.8
GRPO	50.3	47.2	72.3
DAPO	50.8	48.0	72.7
TRAC-GRPO (Ours)	**53.6**	**50.0**	**73.5**

Table 3: Comparison of different methods on Out-of-Domain (OOD) evaluation. The evaluated models are from the Qwen3-8B-Base experiments in Table 1.

Model	Method	Time per step (seconds)
Qwen3-8B-Base	GRPO	**185**
	DAPO	371
	TRAC-GRPO	210
OpenMath-Nemotron-1.5B	GRPO	**418**
	DAPO	552
	TRAC-GRPO	494

Table 4: Comparison of training speed between different methods.

Notably, TRAC-GRPO demonstrates superior training efficiency compared to DAPO. DAPO relies on computationally expensive rejection sampling to ensure a mix of correct and incorrect trajectories for each problem, aiming to prevent zero-gradient issues when all samples share the same outcome reward. In contrast, our token-level TRAC rewards naturally introduce reward differences across trajectories, even among those with identical final outcomes (all correct or all wrong). This eliminates the need for repeated rejection sampling like DAPO, making TRAC-GRPO strictly more efficient than DAPO in practice.

4.3 Ablation Study

Due to space constraints, we present comprehensive ablation studies in Appendix E, covering the effectiveness of the TRAC reward, the impact of each calibration coefficient, the comparison of advantage estimation strategies, the choice of teacher models, etc.

5 Conclusion

In this paper, we propose TRAC, a robust framework designed to mitigate noisy supervision in teacher-guided RL for LLM reasoning. By implementing a multi-granularity calibration mechanism, spanning problem expertise, trajectory discrimination, and token confidence, TRAC effectively filters flawed teacher signals. Extensive evaluations demonstrate that our method consistently achieves state-of-the-art performance.

6 Acknowledgements

This work was supported in part by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R. The work has been supported by Hong Kong Research Grant Council-General Research Fund Scheme (Grant No. 17202422, 17212923, 17215025), Theme-based Research (Grant No. T45-701/22-R), and Strategic Topics Grant (Grant No. STG3/E-605/25-N). Part of the described research work is conducted in the JC STEM Lab of Robotics for Soft Materials funded by The Hong Kong Jockey Club Charities Trust.

7 Limitations

In this paper, we present a novel framework designed to provide robust and effective fine-grained reward signals. While the efficacy of our approach has been rigorously validated through extensive testing in mathematical reasoning tasks, its applicability to multimodal and video reasoning remains to be explored due to current constraints in time and computational resources. We plan to secure additional resources to conduct further verification and expand the scope of our experiments across a broader range of scenarios in future work.

8 Ethical Considerations

We strictly adhere to the licensing terms of all scientific artifacts used in this study. As our methodology involves neither the generation nor the collection of raw training data, but rather relies exclusively on established benchmarks and publicly available datasets vetted by the academic community, this work does not pose any privacy infringement risks. However, we acknowledge that these datasets are predominantly English-centric, reflecting a systemic bias within the current AI ecosystem. In the future, we commit to and call upon the research community to join us in expanding efforts toward more linguistically diverse datasets and evaluation benchmarks to ensure broader inclusivity.

References

AIME. 2024. American invitational mathematics examination.

AMC. 2023. American mathematics competitions.

Anthropic. 2025. Introducing claude sonnet 4.5.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michał Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, and 1 others. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Google DeepMind. 2025. Gemini 3 pro.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Wu Fei, Hao Kong, Shuxian Liang, Yang Lin, Yibo Yang, Jing Tang, Lei Chen, and Xiansheng Hua. 2025. Self-guided process reward optimization with redefined step-wise advantage for process reinforcement learning. *arXiv e-prints*, pages arXiv–2507.

Chinese GaoKao. 2024. Gaokao2023en.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

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

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Alan Lee and Harry Tong. 2025. Token-efficient rl for llm reasoning. *arXiv preprint arXiv:2504.20834*.

Jung Hyun Lee, June Yong Yang, Byeongho Heo, Dongyoon Han, Kyungsu Kim, Eunho Yang, and Kang Min Yoo. 2024. Token-supervised value models for enhancing mathematical problem-solving capabilities of large language models. *arXiv preprint arXiv:2407.12863*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023a. Generative judge for evaluating alignment. *Preprint*, arXiv:2310.05470.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025. Limr: Less is more for rl scaling. *Preprint*, arXiv:2502.11886.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Making large language models better reasoners with step-aware verifier. *Preprint*, arXiv:2206.02336.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2021): Findings*. To appear.

