Don't Tell the Answer, Truly Guide the Reasoning During RL Rollouts

Anonymous ACL submission

Abstract

Reinforcement Learning (RL) has become a key driver for enhancing the long chain-of-thought (CoT) reasoning capabilities of Large Language Models (LLMs). However, prevalent methods like GRPO often fail when task difficulty exceeds model capacity, leading to reward sparsity and inefficient training. While prior work attempts to mitigate this using off-policy data often induce severe distributional mismatches that destabilize policy updates. In this work, we identify a core issue underlying these failures, which we term low training affinity, and introduce *Affinity*, the first quantitative metric for monitoring the compatibility between external guidance and the model's intrinsic policy. To address this, we propose HINT, an adaptive framework designed to enhance reasoning capabilities while explicitly preserving high *Affinity*. First, instead of revealing partial answers, HINT supplies **Meta-Hints**, which act as abstract cognitive scaffolding to guide the model in articulating solutions independently. Second, to ensure stability, we integrate **Affinity-Aware Policy Optimization (AAPO)**, which dynamically modulates the learning objective based on the *Affinity*. Extensive experiments across diverse benchmarks demonstrate that HINT achieves state-of-the-art performance, exhibiting superior stability and robust generalization to out-of-distribution tasks. Code is available on Github[1].

1 Introduction

RL methods, particularly GRPO (Shao et al., 2024), play a pivotal role in advancing long CoT reasoning (Wei et al., 2022). By avoiding the instability and overhead of training a separate value model, GRPO leverages group-based reward aggregation to deliver stable and efficient learning signals. Such RL approaches (Ahmadian et al., 2024; Shao et al., 2024; Hu, 2025; Yu et al., 2025) have become a key driver of progress in reasoning ability, enabling models to explore solution paths on verifiable problems. Building on these advances, recent reasoning models such as DeepSeek-R1 (Guo et al., 2025) and OpenAI-o1 (Jaech et al., 2024) have achieved remarkable performance on complex tasks like mathematical problem solving and programming (Jiang et al., 2024).

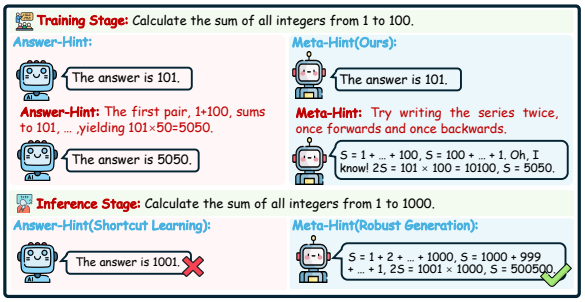


Figure 1: Comparison of Hint Mechanisms and Their Impact on Learning. **Left:** Answer-Hints provide explicit partial solutions. The model maximizes rewards by simply completing this pre-defined path, which leads to **Shortcut Learning**, characterized by the memorization of surface patterns rather than an understanding of the underlying logic. **Right:** In contrast, our Meta-Hints offer high-level cognitive scaffolding, **compelling the model to develop solution path independently** and fostering robust generation.

A critical challenge for GRPO, despite its strong empirical performance, is its tendency to generate sample groups consisting entirely of incorrect answers on tasks whose difficulty exceeds policy-model capacity, resulting in reward sparsity (Zhao et al., 2025; Yue et al., 2025). This sparsity renders gradients uninformative and leaves the policy model without effective learning signal, reducing training efficiency and wasting valuable data.

Leveraging external, off-policy data is a key method for addressing this issue. This method has been implemented in prior work through two main lines of remedies. (I) **Mixed-policy** (Yan

[1]https://anonymous.4open.science/r/HINT-9DD9/

et al., 2025; Zhang et al., 2025a; Fu et al., 2025b): Mixed-policy involves interleaving RL with SFT in a hybrid scheme to stabilize training by leveraging off-policy data. (II) **Using hints** (Li et al., 2025; Liu et al., 2025b; Zhang et al., 2025b): To mitigate reward sparsity and ensure continuous training updates, another common approach is to leverage prompts derived from the ground truth during the rollout phase, guiding the model's exploration along correct trajectories.

Despite their potential benefits, both of these approaches introduce a significant drawback rooted in a substantial distributional mismatch. We unify these methods under the term *Answer-Hints*, defined by their common reliance on ground-truth trajectories. Unfortunately, such trajectories diverge significantly from the intrinsic policy of the model, resulting in severe gradient instability (Yan et al., 2025) and deceptive signals that encourage spurious solution paths as illustrated in Figure 2.

Fundamentally, we characterize this distributional disconnect as a critical deficiency in *training affinity*. This condition arises when the over-reliance on off-policy sources shifts the data distribution too far from the intrinsic policy of the model (Fu et al., 2025a), causing excessive variance in importance sampling ratios that destabilizes optimization. Drawing on the clipping mechanism of PPO as a stability indicator (Schulman et al., 2017), we introduce the *Affinity* metric to formally quantify this dynamic through the lens of clipping frequency and update consistency.

To leverage off-policy data for enhancing model capability while preserving high *Affinity*, the guiding principle must be to **help the model articulate the solution on its own, rather than being directly told the answer.** To this end, we propose HINT, an adaptive framework that provides heuristic Meta-Hints to guide exploration without revealing partial answers. Akin to the Socratic method, this high-level scaffolding fosters robust reasoning skills by prompting the model to navigate challenges independently. To ensure these updates translate into stable improvements, we integrate Affinity-Aware Policy Optimization (AAPO), which dynamically modulates the objective based on the compatibility between guidance and the model's intrinsic distribution. Our contributions are summarized as follows:

- We formally define low training affinity as a key failure mode when integrating off-policy data into on-policy RL frameworks, and propose *Affinity*, a quantitative metric to monitor these dynamics.
- We propose the HINT framework, which synergizes Meta-Hints to guide the model in discovering effective reasoning paths independently, and AAPO to ensure training stability.
- Extensive experiments demonstrate that HINT consistently outperforms Answer-Hints methods across in-domain and out-of-domain benchmarks, exhibiting superior robustness.

2 Related Work

2.1 Reinforcement Learning for Large Language Model Reasoning.

Recent advances in RL approaches have significantly enhanced the reasoning capabilities of LLMs. Large reasoning Models (LRMs) such as OpenAI-o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025), and Kimi-1.5 (Team et al., 2025) achieve state-of-the-art performance on complex reasoning tasks (e.g., mathematics, coding, scientific problem solving) by leveraging Reinforcement Learning from Verifiable Rewards (RLVR) (Liu et al., 2025a; Hu et al., 2025; Cui et al., 2025), where automatically checkable rules provide supervision signals. Compared to earlier methods like SFT or reinforcement learning from human feedback (RLHF), RLVR has shown superior generalization and robustness (Chu et al., 2025; Snell et al., 2025). Building on this paradigm, subsequent studies have proposed improved optimization strategies and structured prompting techniques that further strengthen reasoning capabilities (Schulman et al., 2017; Wang et al., 2020). Despite this progress, a critical failure mode for existing RL methods is reward sparsity, which occurs when all rollouts in a sample fail. Overcoming this challenge is essential for enhancing the stability and sample efficiency of training.

2.2 Improving Rollout Efficiency in RL for LLMs.

A well-known challenge in methods such as GRPO is the vanishing gradient issue. This problem occurs when all trajectories in a sample group are incorrect, as the group advantage collapses to zero, yielding no gradient for policy updates (Shao et al., 2024; Guo et al., 2025). To mitigate this, some works have focused on injecting external, off-policy data to improve training efficiency and stability. This has been explored through two main strate-

gies. Some methods use mixed-policy, replacing a portion of on-policy rollouts with complete, high-quality trajectories from off-policy datasets (Yan et al., 2025; Lin et al., 2025; Xu et al., 2025; Wang et al., 2025). Others employ partial supervision, providing segments of a ground truth to rescue failed rollouts (Li et al., 2025; Liu et al., 2025b; Zhang et al., 2025b). While these approaches effectively improve rollout efficiency, their over-reliance on off-policy data can misguide policy updates, steering the model toward non-generalizable or spurious solution paths.

