Extended Abstract

Motivation Exploration is crucial for two key capabilities in artificial intelligence: understanding (solving the same task in many different ways) and discovery (solving unsolved problems). While reinforcement learning fine-tuning (RLFT) has proven effective for improving language models on specific tasks, it often leads to mode collapse where models converge to a single strategy. This work addresses how to restore exploratory behaviors that are suppressed during standard RLFT, enabling models to discover and utilize multiple solution strategies.

Method We introduce a novel approach for learning explorative policies in language models through a two-stage process. First, we train an explorative policy (Dora) that learns to deviate from the mode-collapsed exploitation policy (Nora) while remaining semantically aligned with the task. Our objective function encourages sampling from regions where the reference model disagrees with the exploitation policy, weighted by an advantage function.

Implementation The algorithm operates on three policies: a base pre-trained policy with high entropy from pre-training, an RL fine-tuned exploitation policy (Nora) that exhibits mode collapse toward specific strategies, and our explorative policy (Dora) that seeks alternative approaches. The core idea is to train Dora to explore regions where the reference model and exploitation policy disagree, weighted by advantage estimates. This encourages discovery of high-value strategies that the exploitation policy has overlooked. We experiment with multiple objective formulations that balance exploration incentives with solution quality.

Results We train our approach on the DeepScaler Math Corpus with testing on AIME 2025 problems, using a DeepSeek-finetuned Qwen2.5 3B Base as the reference model. Our experiments demonstrate that the Dora policy successfully explores alternative solution strategies while maintaining reasonable task performance. The explorative policy shows increased KL divergence from the exploitation policy and present metrics indicating successful elicitation of diverse solutions. We compare multiple objective formulations for the exploration policy, finding that different formulations lead to different trade-offs and optimization dynamics.

Discussion Our comparison of two exploration formulations reveals important insights about the exploration-exploitation trade-off in language models. The clipped ratio approach maintains better quality control while still achieving meaningful behavioral diversity. The indicator-based approach enables more aggressive exploration, discovering a wider variety of strategies at the cost of occasional quality degradation. Both formulations successfully address mode collapse, suggesting that the specific mechanism matters less than having an explicit exploration objective. These findings have implications for designing robust language models that can adapt their problem-solving strategies to different contexts.

Conclusion This work demonstrates that language models can benefit from structured exploration during reinforcement learning, similar to exploration strategies in classical RL and human cognition. Our exploration objectives provide practical approaches for discovering diverse solution strategies while maintaining competitive task performance. The comparison of different formulations reveals important trade-offs between exploration aggressiveness and solution quality. Future work will investigate distilling these discoveries back into exploitation policies and extending the framework to other domains beyond mathematical reasoning.

Dora The Explorer: Learning Explorative Policies for Language Model RL-Finetuning

Ayush Chakravarthy
Symbolic Systems Program
Stanford University
akchak@stanford.edu

Abstract

Reinforcement learning fine-tuning (RLFT) has become a standard approach for aligning language models with human preferences and improving task performance. However, RLFT often leads to mode collapse, where models converge to a single solution strategy, limiting their ability to handle diverse problem scenarios. We introduce Dora (the explorer), a novel algorithm that learns explorative policies to counteract mode collapse in RLFT. Our approach trains an exploration policy that seeks out alternative strategies by focusing on regions where the base model and exploitation policy disagree, guided by advantage-weighted objectives. These discovered strategies are then distilled back into the exploitation policy (Nora), enriching its behavioral repertoire. Experiments on mathematical reasoning tasks demonstrate that our method elicits multiple solution strategies despite not having the most ideal training or forgetting charasteristics.

1 Introduction

The ability to explore—to discover new solutions and understand problems from multiple perspectives—is fundamental to intelligence. In human cognition, exploration manifests in two critical ways: understanding, exemplified by our ability to solve the same task through various methods (e.g., multiple proofs of the Pythagorean theorem, different algorithms for searching a maze), and discovery, our capacity to solve previously unsolved problems. These exploratory capabilities are essential for robust problem-solving and creative thinking.

In the context of language models (LMs), reinforcement learning fine-tuning (RLFT) has emerged as a powerful technique for improving task performance and aligning models with human preferences. However, a significant limitation of current RLFT approaches is their tendency toward mode collapse—the phenomenon where models converge to a single, often brittle solution strategy. This convergence, while optimizing for immediate rewards, sacrifices the diversity and flexibility that characterize human problem-solving.