Yuliang Liu, Junjie Lu, Zhaoling Chen, Chaofeng Qu, Jason Klein Liu, Chonghan Liu, Zefan Cai, Yunhui Xia, Li Zhao, Jiang Bian, Chuheng Zhang, Wei Shen, and Zhouhan Lin. 2025a. Adaptivestep: Automatically dividing reasoning step through model confidence. *Preprint*, arXiv:2502.13943.

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

Jianqiao Lu, Zhiyang Dou, Hongru Wang, Zeyu Cao, Jianbo Dai, Yingjia Wan, and Zhijiang Guo. 2024. Autopsv: Automated process-supervised verifier. *Preprint*, arXiv:2405.16802.

Kevin Lu and 1 others. 2025. On-policy distillation.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2. Notion Blog.

Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, and 1 others. 2025. Exploring the limit of outcome reward for learning mathematical reasoning. *arXiv preprint arXiv:2502.06781*.

Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023. Let's reward step by step: Step-level reward model as the navigators for reasoning. *Preprint*, arXiv:2310.10080.

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. 2024. Llm critics help catch llm bugs. *Preprint*, arXiv:2407.00215.

OpenAI. 2024a. Gpt-5.

OpenAI. 2024b. Learning to reason with llms.

OpenAI. 2025. Introducing openai o3 and o4-mini.

Sarah Pan, Vladislav Lialin, Sherin Muckatira, and Anna Rumshisky. 2023. Let's reinforce step by step. *Preprint*, arXiv:2311.05821.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024a. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024b.

Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Open-R1 Team. 2025a. Openr1-math-220k.

OpenThoughts Team. 2025b. Open Thoughts. https://open-thoughts.ai.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *Preprint*, arXiv:2211.14275.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2024a. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *Preprint*, arXiv:2312.08935.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024b. Self-taught evaluators. *Preprint*, arXiv:2408.02666.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guofu Xie, Yunsheng Shi, Hongtao Tian, Ting Yao, and Xiao Zhang. 2025. Capo: Towards enhancing llm reasoning through verifiable generative credit assignment. *arXiv e-prints*, pages arXiv–2508.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024b. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Preprint*, arXiv:2203.14465.

Thomas Zeng, Shuibai Zhang, Shutong Wu, Christian Classen, Daewon Chae, Ethan Ewer, Minjae Lee, Heeju Kim, Wonjun Kang, Jackson Kunde, Ying Fan, Jungtaek Kim, Hyung Il Koo, Kannan Ramchandran, Dimitris Papailiopoulos, and Kangwook Lee. 2025. Versaprm: Multi-domain process reward model via synthetic reasoning data. *Preprint*, arXiv:2502.06737.

Yizhou Zhang, Ning Lv, Teng Wang, and Jisheng Dang. 2025. Fastgrpo: Accelerating policy optimization via concurrency-aware speculative decoding and online draft learning. *arXiv preprint arXiv:2509.21792*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Potential Risks

The proposed method demonstrates significant potential in augmenting the logical inference of large-scale models across various domains. However, we must candidly address the ethical challenges inherent in such technological leaps. The enhanced reasoning power could, if misappropriated, be used to facilitate intrusive surveillance systems or catalyze the production of malicious content. Addressing these challenges requires a multi-stakeholder approach: we call upon fellow researchers to prioritize safety alignment and emphasize that technological innovation must be complemented by rigorous legal and regulatory safeguards. Only through collective oversight and responsible deployment can we prevent the weaponization of these advanced AI capabilities.

B AI Usage Declaration

The authors affirm that the manuscript was originally conceived and drafted by the research team. The generative AI tool, ChatGPT, was employed exclusively for linguistic refinement, including proofreading and grammatical corrections, to ensure clarity and professional expression. The final version has been thoroughly reviewed and approved by all authors, who remain fully accountable for the integrity and accuracy of the content.