3 Methods

3.1 Preliminary

Following recent work (Yu et al., 2025; Yan et al., 2025), we build upon GRPO (Guo et al., 2025) and omit the KL penalty term. For each prompt, GRPO draws a group of n rollouts and computes a group-normalized advantage for every token. Mathematically, GRPO optimizes the behavior of model through the following objective function:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim Q,\{y^{(i)}\}_{i=1}^{n}\sim\pi_{\text{old}}(\cdot|q)} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|} \min\left[r_t^{(i)}(\theta)\hat{A}_t^{(i)}, \text{clip}(r_t^{(i)}(\theta), 1\pm\epsilon)\hat{A}_t^{(i)}\right], \quad (1)$$

where $r_t^{(i)}(\theta) = \frac{\pi_\theta(y_t^{(i)}|q,y_{<t}^{(i)})}{\pi_{\text{old}}(y_t^{(i)}|q,y_{<t}^{(i)})}$ is the importance sampling ratio between the current policy and the behavior policy.

Let $\{R_i\}_{i=1}^n$ denote the sequence-level rewards assigned to these rollouts. The token-level advantages $\hat{A}_t^{(i)}$ are computed by normalizing each trajectory's reward within the group:

$$\hat{A}_t^{(i)} = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^n)}{\text{std}(\{R_j\}_{j=1}^n) + \varepsilon}.$$

When all rollouts in a group are assigned identical rewards, $R_i - \text{mean}(\{R_j\}_{j=1}^n)$ becomes zero for every i, causing every advantage $\hat{A}_t^{(i)}$ to collapse to zero. Such prompts therefore provide no learning signal during training. Conversely, prompts that produce non-identical rewards across the group yield non-zero advantages and therefore generate meaningful gradients.

3.2 Quantifying the Quality of Exploration

While strategies like Answer-Hints mitigate sparsity, they often induce the "Illusion of High Rewards", a phenomenon where training rewards

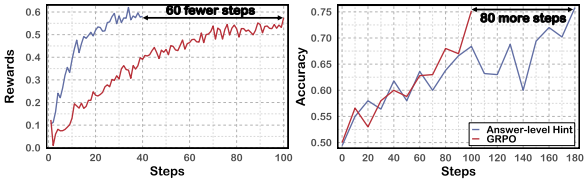


Figure 2: The *Illusion* of High Rewards During Training with the Answer-Hints Method. **Left:** Training rewards surge rapidly. **Right:** Test accuracy on MATH-500 stagnates. This discrepancy indicates that reward signals alone cannot reliably represent the actual training state.

surge while generalization stagnates (Figure 2). This discrepancy arises because strong external guidance creates distributional mismatches, inflating reward metrics while yielding uninformative or unstable gradients. Consequently, relying solely on rewards is deceptive. To capture the true training dynamics, we must look beyond reward accumulation and introduce rigorous metrics that quantify both the effectiveness and stability of policy updates.

Effective Update Ratio (EUR). EUR quantifies how many token-level updates remain unclipped under the clipped objective. Recall from Eq. (1) that GRPO generates token-level advantages $\hat{A}_t^{(i)}$ for each rollout and computes the importance sampling ratio $r_t^{(i)}(\theta)$ between the updated policy and the behavior policy. We write $\ell_t^{(i)}(\theta) = \log r_t^{(i)}(\theta)$ as the log-importance ratio, which provides a local measure of policy deviation.

We define the trust-region set as

$$\mathcal{I} = \{(i,t) : |\ell_t^{(i)}(\theta)| \le \delta\}, \quad (2)$$

which corresponds exactly to the unclipped updates. Using this notation, EUR is defined as

$$\text{EUR} = \frac{\sum_{(i,t)\in\mathcal{I}}|\hat{A}_t^{(i)}|}{\sum_{i,t}|\hat{A}_t^{(i)}|}. \quad (3)$$

In Appendix A.1, we formally show that EUR provides a principled estimate of unclipped gradient contributions and serves as a proxy for controlling the upper bound of policy divergence. Consequently, a high EUR signifies stable and meaningful policy improvement, whereas a low EUR warns that the optimizer is effectively stalling due to suppressed gradients.

Update Consistency (UC). UC quantifies the variability of the unclipped updates, where larger values indicate greater inconsistency in

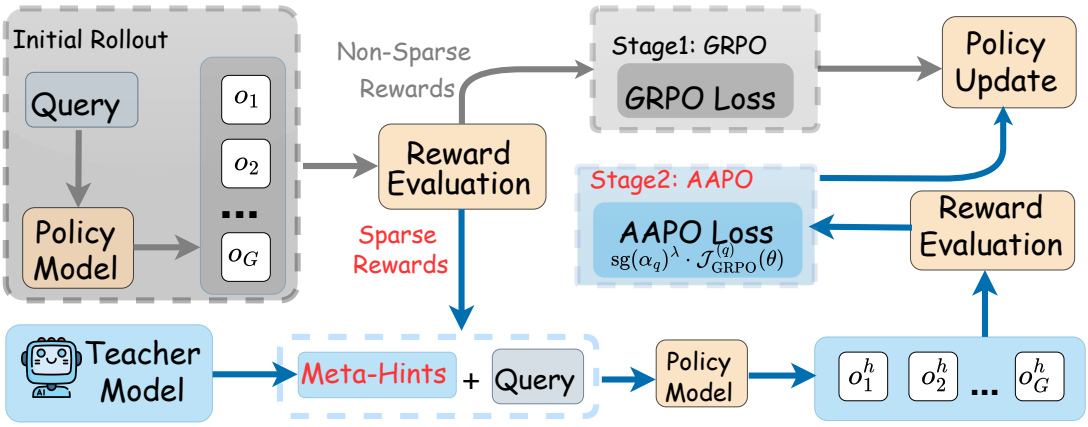


Figure 3: The HINT Framework: An Adaptive Two-Stage Rollout Process. HINT operates in two stages. **(I) Standard Rollout:** The model first samples trajectories from the original problem. If the rewards are non-sparse, the process follows the standard GRPO update path. **(II) Hint-Augmented Rescue:** If rewards are sparse (all trajectories are incorrect), the HINT mechanism is activated. The model re-rolls out conditioned on a **Meta-Hint** to guide exploration toward correct solutions. Crucially, to mitigate the potential instability introduced by external guidance, these updates are optimized via **Affinity-Aware Policy Optimization (AAPO)**, which dynamically gates gradients based on update affinity to filter out noise.

their deviation magnitudes. Using the trust-region set $\mathcal{I}$ defined in Eq. (2), we compute the advantage-weighted mean log-ratio as

$$\mu_\ell = \frac{\sum_{(i,t)\in\mathcal{I}} |\hat{A}_t^{(i)}|\, \ell_t^{(i)}(\theta)}{\sum_{(i,t)\in\mathcal{I}} |\hat{A}_t^{(i)}|},$$

with this quantity in place, UC is defined as

$$\mathrm{UC} = \sqrt{\frac{\sum_{(i,t)\in\mathcal{I}} |\hat{A}_t^{(i)}|\, \big(\ell_t^{(i)}(\theta) - \mu_\ell\big)^2}{\sum_{(i,t)\in\mathcal{I}} |\hat{A}_t^{(i)}|}}. \quad (4)$$

In Appendix A.2, we formally demonstrate that this metric is closely related to the variance of the local KL divergence. Thus, UC provides a principled indicator of update stability within the trust region, allowing us to distinguish coherent policy improvements from noisy, destabilizing steps.

Affinity. Affinity quantifies the joint quality of policy optimization by synthesizing the volume of effective updates with the stability of their deviation magnitudes. Effective training requires balancing the quantity of unclipped updates against the variance of their divergence, as neither EUR nor UC is sufficient in isolation. Using the trust-region threshold δ from Eq. (2) and setting a temperature parameter $\tau = \delta/2$, we define

$$\textit{Affinity} = \mathrm{EUR} \cdot \exp\Big(-\frac{\mathrm{UC}}{\tau}\Big). \quad (5)$$

This formulation modulates the update volume EUR with an exponential decay based on the consistency metric UC. In Appendix A.3, we provide further theoretical derivations for this composite design. Consequently, *Affinity* serves as a robust scalar indicator that yields a high score only when the optimization is both sufficiently active and stable, effectively filtering out updates that are either negligible in volume or excessively noisy.

3.3 HINT: Helping Ineffective Rollouts Navigate Towards Effectiveness

Incorporating off-policy data into on-policy RL requires maintaining high *Affinity* to ensure that external guidance translates into stable and effective policy updates. However, prior methods consistently suffer from low *Affinity*, as their reliance on Answer-Hints creates severe distributional mismatches that destabilize the training process.