C Related Work

C.1 LLM Reasoning

Recent years have witnessed remarkable strides in the reasoning capabilities of Large Language Models (LLMs) (OpenAI, 2024b, 2025; Shao et al., 2024a; Guo et al., 2025; DeepMind, 2025; Wei et al., 2022; Anthropic, 2025; Yang et al., 2024a), leading to significant performance gains in critical domains such as mathematics (AIME, 2024; AMC, 2023; GaoKao, 2024; Cobbe et al., 2021) and programming (Jain et al., 2024). For a long time, the Chain-of-Thought (CoT) (Wei et al., 2022) paradigm has served as the foundational framework for LLM reasoning, enabling models to generate step-by-step intermediate thoughts before arriving at a final answer. This capability is typically elicited through meticulously designed prompts (Wei et al., 2022) or specifically formatted datasets (Shao et al., 2024a; Zelikman et al., 2022). Building on this, advanced reasoning topologies, such as the Tree-of-Thought (ToT) (Yao et al., 2023) and Graph-of-Thought (GoT) (Besta et al., 2024), have been proposed to further enhance reasoning structures. A major milestone was reached with OpenAI's o1 (OpenAI, 2024b), which introduced a self-reflection mechanism within the CoT process, allowing for extended and more deliberate thinking. DeepSeek-R1 (Guo et al., 2025) further advanced this field as the first open-source Long-CoT work, demonstrating that models can spontaneously evolve reflection-based reasoning patterns through Reinforcement Finetuning (RFT) using verifiable reward. Consequently, a growing body of research has emerged focusing on, such as, curating high-quality data for RFT (Ye et al., 2025; Luo et al., 2025; Hu et al., 2025; Team, 2025a,b; Li et al., 2025), and improving RL algorithms (Yu et al., 2025; Hu, 2025; Zhang et al., 2025; Lu et al., 2025; Liu et al., 2025b).

C.2 Dense Rewarding

Reinforcement Fine-Tuning (RFT) (Schulman et al., 2017; Shao et al., 2024a) optimizes LLM reasoning (Guo et al., 2025; Yang et al., 2024a,b, 2025; DeepMind, 2025; OpenAI, 2024b) through verifiable scalar rewards. DeepSeek-R1 (Guo et al., 2025) demonstrated the power of binary outcome rewards for verifiable tasks. Such rewards can range from the objective correctness of answers in mathematical or coding tasks to subjective LLM-based judgments in unstructured scenarios such as writing and translation (Wang et al., 2024b; Li et al., 2023a; McAleese et al., 2024; Zheng et al., 2023).

However, outcome rewards suffer from a severe limitation: reward sparsity (Uesato et al., 2022). The vast majority of tokens and intermediate reasoning steps generated by the model do not receive customized reward signals. Beyond traditional outcome rewards, Process Reward Models (PRMs) (Uesato et al., 2022; Li et al., 2023b; Ma et al., 2023; Pan et al., 2023; Lu et al., 2024; Liu et al., 2025a; Zeng et al., 2025) offer a more granular feedback signal and have become a focal point of recent research. The development of PRMs necessitates step-specific supervision, which can be acquired through manual annotation (Lightman et al., 2023) or via automated pipelines (Wang et al., 2024a). In the latter case, the system performs multiple rollouts originating from a given reasoning step; the resulting accuracy of these trajectories then quantifies the quality of that particular step. Although process rewards are more fine-

grained than outcome rewards, they remain relatively coarse, as each step still encompasses many tokens (Yu et al., 2025). Several recent studies have investigated token-level reward mechanisms (Cui et al., 2025; Lyu et al., 2025; Lee and Tong, 2025; Fei et al., 2025; Xie et al., 2025; Lee et al., 2024) to provide more granular supervision. Notably, PRIME (Cui et al., 2025) and SPRO (Fei et al., 2025) employ the difference in log probabilities between the reward and reference models to derive a per-token reward. Methods like TVM (Lee et al., 2024) and T-SPMO (Lee and Tong, 2025) explore token-level rewards through different lenses: the former optimizes a scalar-head LLM to predict success probabilities, whereas the latter utilizes the statistical difference in average rewards between prefixes to isolate a token's individual contribution.