Formally, we distinguish between Answer-Hints, which directly reveal intermediate steps or solutions, and *Meta-Hints*, which provide abstract strategic scaffolding. Drawing from cognitive psychology, research on feedback levels demonstrates that process-oriented guidance promotes deeper understanding and generalization, whereas task-level feedback often leads to superficial dependence (Hattie and Timperley, 2007). Despite this theoretical consensus, prior exploration methods in RL predominantly rely on Answer-Hints, thereby

failing to activate the intrinsic problem-solving capabilities of the model. To improve *Affinity*, we guide the model toward productive reasoning trajectories using Meta-Hints.

To explicitly leverage this improved *Affinity* for stable optimization, simply augmenting the data is insufficient; we require an objective that dynamically adapts to the quality of each update. To this end, we propose Affinity-Aware Policy Optimization (AAPO), which re-weights the objective using the group-level affinity score α_q:

$$\mathcal{J}_{\text{AAPO}}(\theta) = \mathbb{E}_q \left[\text{sg}(\alpha_q)^\lambda \cdot \mathcal{J}^{(q)}_{\text{GRPO}}(\theta) \right], \quad (6)$$

where $\mathcal{J}^{(q)}_{\text{GRPO}}(\theta)$ is the standard objective term defined in Eq. (1). The affinity score α_q is derived directly from Eq. (5), serving as a unified metric that synthesizes both the quantity and consistency of effective updates. Crucially, we apply the stop-gradient operator $\text{sg}(\cdot)$ to ensure that α_q functions strictly as a scalar coefficient, which prevents the model from maximizing the objective by freezing parameters to artificially boost stability. The hyper-parameter $\lambda \geq 1$ acts as a sensitivity coefficient, which suppresses the gradient contribution from unstable updates characterized by low affinity scores while preserving high-quality learning signals.

Formally, as illustrated in Figure 3, the HINT framework operates as an adaptive two-stage process that dynamically selects the optimization objective based on rollout outcomes. In the first stage, for a problem q, the model samples a set of trajectories $\{o_1, \ldots, o_G\}$ which are evaluated to obtain rewards $\{r_1, \ldots, r_G\}$. If these rewards are non-sparse (i.e., at least one is correct), the data is treated as on-policy, and we update the model using the standard GRPO objective. Conversely, if the initial rewards are sparse, we activate the rescue stage by constructing a hint-augmented query q_h with a Meta-Hint h to resample a new set of trajectories $\{o^h_1, \ldots, o^h_G\}$. To mitigate the potential instability introduced by this external guidance, we optimize these regenerated trajectories using the $\mathcal{J}_{\text{AAPO}}$ objective defined in Eq. (6).

Crucially, while q_h guides the rollout, the gradient is computed against the original query q, ensuring the model learns to solve the task independently without relying on hints as input features.

4 Experiments

4.1 Setup

Experimental Setup. Our experiments are conducted using Qwen2.5-7B, Qwen2.5-3B (Team, 2024) and LLaMa3.1-8B (Dubey et al., 2024) as backbone models. To ensure a fair and controlled comparison, we constructed a high-quality training set derived from the DAPO-Math-17K dataset (Yu et al., 2025). This process involved using Qwen2.5-72B-Instruct (Team, 2024) to generate four distinct reasoning trajectories for each problem. These outputs were then validated for correctness with Math Verify[2], from which we retained 10k fully correct samples to form our final training data. For baseline methods that require a ground-truth reference solution, we designated the shortest of the four correct trajectories for each problem.

Benchmarks. We evaluate the generalization ability of HINT on seven datasets, covering both in-distribution and out-of-distribution scenarios, without using any hint during evaluation. For mathematical reasoning, we adopt AIME24[3], MATH-500 (Hendrycks et al., 2021), OlympiadBench (He et al., 2024), and Minerva (Lewkowycz et al., 2022), which are widely used benchmarks. Since the test sets of AIME24 are relatively small, we report avg@32, while for the other datasets we use pass@1. To assess complex reasoning and out-of-distribution generalization, we further evaluate on ARC-Challenge (Clark et al., 2018), GPQA-Diamond (Rein et al., 2024), and MMLU-Pro (Wang et al., 2024). To demonstrate HINT effectiveness, we conduct systematic experiments across multiple benchmarks.

Baselines. We compare HINT against several existing methods, including: (1)**GRPO** (Guo et al., 2025): The vanilla Group Relative Policy Optimization algorithm. (2)**SFT**: Standard Supervised Fine-Tuning. (3)**LUFFY** (Yan et al., 2025): A hybrid approach that combines on-policy and off-policy training, ensuring that each sampled batch contains at least one correct trajectory. (4)**BREAD** (Zhang et al., 2025b): A binary search–based method that identifies a hint length such that the model's rollouts are neither all correct nor all incorrect, and uses this balanced point as the hint for training. Further experimental details can be found in Appendix B for full reproducibility.

[2]https://github.com/huggingface/Math-Verify

[3]https://huggingface.co/datasets/math-ai/aime24

Table 1: Main Performance Comparison of HINT against Baselines. HINT demonstrates significant performance gains on in-distribution datasets, improving the Qwen2.5-7B, Qwen2.5-3B, and LLaMa3.1-8B models by **14.5%**, **17.7%**, and **9.1%** in average accuracy, respectively. Furthermore, **the method consistently outperforms baselines on out-of-distribution data, highlighting its strong generalization capabilities.**

Methods	In-Distribution				Avg	Out-of-Distribution			Avg
	AIME24	Math	Olympiad	Minerva		ARC	GPQA	MMLU	
Qwen2.5-7B									
Vanilla	9.8	50.2	34.0	19.5	28.4	85.3	25.6	46.0	52.3
SFT	11.2	72.8	36.2	28.8	37.3	85.1	25.6	46.2	52.3
LUFFY (NIPS'25)	13.4	77.0	38.6	34.2	40.8	86.0	26.8	48.8	53.9
BREAD (ICML'25)	14.0	77.4	38.0	31.0	39.9	88.2	30.2	49.3	55.9
GRPO	13.9	76.8	38.0	31.0	39.9	88.0	29.4	48.0	55.1
GRPO + Meta-Hints	14.4	79.6	40.2	34.0	42.1	88.8	30.4	**50.2**	56.5
HINT (Ours)	**14.6**	**80.4**	**42.2**	**34.4**	**42.9**	**89.0**	32.8	**50.2**	**57.3**
Qwen2.5-3B									
Vanilla	2.9	39.8	12.0	9.8	16.1	44.8	11.4	28.8	28.3
SFT	5.3	54.8	20.6	19.6	25.1	46.4	11.0	32.0	29.8
LUFFY (NIPS'25)	5.8	62.2	29.6	22.2	30.0	70.2	15.2	34.2	39.9
BREAD (ICML'25)	6.3	62.0	29.0	24.4	30.4	72.0	18.2	**36.3**	42.2
GRPO	6.0	60.4	26.0	23.6	29.0	74.4	16.0	36.2	42.2
GRPO + Meta-Hints	6.8	66.4	30.4	25.2	32.2	77.6	18.0	35.0	43.5
HINT (Ours)	**7.4**	**68.8**	**32.8**	**26.0**	**33.8**	**78.8**	**20.4**	35.5	**44.9**
LLaMa3.1-8B									
Vanilla	0.0	9.4	2.1	3.2	3.7	0.0	0.0	0.0	0.0
SFT	0.2	14.4	4.4	8.4	6.9	52.4	18.3	26.5	32.4
LUFFY (NIPS'25)	0.5	25.2	**7.4**	14.4	11.9	66.8	25.5	33.3	41.9
BREAD (ICML'25)	0.7	23.0	7.0	**16.6**	11.8	70.4	27.2	33.9	43.8
GRPO	0.5	23.2	6.3	12.2	10.6	70.0	26.4	33.0	43.1
GRPO + Meta-Hints	0.5	26.8	6.6	14.4	12.1	74.8	28.0	36.4	46.4
HINT (Ours)	**1.0**	**28.0**	7.0	15.2	**12.8**	**75.3**	**30.4**	**39.0**	**48.2**

4.2 Main results

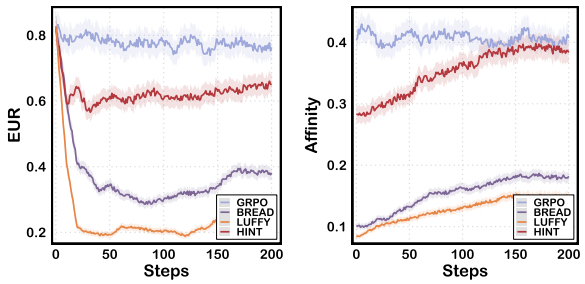
Table 1 presents a comprehensive comparison of HINT against several mainstream baselines, encompassing two Answer-Hints methods. Overall, HINT demonstrates remarkable effectiveness across all model scales, achieving average accuracy improvements of **14.5%**, **17.7%**, and **9.1%** on in-distribution datasets for Qwen2.5-7B, Qwen2.5-3B, and LLaMa3.1-8B, respectively. Our detailed analysis reveals three key findings regarding the efficacy of our data strategy, the necessity of our optimization objective, and the generalization capabilities of our method.

Meta-Hints foster genuine reasoning over answer memorization. First, our results demonstrate that process-oriented guidance is fundamentally more effective than answer-centric supervision. Across almost all benchmarks, the *GRPO + Meta-Hints* variant consistently outperforms strong baselines like LUFFY and BREAD, which rely on Answer-Hints. This performance gap indicates that unlike Answer-Hints, which risk inducing shortcut learning by forcing the model to mimic rigid paths, Meta-Hints effectively act as "cognitive scaffolding." By constraining the reasoning space rather than dictating the exact solution, they guide the model to discover valid paths independently, thereby converting sparse-reward failures into constructive training signals that foster genuine reasoning capabilities.

AAPO is essential for maximizing off-policy utility and stability. Second, while Meta-Hints provide high-quality data, the consistent performance gain from upgrading to the full HINT framework confirms the critical necessity of the AAPO objective. HINT consistently outperforms the *GRPO + Meta-Hints* baseline, exemplified by a notable increase in the average accuracy of Qwen2.5-7B from 42.1% to 42.9%. This indicates that simply augmenting data introduces inevitable off-policy noise that standard GRPO cannot fully handle. AAPO effectively mitigates this by dynamically gating gradients based on the affinity score α_q, ensuring that the model absorbs the strategic guidance from hints while filtering out unstable updates that could destabilize the policy.

HINT activates generalized reasoning capabilities beyond mathematical memorization. Finally, HINT exhibits robust out-of-distribution (OOD) generalization, suggesting that the model acquires transferable reasoning skills rather than merely overfitting to mathematical patterns. On OOD benchmarks such as ARC, GPQA, and MMLU, HINT achieves significant gains, most notably propelling the LLaMa3.1-8B model from a near-zero baseline to a competitive average of 48.2%, which confirms that our method operates on a conceptual level whereby the internalization of meta-cognitive strategies, such as problem decomposition, allows the model to transfer abstract reasoning paradigms to novel and unseen domains.

4.3 Training Dynamics



(a) EUR Metric (b) *Affinity* Metric

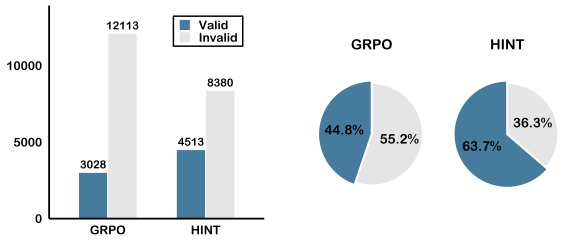
Figure 4: Comparative analysis of training dynamics. **(a)** HINT maintains a consistently high EUR, preventing the collapse seen in baselines. **(b)** Consequently, HINT achieves significantly higher *Affinity*.

To investigate the impact of various strategies on training stability, we tracked the EUR and *Affinity* metrics throughout the training process. The comparative results are plotted in Figure 4, while the analysis of UC is detailed in Appendix C.1.

HINT prevents the "EUR Collapse" typical of off-policy learning. As illustrated in Figure 4a, traditional off-policy methods suffer from a severe "EUR Collapse", where the EUR plummets to near 0.2, indicating excessive clipping and wasted samples. In sharp contrast, HINT avoids this failure mode. While there is a slight initial adjustment, HINT maintains a high steady-state EUR that is much closer to GRPO than to other on-policy methods using Answer-Hints. This confirms that HINT successfully keeps the policy updates within the trust region, ensuring high sample efficiency.

High *Affinity* validates the effectiveness of Meta-Hints. As presented in Figure 4b, HINT is the only off-policy method that achieves and sustains high *Affinity* scores. While other methods stagnate at low *Affinity* levels due to severe distributional mismatches, the *Affinity* of HINT steadily increases and tracks the GRPO baseline. This empirically validates our core proposition that the alignment between Meta-Hints and the intrinsic distribution of the model facilitates the effective assimilation of external guidance, thereby preventing it from being treated as adversarial interference.

4.4 Does hinting truly enhance sample efficiency?



(a) Throughput under fixed time budget (b) Validity rollout distribution over a full training epoch

Figure 5: Efficiency analysis. **(a)** HINT produces a higher net volume of valid trajectories despite lower total generation speed. **(b)** HINT significantly increases the proportion of effective training data.

HINT significantly enhances both sampling efficiency and the density of effective supervision. To quantify this, we conducted a two-fold analysis evaluating generation speed under a fixed 8-hour budget and global data distribution over a full training epoch. As illustrated in Figure 5a, although the inference overhead of Meta-Hints results in fewer total samples, HINT successfully yields a substantially higher volume of valid samples, defined as rollouts containing correct reasoning steps. Specifically, it produces **1,485 more valid samples** compared to the baseline, confirming that the computational cost of generating hints is outweighed by the gain in exploration success. Furthermore, Figure 5b reveals a distinct contrast in the global data distribution where HINT elevates the validity rate from 44.8% to 63.7%, representing an absolute gain of **18.9%**. This shift indicates that our method effectively steers the model toward productive regions of the solution space. By reducing the prevalence of uninformative trajectories, HINT ensures that gradient updates are derived from high-quality data, thereby maximizing the utility of the limited computational budget.

4.5 Does external feedback affect generation diversity?

Table 2: Quantitative analysis of exploration diversity using average entropy. **HINT promotes broader exploration compared to Answer-Hints.** Here, "w/ Off." denotes trajectories augmented with guidance, while "w/o Off." refers to standard on-policy rollouts.

	w/ Off.	w/o Off.	All
GRPO	–	0.143	0.143
LUFFY	–	0.174	0.174
BREAD	0.128	0.183	0.162
HINT	**0.188**	**0.198**	**0.193**

HINT fosters broad and diverse exploration rather than converging to a narrow set of solutions. To quantify this, we analyzed the output distribution using average entropy as a metric for exploration breadth, with results detailed in Table 2. The data reveals a fundamental contrast between the methods. Strategies relying on Answer-Hints, such as BREAD, exhibit the lowest entropy of 0.128 on the off-policy subset. This indicates that providing explicit answers acts as a rigid constraint that stifles exploration and forces the model to mimic a single fixed path. In contrast, HINT maintains a significantly higher entropy of 0.188 even under guidance, confirming that abstract Meta-Hints guide the reasoning process without restricting the specific trajectory. Crucially, this benefit extends to standard on-policy rollouts where HINT achieves the highest entropy of 0.198 among all methods. This demonstrates that HINT effectively prevents the tendency to overfit to repetitive solutions often observed in standard RL, enabling the model to learn a generalized policy capable of diverse reasoning.

4.6 Does HINT scale with reasoning complexity?

HINT acts as a vital cognitive scaffold that specifically enhances performance on complex reasoning tasks. To verify this, we stratified the performance on the MATH-500 benchmark across five difficulty levels as illustrated in Figure 6. On simpler tasks classified as Levels 1 and 2, all methods exhibit high competency with accuracy rates exceeding 92%. BREAD slightly outperforms the others in this regime. This suggests that answer-level hints are sufficient for recalling basic mathematical patterns.