D Experimental Settings

D.1 Model

Our experiments encompass a diverse set of models across different training stages: (1) Pre-trained base models, including Qwen2.5-Math-7B-Base and the Qwen3 series (1.7B, 4B, and 8B variants); (2) Instruction-tuned models (SFT), specifically DeepSeek-R1-Distill-Qwen-1.5B and OpenMath-Nemotron-1.5B; and (3) Fully post-trained models (SFT + RL), such as Qwen3-1.7B and DeepScaleR-1.5B-Preview.

D.2 Training Datasets

Our training pipeline utilizes two primary datasets to facilitate the reinforcement learning process: DeepScaleR-40K and OpenMathReasoning. We maintain a clear distinction in data application based on the target model architecture to ensure compatibility with their respective pre-training or instruction-tuning stages. Specifically, for experiments involving OpenMath-Nemotron-1.5B, we employ the OpenMathReasoning dataset as originally proposed by the authors. For all other model configurations, including our base and instruction-tuned variants, we utilize DeepScaleR-40K as the standard training corpus. This curated selection of high-quality reasoning data provides the necessary diversity and depth to support the emergence of robust long-chain reasoning capabilities across the evaluated model series.

D.3 Training Settings

For our experimental configuration, we maintain a consistent optimization framework across all runs while tailoring specific computational constraints to the model type. For all base models, we employ a global batch size of 256 and a maximum response length of 8,192 tokens; however, for instruction-tuned models, these are adjusted to a batch size of 128 and an extended sequence length of 16,384 tokens to accommodate longer reasoning trajectories. Across all experiments, we utilize a constant learning rate of 1e-6 without decay, a weight decay of 0.01, and a PPO mini-batch size of 64. Notably, we omit the KL penalty term in the objective function and set the entropy coefficient to 0. During on-policy sampling, we use a temperature of 1.0 and generate 8 trajectories per data point to ensure a robust estimation of reward signals.

D.4 Evaluation Settings

For our experimental evaluation, we conduct comprehensive testing across both in-domain and out-of-distribution (OOD) benchmarks to assess the reasoning performance and generalizability of the models. In-domain evaluation is performed on six challenging mathematical reasoning benchmarks, including AIME24 (AIME, 2024), AIME25, AMC23 (AMC, 2023), MATH-500 (Hendrycks et al., 2021), MinervaMath (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). To investigate whether the enhanced reasoning capabilities can transfer to other domains, we also conduct OOD evaluations on the GPQA benchmark (Rein et al., 2024), which consists of expert-level STEM questions in biology, physics, and chemistry, and RiddleSense (Lin et al., 2021), a benchmark focused on complex logical puzzles. During evaluation, we report the average Pass@1 Accuracy by sampling 8 trajectories for each model. The sampling parameters are consistently set with a temperature of 0.6, a top-p value of 0.95, and a maximum sequence length of 32k tokens for all models. Specifically, following the official Qwen3 configuration, we set top-$k = 20$ for all Qwen3 series models, while a top-$k = -1$ is applied to all other models.

D.5 Implementation Details of TRAC

Regarding the implementation of our proposed TRAC framework, we maintain a default trade-off weight of $\omega = 1$ in Eq. 11 to balance the learn-

ing objectives. The dynamic adjustment of the importance weights relies on several key hyperparameters across the expert and confidence modules. Specifically, for the expert-based filtering in Eq. 5, we define $\varepsilon^*_{\text{max}} = 1.0$ and $\varepsilon^*_{\text{min}} = 0.1$ as the upper and lower thresholds for expert-guided reliability, respectively. Similarly, in Eq. 9, $\varepsilon^{\text{conf}}_{\text{max}} = 1.5$ and $\varepsilon^{\text{conf}}_{\text{min}} = 0.2$ represent the bounds for the model's self-confidence scores. These boundary values were determined through a coarse-to-fine grid search on a validation subset to ensure optimal stability during reinforcement learning. For our primary experiments, we utilize Qwen3-8B as the default teacher model to provide supervisory signals. Our ablation studies further explore the impact of teacher scaling by comparing this default with Qwen3-32B; results indicate that while the larger 32B teacher provides marginal performance gains, the 8B variant offers a more computationally efficient balance for large-scale training.

E Ablation Study

We perform a comprehensive ablation study to investigate the contribution of each component in our framework. All experiments are conducted using the Qwen3-8B-Base model trained on the DeepScaleR-40K dataset. The results are summarized in Table 5.