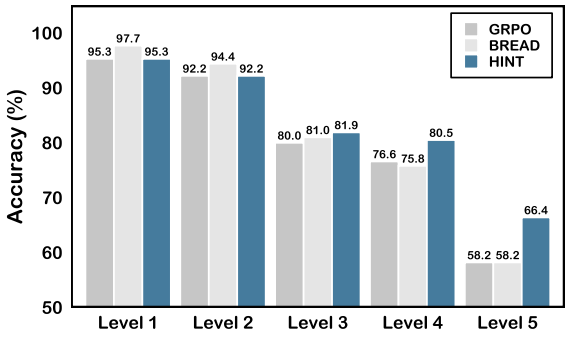


Figure 6: Performance comparison across different difficulty levels on the MATH-500 benchmark. While baseline methods plateau on hard tasks, HINT demonstrates **widening performance gaps as difficulty increases**, achieving an **8.2% absolute gain** on Level 5 problems.

However, a distinct performance divergence emerges as complexity increases. On the most challenging Level 5 problems, both GRPO and BREAD stagnate at an identical accuracy of 58.2%. In contrast, HINT achieves a robust 66.4% and marks a substantial absolute gain of 8.2%. This indicates that the benefits of Meta-Hints are positively correlated with task complexity. While the intrinsic policy suffices for simple scenarios, the strategic guidance of Meta-Hints becomes decisive for managing exploding search spaces. This breakdown confirms that HINT unlocks the potential of the model to navigate deep reasoning paths that are otherwise inaccessible to standard RL exploration.

5 Conclusion

We address the fundamental trade-off between exploration efficiency and update stability in RL for reasoning. We revealed that conventional Answer-Hints often induce low *Affinity*, leading to unstable gradients despite high rewards. Our solution, HINT, resolves this conflict by combining Meta-Hints for high-level conceptual guidance with AAPO, a novel optimization objective that dynamically filters noise based on our proposed affinity metric. Empirical results confirm that HINT significantly outperforms strong baselines, particularly in scenarios requiring generalization beyond the training distribution. By providing a principled mechanism to leverage off-policy data without compromising stability, our work offers a robust foundation for training the next generation of reasoning models. Future directions include applying HINT to broader domains and exploring its synergy with iterative self-correction mechanisms.

6 Limitations

Despite the promising results, our work has several limitations that we plan to address in future research. First, due to computational constraints, our experimental evaluation is primarily conducted on models with parameters ranging from 3B to 8B. While HINT demonstrates consistent gains across these scales, its efficacy on significantly larger models (e.g., 70B or larger) remains to be empirically verified. Second, the current implementation of HINT focuses exclusively on text-based reasoning tasks. We have not yet explored its application to multimodal scenarios, such as visual mathematical problem solving, where integrating visual cues into the meta-hint generation process presents a unique challenge. We leave the extension of our framework to larger-scale models and multimodal domains as directions for future work.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

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

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.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, and 1 others. 2025a. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. *arXiv preprint arXiv:2505.24298*.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. 2025b. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *arXiv preprint arXiv:2506.19767*.

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*.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

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*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*.

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.

Jiazheng Li, Hong Lu, Kaiyue Wen, Zaiwen Yang, Jiaxuan Gao, Hongzhou Lin, Yi Wu, and Jingzhao Zhang. 2025. Questa: Expanding reasoning capacity in llms via question augmentation. *arXiv preprint arXiv:2507.13266*.

Zhihang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. 2025. Cppo: Accelerating the training of group relative policy optimization-based reasoning models. *arXiv preprint arXiv:2503.22342*.

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

Ziru Liu, Cheng Gong, Xinyu Fu, Yaofang Liu, Ran Chen, Shoubo Hu, Suiyun Zhang, Rui Liu, Qingfu Zhang, and Dandan Tu. 2025b. Ghpo: Adaptive guidance for stable and efficient llm reinforcement learning. *arXiv preprint arXiv:2507.10628*.

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, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.

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, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Yuhui Wang, Hao He, and Xiaoyang Tan. 2020. Truly proximal policy optimization. In *Uncertainty in artificial intelligence*, pages 113–122. PMLR.

Zhenting Wang, Guofeng Cui, Yu-Jhe Li, Kun Wan, and Wentian Zhao. 2025. Dump: Automated distribution-level curriculum learning for rl-based llm post-training. *arXiv preprint arXiv:2504.09710*.

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.

Yixuan Even Xu, Yash Savani, Fei Fang, and Zico Kolter. 2025. Not all rollouts are useful: Down-sampling rollouts in llm reinforcement learning. *arXiv preprint arXiv:2504.13818*.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*.

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*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. 2025a. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. *arXiv preprint arXiv:2508.11408*.

Xuechen Zhang, Zijian Huang, Yingcong Li, Chenshun Ni, Jiasi Chen, and Samet Oymak. 2025b. Bread: Branched rollouts from expert anchors bridge sft & rl for reasoning. *arXiv preprint arXiv:2506.17211*.

Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. 2025. Echo chamber: Rl post-training amplifies behaviors learned in pretraining. *arXiv preprint arXiv:2504.07912*.

Appendix

A Theoretical Foundations of EUR, UC, and Affinity

A.1 Proofs for EUR

In this section, we provide the theoretical justification for the two main claims made in the main paper regarding the EUR: (I) EUR estimates the fraction of unclipped PPO gradient contributions (Schulman et al., 2017); (II) EUR serves as a proxy for bounding policy divergence in the sense of TRPO's monotonic improvement guarantee (Schulman et al., 2015).

A.1.1 Preliminaries

For each token step i, let

$$r_i = \frac{\pi_\theta(a_i \mid s_i)}{\pi_{\theta_{\text{old}}}(a_i \mid s_i)}, \qquad \ell_i = \log r_i.$$

PPO optimizes a clipped surrogate objective (Schulman et al., 2017), defined as

$$L_{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_i\Big[\min\big(r_i A_i,\ \text{clip}(r_i, 1 \pm \varepsilon)\, A_i\big)\Big], \quad (7)$$

and then maximizes $L_{\text{CLIP}}(\theta)$ with respect to θ.

Let $\mathcal{I} = \{i : |r_i - 1| \le \varepsilon\}$ denote the set of unclipped updates and $\mathcal{C}$ the clipped ones. The gradient of (7) decomposes as:

$$\nabla_\theta L_{\text{CLIP}} = \mathbb{E}[\,\nabla_\theta(r_i A_i)\,\mathbf{1}(i \in \mathcal{I})\,] + \mathbb{E}[\,\nabla_\theta(r_i^{\text{clip}} A_i)\,\mathbf{1}(i \in \mathcal{C})\,].$$

As noted in Schulman et al. (2017), gradients from clipped terms either vanish or are directionally distorted, while terms in $\mathcal{I}$ preserve the correct policy gradient direction.

The Effective Update Ratio is defined in the main paper as:

$$\text{EUR} = \frac{\sum_i |A_i|\,\mathbf{1}(|\ell_i| \le \delta)}{\sum_i |A_i|}.$$

A.1.2 Proof of Claim (i): EUR estimates the fraction of unclipped PPO gradient contributions

We demonstrate that EUR provides a principled empirical estimate of the proportion of gradient contributions arising from unclipped PPO updates. Recall that, for token-level PPO, the unclipped surrogate gradient at position i, denoted as g_i, is given by:

$$g_i = \nabla_\theta(r_i A_i) = A_i\, r_i\, \nabla_\theta \log \pi_\theta(a_i \mid s_i), \quad (8)$$

where $r_i = \frac{\pi_\theta(a_i|s_i)}{\pi_{\theta_{\text{old}}}(a_i|s_i)}$. For updates within the trust region (i.e., $i \in \mathcal{I}$ with $|\ell_i| \le \delta$), we have $r_i = e^{\ell_i} \approx 1$ given that ℓ_i is small. Consequently, the gradient magnitude simplifies to:

$$\|g_i\| \approx |A_i|\,\|\nabla_\theta \log \pi_\theta(a_i \mid s_i)\|.$$

Since $\|\nabla_\theta \log \pi_\theta(a_i \mid s_i)\|$ is locally bounded and relatively stable across nearby policy iterates, variations in $\|g_i\|$ are dominated by variations in $|A_i|$. Thus, the total contribution of unclipped updates to the gradient is proportional to:

$$\mathbb{E}\big[\,|A_i|\,\mathbf{1}(i \in \mathcal{I})\,\big].$$

Similarly, the total gradient magnitude (including both clipped and unclipped updates) is proportional to $\mathbb{E}[\,|A_i|\,]$. Therefore, the fraction of gradient contributions originating from unclipped updates is:

$$\text{EUR} \approx \frac{\mathbb{E}\big[\,|A_i|\,\mathbf{1}(i \in \mathcal{I})\,\big]}{\mathbb{E}\big[\,|A_i|\,\big]}.$$

By construction, this matches our definition of EUR, confirming it as an effective estimator for the fraction of gradient contributions unsuppressed by clipping.

A.1.3 Proof of Claim (ii): EUR controls policy divergence in the TRPO sense

TRPO (Schulman et al., 2015) establishes a monotonic improvement lower bound dependent on the KL divergence:

$$\eta(\theta) \;\ge\; L_{\theta_{\text{old}}}(\theta) - C \cdot D_{\text{KL}}^{\max}(\pi_{\theta_{\text{old}}}, \pi_\theta),$$

where C is a constant dependent on γ and ϵ. The token-level empirical KL divergence can be approximated by the expectation of log-ratios:

$$D_{\text{KL}}(\pi_{\theta_{\text{old}}} \| \pi_\theta) \approx \mathbb{E}_{s,a\sim\pi_{\text{old}}}\big[\,|\ell_i|\,\big].$$

Recall that EUR is the advantage-weighted fraction of updates within the trust region ($|\ell_i| \le \delta$). Let $\mathcal{C} = \{i : |\ell_i| > \delta\}$ denote the set of clipped updates. The relationship between EUR and the probability mass of $\mathcal{C}$ depends on the distribution of advantages.

Assumption 1. *The expected magnitude of advantages for clipped updates is lower bounded by a factor of the global expected magnitude, i.e.,* $\mathbb{E}[|A_i| \mid i \in \mathcal{C}] \ge \alpha \mathbb{E}[|A_i|]$ *for some* $\alpha > 0$.

Under this mild assumption, we can relate EUR to the probability of clipping $P(\mathcal{C})$:

$$\begin{aligned} 1 - \mathrm{EUR} &= \frac{\sum_{i \in \mathcal{C}} |A_i|}{\sum_{\text{all}} |A_i|} \\ &\approx \frac{P(\mathcal{C}) \cdot \mathbb{E}[|A_i| \mid \mathcal{C}]}{\mathbb{E}[|A_i|]} \\ &\geq \alpha P(\mathcal{C}). \end{aligned}$$

This implies $P(\mathcal{C}) \leq \frac{1-\mathrm{EUR}}{\alpha}$. Conversely, the contribution to the KL divergence from clipped samples is lower bounded:

$$\begin{aligned} D_{\mathrm{KL}} &\geq P(\mathcal{C}) \cdot \min_{i \in \mathcal{C}} |\ell_i| \\ &> P(\mathcal{C}) \cdot \delta. \end{aligned}$$

If EUR is low (close to 0), the advantage mass is concentrated in $\mathcal{C}$. Unless the advantages in $\mathcal{C}$ are negligibly small (which contradicts meaningful exploration), a low EUR implies a significant $P(\mathcal{C})$, forcing D_{KL} to exceed the trust region boundary δ. Therefore, maintaining a high EUR is a necessary proxy for constraining D_{KL} and preserving the validity of the TRPO bound.

A.1.4 Summary

Taken together, the results above show that EUR simultaneously quantifies the fraction of gradient mass preserved by the unclipped PPO surrogate and provides a practical handle on the policy divergence term appearing in TRPO's monotonic improvement bound. Consequently, a high EUR indicates that most updates lie within a stable trust-region regime where policy gradients remain informative, whereas a low EUR reveals that clipped updates dominate the optimization process, leading to vanishing effective gradients and ineffective learning.

A.2 Proofs for UC

In this section, we provide the theoretical justification for the UC metric introduced in the main paper. We show that UC can be interpreted as (I) an advantage-weighted measure of variability in local log-importance ratios among unclipped updates, and (II) a proxy for the variance of the local KL divergence, which is closely tied to the stability of policy updates.

A.2.1 Preliminaries

Recall that for each token step i, we define

$$r_i = \frac{\pi_\theta(a_i \mid s_i)}{\pi_{\theta_{\text{old}}}(a_i \mid s_i)}, \qquad \ell_i = \log r_i,$$

and the trust-region condition $|\ell_i| \leq \delta$ identifies the set of unclipped updates:

$$\mathcal{I} = \{\, i : |\ell_i| \leq \delta \,\}.$$

The token-level advantages are denoted by A_i, and we use the absolute values $|A_i|$ as importance weights on the contribution of each token.

Within the set $\mathcal{I}$, we define the advantage-weighted mean log-ratio:

$$\mu_\ell = \frac{\sum_{i \in \mathcal{I}} |A_i| \, \ell_i}{\sum_{i \in \mathcal{I}} |A_i|},$$

and the UC is given by the advantage-weighted standard deviation:

$$\mathrm{UC} = \sqrt{\frac{\sum_{i \in \mathcal{I}} |A_i| \, (\ell_i - \mu_\ell)^2}{\sum_{i \in \mathcal{I}} |A_i|}}. \tag{9}$$

A.2.2 UC as a measure of variability among effective updates

As shown in (9), UC is precisely the standard deviation of the log-importance ratios ℓ_i over the set of effective updates $\mathcal{I}$. A small UC indicates that the ℓ_i values within $\mathcal{I}$ are tightly concentrated around their weighted mean μ_ℓ, implying that the magnitudes of the effective updates are consistent and that the resulting policy changes are approximately uniform across token positions. In contrast, a large UC reflects substantial variability among the ℓ_i values: some effective updates correspond to very small log-ratios (i.e., conservative steps), while others lie close to the trust-region boundary (i.e., aggressive steps). Such heterogeneity results in uneven and potentially unstable policy updates.

Formally, define the normalized weights

$$\tilde{w}_i = \frac{|A_i|}{\sum_{j \in \mathcal{I}} |A_j|}, \quad i \in \mathcal{I}.$$

Then (9) can be rewritten as

$$\mathrm{UC}^2 = \sum_{i \in \mathcal{I}} \tilde{w}_i (\ell_i - \mu_\ell)^2,$$

which is the weighted variance of ℓ_i under the empirical distribution induced by the advantages $|A_i|$. Thus UC quantifies how "spread out" the log-ratios are among those updates that are not clipped.

A.2.3 Relation between UC and gradient variance

We now connect UC to the variance of the policy gradient updates. Consider the gradient contribution magnitude for a single token i within the trust

region ($i \in \mathcal{I}$), defined as $X_i = A_i r_i \approx A_i(1+\ell_i)$. The stability of training depends on the variance of this update scale. Assuming that the advantage A_i and the log-ratio ℓ_i are uncorrelated within the local trust region, we evaluate $\mathrm{Var}(X_i)$ using the standard variance decomposition approximation:

$$\begin{aligned}\mathrm{Var}(g_i) &\propto \mathrm{Var}\big(A_i(1+\ell_i)\big) \\ &\approx \mathrm{Var}(A_i) + \mathrm{Var}(A_i\ell_i).\end{aligned}$$

The first term, $\mathrm{Var}(A_i)$, represents the inherent variance of the reward structure (baseline variance), which is irreducible by policy constraints. The second term captures the variance introduced by the policy shift. Applying the product variance decomposition to $A_i\ell_i$:

$$\begin{aligned}\mathrm{Var}(A_i\ell_i) \approx \mathbb{E}[A_i^2]\mathrm{Var}(\ell_i) \\ + \mathbb{E}[\ell_i]^2\mathrm{Var}(A_i).\end{aligned} \tag{10}$$

Inside the trust region, ℓ_i is centered near 0, making the term $\mathbb{E}[\ell_i]^2$ negligible. Thus, the dominant component of the induced variance simplifies to:

$$\mathrm{Var}_{\text{induced}} \approx \mathbb{E}[A_i^2] \cdot \mathrm{Var}(\ell_i).$$

Recall that UC^2 is defined as the advantage-weighted variance of ℓ_i. Although strictly distinct from the unweighted $\mathrm{Var}(\ell_i)$, they are empirically aligned. As shown in (10), UC acts as a multiplicative gain on the gradient variance. A high UC amplifies the gradient noise proportional to the squared advantages $\mathbb{E}[A_i^2]$, thereby destabilizing the update direction. Consequently, minimizing UC is theoretically justified to dampen the variance of policy updates specifically arising from diverse importance ratios.