E.1 Effectiveness of Token-level Teacher Reward

We first evaluate the fundamental impact of introducing dense teacher signals $R^{\text{teacher}}_{i,j,t}$ (defined in 1) into the sparse-reward GRPO framework. Comparing the baseline GRPO (Table 5 line 1) with the uncalibrated teacher reward setting (Table 5 line 2), we observe that simply adding raw teacher rewards yields only marginal improvements (+0.7% on AIME24).

E.2 Effectiveness of Calibration Mechanism

The pivotal contribution of our method lies in the reliability of the supervision. By applying our TRAC calibration mechanism to the raw teacher signals, we observe a dramatic performance surge. As Table 5, comparing line 6 (Full TRAC) with line 2 (Uncalibrated), the performance jumps from 26.3 to 29.5 on AIME24, a substantial +3.2 gain solely attributable to calibration. This stark contrast highlights that the efficacy of fine-grained rewarding depends less on the mere presence of dense rewards and more on their quality and reliability. TRAC effectively purifies the noisy teacher feedback, ensuring that the student only learns from high-quality, trustworthy signals.

E.3 Efficacy of Each Calibration Coefficient

To dissect the source of these gains from the calibration mechanism, we analyze the impact of each calibration coefficient individually by applying it one at a time to the raw teacher reward (line 2). As shown in Table 5,

- Problem-wise Expertise (λ^{expert}): As line 3, filtering out problems where the teacher is incompetent provides the most substantial individual boost (+1.6 on AIME24 over line 2), validating that preventing the "blind leading the blind" is crucial.
- Trajectory-wise Discrimination (λ^{disc}): Line 4 demonstrates that ensuring the teacher's preference aligns with ground-truth correctness improves performance by +1.3 on AIME24, confirming the importance of consistent ranking.
- Token-wise Confidence (λ^{conf}): Line 5 shows that down-weighting uncertain tokens brings a +0.7 gain, proving the benefit of fine-grained uncertainty management.

Most importantly, combining all three coefficients (line 6) yields a performance gain that exceeds any single component, demonstrating the synergistic and complementary effect of multi-granular calibration. In summary, each coefficient is indispensable, as they target distinct sources of noise across different granularities that cannot be substituted by one another."

E.4 Advantage Estimation Strategy for Incorporating TRAC

A critical design choice in our framework is how to integrate the dense teacher-derived rewards with the sparse outcome rewards. We compare our proposed "Reshaped Advantage Estimation" (which decouples the two terms) against a straightforward "Naive Advantage Estimation" strategy.

Definition of Naive Strategy. The Naive strategy follows the standard GRPO paradigm by first aggregating all reward signals into a unified scalar and then performing group-wise normalization. Specifically, it computes the total reward $R_{i,j,t}$ as the weighted sum of the outcome reward and the calibrated teacher reward. Since GRPO requires a trajectory-level scalar for baseline computation, it

	Teacher Reward	Calibration Mechanism λ^{expert}	λ^{disc}	λ^{conf}	Advantage Estimation	Teacher Model	AIME24	AIME25
1	-	-	-	-	reshaped	-	25.6	20.2
2	✓	-	-	-	reshaped	Qwen3-8B	26.3	21.5
3	✓	✓	-	-	reshaped	Qwen3-8B	27.9	23.0
4	✓	-	✓	-	reshaped	Qwen3-8B	27.6	22.1
5	✓	-	-	✓	reshaped	Qwen3-8B	27.0	22.5
6	✓	✓	✓	✓	reshaped	Qwen3-8B	29.5	24.0
7	✓	✓	✓	✓	naive	Qwen3-8B	26.8	22.2
8	✓	✓	✓	✓	reshaped	Qwen3-32B	29.9	24.3

Table 5: Ablation study for key designs in our method on the (a) effectiveness of calibration mechanism, (b) impact of each calibration coefficient, (c) advantage estimation strategy for incorporating teacher-derived rewards, and (d) choice of teacher model.