A.2.4 Relation between UC and local KL variability

We next relate UC to the variability in local KL divergence. The per-state KL divergence between the old and new policy can be expressed as:

$$\begin{aligned}D_{\mathrm{KL}}\big(\pi_{\theta_{\text{old}}}(\cdot \mid s) \,\|\, \pi_\theta(\cdot \mid s)\big) \\ = \mathbb{E}_{a\sim\pi_{\theta_{\text{old}}}(\cdot|s)}\big[\log r(a,s)\big].\end{aligned}$$

At the token level, the empirical KL is estimated by averaging ℓ_i over samples from $\pi_{\theta_{\text{old}}}$. Thus, the variability of ℓ_i within $\mathcal{I}$ directly reflects how much the local per-state KL fluctuates around its mean.

Since the monotonic improvement bound of TRPO (Schulman et al., 2015) relies on controlling the KL divergence, large fluctuations in ℓ_i (i.e., a high UC) suggest that certain states experience near-boundary policy shifts, even if the average KL remains small. This phenomenon effectively weakens the trust-region assumption and may induce oscillatory learning dynamics. By contrast, a low UC ensures that per-token KL changes are not only small on average but also uniformly bounded, leading to more reliable surrogate optimization.

A.2.5 Summary

In summary, UC captures the internal stability of policy updates within the trust region by measuring the advantage-weighted variance of log-importance ratios among unclipped samples. A low UC implies that effective updates move the policy in a coherent and conservative manner, whereas a high UC reveals that updates, though nominally "valid," are heterogeneous and prone to inducing instability. Together with EUR, UC provides a complementary view of both the quantity and the quality of effective policy updates during training.

A.3 Theoretical Discussion of Affinity

In this section, we provide the theoretical motivation for combining EUR and UC into the unified *Affinity* metric introduced in the main paper. We demonstrate that *Affinity* captures the joint requirements for effective and stable policy updates in PPO-style RL and relate its formulation to the principles underlying trust-region optimization.

A.3.1 Preliminaries

We briefly recall the definitions of EUR and UC. Let $\ell_i = \log \frac{\pi_\theta(a_i|s_i)}{\pi_{\theta_{\text{old}}}(a_i|s_i)}$ denote the log-importance ratio at token step i, and let $\mathcal{I} = \{i : |\ell_i| \le \delta\}$ be the set of unclipped updates. EUR measures the fraction of effective updates:

$$\mathrm{EUR} = \frac{\sum_i |A_i| \mathbf{1}(i \in \mathcal{I})}{\sum_i |A_i|}.$$

UC quantifies the internal variability of those updates. Defined formally:

$$\mathrm{UC} = \sqrt{\frac{\sum_{i\in\mathcal{I}} |A_i|(\ell_i - \mu_\ell)^2}{\sum_{i\in\mathcal{I}} |A_i|}},$$

$$\mu_\ell = \frac{\sum_{i\in\mathcal{I}} |A_i|\ell_i}{\sum_{i\in\mathcal{I}} |A_i|}.$$

A.3.2 Rationale for combining EUR and UC

As shown in Appendix A.1, EUR provides an unbiased estimate of the proportion of gradient mass

preserved by the unclipped PPO surrogate. Hence, a high EUR indicates that most updates meaningfully contribute to the policy gradient. However, EUR alone cannot ensure stability: if the log-ratios within $\mathcal{I}$ vary widely (high UC), many of those "effective" updates may be close to the trust-region boundary, potentially inducing oscillatory policy shifts.

Appendix A.2 further demonstrates that UC approximates the variance of token-level policy divergence, characterizing the consistency of unclipped gradients. Yet, UC alone is insufficient: a perfectly consistent set of updates (low UC) yields little value if EUR is small, as most gradients would be clipped, resulting in negligible policy movement.

Therefore, a high-quality update requires satisfying both conditions simultaneously: a sufficiently large proportion of effective updates (high EUR) and low variability among them (low UC).

A.3.3 Affinity as a joint stability-efficiency indicator

To encode this joint requirement into a single scalar, we define the *Affinity* metric:

$$\text{Affinity} = \text{EUR} \cdot \exp\left(-\frac{\text{UC}}{\tau}\right), \quad \tau = \frac{\delta}{2}.$$

This multiplicative formulation is motivated by two key factors:

Logical conjunction. The product structure ensures that a failure in either condition (low EUR or high UC) produces a proportionally low *Affinity*. This captures the fact that effective PPO-style updates necessitate the simultaneous satisfaction of both conditions.

Exponential penalty on inconsistency. Since UC measures the weighted variance in log-ratios, the term $\exp(-\text{UC}/\tau)$ acts analogously to an inverse smoothness regularizer, sharply penalizing updates near the trust-region boundary. The temperature term $\tau = \delta/2$ scales the penalty, ensuring it becomes substantial when UC approaches the limit of the trust region.

A.3.4 Relationship to trust-region optimization

Trust-region methods (including TRPO) rely on bounding the KL divergence to guarantee monotonic policy improvement. While EUR controls the fraction of updates satisfying the trust-region condition (reflecting the mean KL contribution), UC characterizes the variability of the local KL divergence within that region. Consequently, *Affinity* integrates both aspects: high *Affinity* indicates that the empirical KL is not only small (ensured by high EUR) but also stable across updates (ensured by low UC), aligning with the conditions under which trust-region guarantees are most effective.

A.3.5 Summary

Affinity synthesizes two complementary perspectives on PPO update quality: **(I) the proportion of effective updates (EUR)**, and **(II) the consistency of those updates (UC)**. The multiplicative formulation in (11) captures the synergy required for reliable policy improvement, providing a practical scalar diagnostic for monitoring exploration efficiency and training stability.

B Experimental Details

B.1 Detailed Setup

Platform. All of our experiments are conducted on workstations equipped with 8 NVIDIA A100 PCIe GPUs with 80GB memory.

Training Data. The training was performed using a carefully selected subset of the DAPO-Math-170K dataset (Yu et al., 2025). As the original dataset lacks ground-truth solutions, we curated our own by first using Qwen2.5-72B-Instruct to generate four reasoning trajectories for each problem. After validating the final answers with *Math-verify*, we compiled a high-quality training set of 30k problems for which all four generated trajectories were correct. For baselines requiring a ground truth, the most token-efficient of these four correct trajectories was designated as the ground truth. For our methods, we pre-generated the required heuristic hints for the entire 30k-sample training set using Qwen2.5-72B-Instruct. The prompts used in the above process will be detailed in Section B.2.

Important Parameters of HINT. HINT is implemented based on the open-source Rl framework lsrl[4]. The RL algorithm employs the GRPO advantage estimator with no KL penalty (kl_coef is set to 0.0). The clipping parameter ϵ is set to 0.2. For each group, 8 answers are generated, and the training batch size is set to 2. Distributed training utilizes the DeepSpeed library with the *AdamW* optimizer and a learning rate of 1e-6. The *train batch size* is set to 8, *gen batch size* is set to 32, *accum steps* is set to 64, *gen update steps* is set to 128, *temperature* is set to 0.9, *max response* is set to 4096. Mixed-precision training with BF16 is enabled. Memory optimization employs ZeRO Stage 2, with optimizer state offloading to CPU.

Important Parameters of Other Baselines. For baselines with publicly available code repositories, we utilized their official implementations and the parameters specified in their respective publications. For methods without public code, such as BREAD(Zhang et al., 2025b) and QuestA(Li et al., 2025), we reproduced their results using the lsrl framework, strictly adhering to the experimental parameters detailed in their papers.

Reward Setup. For our experiments, we employ a sparse, binary reward function. The reward is determined exclusively by the correctness of the final answer in a model's generated trajectory. We use the *Math-Verify* tool for automatic verification, assigning a reward of **+1** for a correct final answer and **0** for an incorrect one.

[4]https://github.com/lsdefine/lsrl

B.2 Prompt List

Prompt Template for GRPO

System: You are a helpful AI assistant. A conversation takes place between the User and the Assistant. The User asks a question, and the Assistant solves it. Please help me solve this question. Wrap only the final answer in *boxed{}*.

Question: [Question]

User:

Prompt Template for HINT

System: You are a helpful AI assistant. A conversation takes place between the User and the Assistant. The User asks a question, and the Assistant solves it. Please help me solve this question. Wrap only the final answer in *boxed{}*.

Hint: Here are some key information provided to assist you in solving the problem: [Hint]

Question: [Question]

User:

Prompt Template for Generating hints

System:

* Role and Goal

You are a top-tier problem-solving expert and a master educator. Your goal is not to solve the problem, but to distill the single most critical "Core Insight" or "Aha! Mo-

ment" required to find the solution.