typically averages the token-level rewards over the trajectory length $L_{i,j}$:

$$R_{i,j,t} = R_{i,j}^{\text{outcome}} + \omega \cdot R_{i,j,t}^{\text{trac}}, \tag{13}$$

$$\overline{R}_{i,j} = \frac{1}{L_{i,j}} \sum_{t=1}^{L_{i,j}} R_{i,j,t}, \tag{14}$$

$$A_{i,j,t}^{\text{naive}} = \frac{R_{i,j,t} - \texttt{mean}\big(\{\overline{R}_{i,j}\}_{j=1}^{G}\big)}{\texttt{std}\big(\{\overline{R}_{i,j}\}_{j=1}^{G}\big)}. \tag{15}$$

where ω is a trade-off coefficient. $\texttt{mean}(\cdot)$ and $\texttt{std}(\cdot)$ are mean and standard deviation operation, respectively. This approach couples the teacher's feedback directly into the normalization statistics of the group.

Theoretical Flaw: Unfair Penalization. We identify a critical flaw in this coupled "Naive Advantage Estimation" due to the selective nature of TRAC. Our calibration mechanism intentionally suppresses the teacher reward ($R_{i,j,t}^{\text{trac}} \to 0$) when the teacher is deemed unreliable (*e.g.,* lack of expertise or high uncertainty). Crucially, this suppression stems from the *teacher's incompetence*, not the *policy model's error*. However, in the Naive formulation, a trajectory with suppressed teacher rewards will have a lower total score $\overline{R}_{i,j}$ compared to others where the teacher was active. Consequently, during group normalization, these trajectories are unfairly penalized (assigned negative advantages) simply because the teacher failed to provide guidance. This introduces a systematic bias, as the penalty for an imperfect teacher is wrongly transferred to the student policy.

Empirical Verification. The experimental results in Table 5 strongly support this analysis. Comparing line 6 (Reshaped) with line 7 (Naive) reveals a stark performance gap: using the naive strategy causes a significant performance drop of -2.7 on AIME24 (29.5 $\rightarrow$ 26.8). This confirms that our Reshaped Advantage Estimation is essential. By treating the teacher reward as an additive shaping term after the outcome-based normalization (as detailed in Sec. 3.2), we preserve the unbiased nature of the group baseline while effectively incorporating dense supervision, thereby avoiding the pitfalls of the naive coupled approach.

E.5 Choice of Teacher Model

We investigate the impact of different teacher models. As shown in Table 5, the stronger Qwen3-32B teacher (line 8) naturally yields slightly better performance than the Qwen3-8B teacher (line 6) (29.9 *vs* 29.5). It is remarkable that the 8B teacher already achieves competitive results, significantly outperforming the baseline. This indicates that TRAC is highly effective even in the scenario where a stronger external teacher is unavailable, as the calibration mechanism effectively extracts the model's own high-confidence knowledge to correct its errors. Therefore, we apply Qwen3-8B as the teacher model for training 1.5B $\sim$ 8B level model in our experiments. When scaling up to larger and more capable policy models (*e.g.,* > 8B model), employing a correspondingly stronger teacher is advisable to ensure the quality and upper bound of the supervision.

F Comparison with Other Dense Rewards

We compare our TRAC against other existing dense rewards, including explicit Process Reward Models (PRMs) and alternative token-level formulations. As shown in Table 6, our TRAC achieves superior performance across benchmarks, outperforming existing process-level and token-level rewards. Notably, our experiments reveal that calibrated sig-

Reward Granularity	Method	Fine-grained Reward Source	AIME24	AIME25
-	-	-	13.8	8.7
trajectory-level	GRPO	-	23.7	27.1
process-level	GRPO	Qwen2.5-Math-PRM-7B	24.7	22.0
	GRPO	Skywork-o1-PRM-7B	24.9	21.5
token-level	PRIME	self-trained RM	25.8	20.4
	SPRO	policy model (self-guided)	24.6	21.2
	TRAC-GRPO	Qwen3-8B	**27.1**	**23.0**

Table 6: Comparison of different dense rewards (process-level or token-level) for RL. All the experiments are trained on DeepScaleR-40K dataset and Qwen3-1.7B model (non-thinking). The first row corresponds to the baseline model before training. "RM" is an abbreviation for "Reward Model".

nals derived from off-the-shelf open-source strong LLMs already eclipse the effect offered by current open-source PRMs. This finding validates the promising potential of the teacher-derived reward paradigm, suggesting it offers a more scalable and effective pathway for dense supervision than training specialized reward models from scratch.