* Core Task

You will be given a [Question] and its final [Answer]. Your sole job is to reverse-engineer the most likely solution path and identify the crucial "mental bridge"—the non-obvious insight, change in perspective, or core principle—that unlocks the problem.

* Thinking Framework

Analyze the Gap: First, understand the [Question] and look at the [Answer]. The core difficulty lies in the conceptual space between them. What makes bridging this gap non-trivial? Reconstruct the "Hidden" Step: Mentally construct the most elegant solution path. In that path, what is the single most pivotal, non-obvious leap of logic or application of a principle that a student is most likely to miss? Distill the Insight: Condense this pivotal leap into an extremely short, potent, and core-focused sentence. This sentence is the key that unlocks the door, not the map of the room.

* Constraints

Absolute Brevity: The insight must be a single sentence, ideally under 20 words. No Spoilers: The insight must not reveal any part of the [Answer] or the specific numbers used to calculate it. Inspirational, Not Instructional: It should inspire thought ("heuristic"), not provide a step-by-step recipe ("algorithmic"). Target the Crux: It must address the most critical linchpin that makes the entire solution possible.

* Output Format

Directly output the single, distilled "Core Insight". Do not include any other explanations, headings, or conversational text.

User:

Question:

[Question]

Answer:

[Answer]

Prompt Template for Generating Ground Truth

System: You are a helpful AI assistant. A conversation takes place between the User and the Assistant. The User asks a question, and the Assistant solves it. Please help me solve this question. Wrap only the final answer in *boxed{}*.

Question: [Question]

User:

Prompt Template for Evaluation

System: You are a helpful AI assistant. A conversation takes place between the User and the Assistant. The User asks a question, and the Assistant solves it. Please help me solve this question. Wrap only the final answer in *boxed{}*.

Question: [Question]

User:

C Further Analysis

C.1 UC Analysis

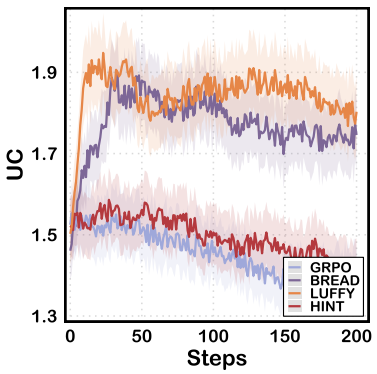


Figure 7: Analysis of Update Consistency (UC). HINT exhibits low UC, mirroring the stability of on-policy GRPO, whereas baselines show significant variance spikes.

In addition to EUR and Affinity, we analyzed Update Consistency (UC) to evaluate the variance of gradient estimates. As shown in Figure 7, there is a clear contrast in stability between the methods. **HINT maintains on-policy-level stability.** Baselines like BREAD and LUFFY quickly spike to high UC values with significant variance, reflecting unstable gradient estimates caused by large importance sampling weights. Remarkably, the UC curve of HINT remains low and stable, almost overlapping with that of standard GRPO. This demonstrates that despite incorporating external guidance, HINT preserves the low-variance training dynamics characteristic of on-policy learning, thereby guaranteeing convergence stability.

C.2 Details of HINT's Entropy

HINT Encourages Sustained Exploration. The entropy of the generation distribution serves as a key indicator of exploration diversity. As illustrated in Figure 8, HINT avoids the rapid entropy collapse observed in GRPO during the early stages of training. Instead, HINT maintains a consistently high level of entropy, indicating that the model actively explores when first introduced to the hints. This period of high exploration corresponds directly to the "EUR collapse" phase (discussed in Section 4.3), explaining that while the model initially resists the off-policy guidance, it is nevertheless engaged in a productive and diverse search of the solution space.

During the middle stages of training, HINT's entropy does not decrease monotonically. It exhibits periodic increases. We attribute this to the model

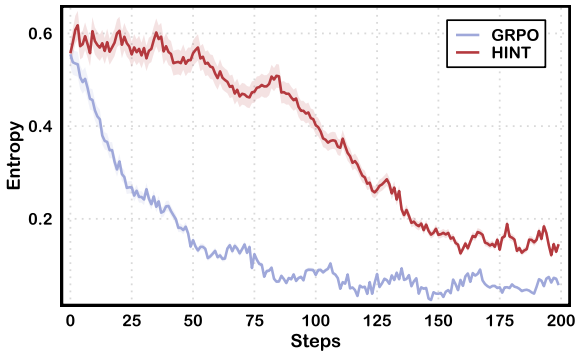


Figure 8: **HINT Prevents Entropy Collapse and Encourages Sustained Exploration.** HINT maintains a high entropy level, especially in the early stages, and stabilizes at a significantly higher value. This demonstrates that HINT's heuristic guidance fosters more continuous and diverse exploration, preventing premature policy convergence.

encountering novel types of hints and adapting its exploratory behavior to learn how to utilize them. Crucially, even after the policy stabilizes in the later stages, HINT maintains a significantly higher entropy level than GRPO. This provides strong evidence that HINT's heuristic guidance successfully fosters more continuous and diverse exploration, preventing the policy from prematurely converging to a deterministic state.

C.3 Details of HINT's Accuracy

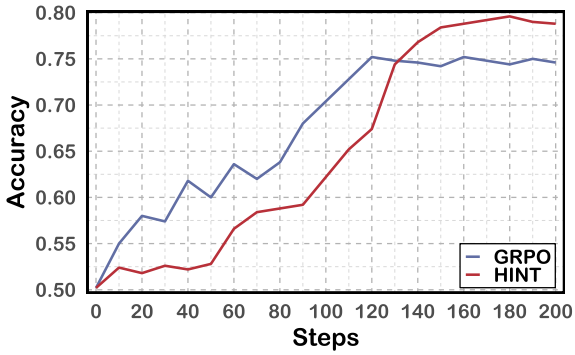


Figure 9: Accuracy of Different Methods. **HINT Achieves Higher Final Accuracy Despite Slower Initial Convergence.**

Our results reveal an interesting trade-off: while the off-policy guidance from HINT may initially slow the rate of convergence, it ultimately enables the model to achieve a higher performance ceiling. As shown in Figure 9, HINT initially exhibits a slower rate of accuracy improvement compared to GRPO. This initial lag is consistent with the early

training stages where the model shows resistance to the heuristic hints and has not yet learned to leverage them effectively. However, as training progresses, the model begins to adapt and utilize the guidance. This leads to an accelerated learning rate after approximately 100 steps, with HINT's accuracy eventually surpassing GRPO's and reaching a higher final value. This dynamic suggests that the model requires an adaptation period to master the use of heuristic hints, but once learned, this skill allows it to develop stronger and more robust capabilities.

D LLM Usage

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text.

It is important to note that the LLM was not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis.

The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

D.1 Case Study

Case Study

Question: Let a and b be positive integers such that all but 2009 positive integers are expressible in the form $ma + nb$, where m and n are nonnegative integers. If 1776 is one of the numbers that is not expressible, find $a + b$.

Meta-Hint: How many positive integers cannot be expressed in the form $ma+nb$ with nonnegative integers m, n?

Before the Hint:

By the two-coin Frobenius result, for coprime positive integers a, b the number of positive integers not representable as $ma + nb$ equals $\frac{(a-1)(b-1)}{2}$.

So

$$\frac{(a-1)(b-1)}{2} = 2009$$
$$(a-1)(b-1) = 4018$$

Factor $4018 = 41 \cdot 98$. Take $a = 42, b = 99$.

Hence $a + b = 141$.

Answer: $\boxed{141}$ ✖

After the Hint:

By the two-coin Frobenius result, for coprime positive integers a, b the number of positive integers not representable as $ma + nb$ equals $\frac{(a-1)(b-1)}{2}$.

So

$$\frac{(a-1)(b-1)}{2} = 2009$$
$$(a-1)(b-1) = 4018$$

Candidates for $(a-1, b-1)$: $(2, 4019)$, $(3, 2010)$, $(8, 575)$, $(15, 288)$, $(42, 99)$, $(50, 83)$.

Require $\gcd(a, b) = 1$, leaving: $(2, 4019)$, $(8, 575)$, $(50, 83)$.

Check 1776: representable for first two, not for $(50, 83)$.

Thus $a + b = 133$.

Answer: $\boxed{133}$ ✔