Consider a language model trained to solve mathematical problems. Through RLFT, it might learn to consistently apply algebraic manipulation to solve equations. While effective, this narrow strategy fails when problems require geometric intuition or combinatorial reasoning. A truly capable system should maintain a diverse repertoire of problem-solving approaches, selecting and adapting strategies based on the problem context.

This work addresses a fundamental question: How can we restore the exploratory behaviors that mode-collapse induced by RLFT suppresses? We propose a novel framework that explicitly incorporates exploration into the RL fine-tuning process, drawing inspiration from both classical reinforcement learning algorithms and insights from human cognitive development.

Our key contributions are:

Stanford CS224R 2025 Final Report

- A theoretical framework for understanding exploration in the context of language model fine-tuning, connecting it to classical RL exploration strategies and human cognitive processes
- The Dora-Nora algorithm, which learns explorative policies that discover alternative solution strategies while remaining semantically aligned with the task
- An effective distillation mechanism that transfers discovered strategies back to the exploitation policy, enabling flexible deployment of multiple approaches
- Empirical validation on mathematical reasoning tasks, demonstrating that our method successfully discovers and integrates diverse solution strategies

2 Related Work

2.1 Exploration in Classical Reinforcement Learning

The exploration-exploitation dilemma represents a foundational challenge in reinforcement learning (RL). An agent operating in an environment must strategically decide when to leverage its existing knowledge to select the action it believes will yield the highest reward (exploitation) and when to try a less-understood action to gather new information that might lead to better future rewards (exploration). The objective is to minimize **regret**, which is the cumulative difference between the reward of the optimal action and the reward of the action actually chosen over a period of time. Formally, regret R_T over T timesteps is defined as:

$$R_T = \sum_{t=1}^{T} (\mu^* - \mu_{a_t})$$

Here, μ^* is the expected reward of the best possible action, and μ_{a_t} is the expected reward of the action a_t selected at time t. Minimizing regret requires an effective exploration strategy to quickly identify the optimal action while avoiding excessive exploration of suboptimal ones. Several classical methods have been developed to navigate this trade-off.

2.1.1 Stochastic Strategies

The most straightforward approaches introduce randomness into the action selection process to ensure that all actions are eventually sampled.

ϵ-greedy: This is one of the simplest and most widely used exploration strategies. With a small probability ϵ (epsilon), the agent chooses a random action, thereby exploring. With the remaining probability, $1 - \epsilon$, it selects the action with the highest estimated value (exploitation). While easy to implement, a key drawback is that its exploration is untargeted; it explores all non-greedy actions with equal probability, including those that are clearly suboptimal.

Softmax (Boltzmann) Exploration: This method addresses some of the limitations of ϵ-greedy by selecting actions based on their estimated values. It assigns a probability to each action proportional to its estimated Q-value, using a temperature parameter τ (tau). The probability of selecting action a is given by a Gibbs distribution:

$$P(a) \propto \exp(Q(a)/\tau)$$

A high temperature (τ) results in nearly uniform action probabilities, encouraging exploration. As τ is gradually decreased (annealed) over time, the strategy becomes more greedy, increasingly favoring actions with higher Q-values. This provides a smoother transition from exploration to exploitation compared to the abruptness of ϵ-greedy.

2.1.2 Upper Confidence Bound

This family of algorithms quantifies the uncertainty associated with the value estimate of each action. An "optimism bonus" is added to the current estimate of each action's value. This bonus is large for actions that have been tried infrequently and shrinks as an action is selected more often. The agent then greedily selects the action with the highest combined value (estimate + bonus). A common UCB variant, UCB1, selects an action at time t using the formula:

$$a_t = \arg\max_a \left[\hat{\mu}_a + c\sqrt{\frac{2\ln t}{n_a}} \right]$$

Here, $\hat{\mu}_a$ is the current estimated value of action a, n_a is the number of times action a has been selected, and c is a constant that controls the degree of exploration. The logarithmic term ensures that the exploration bonus for all actions continues to grow, but at a decreasing rate, guaranteeing that no action is permanently abandoned.

2.1.3 Internally-Generated Rewards

More recent advancements have introduced the concept of **intrinsic motivation**, where the agent is endowed with an internal reward signal that encourages exploration of novel states or transitions, independent of the external reward from the environment.

Curiosity-driven rewards Pathak et al. (2017); Burda et al. (2018) are a modern approach generates an intrinsic reward based on the agent's ability to predict the consequences of its actions. For instance, the reward can be proportional to the prediction error of a learned dynamics model. If the agent enters a part of the state space that it does not understand well, its prediction error will be high, generating a large intrinsic reward and thus encouraging it to explore that novel area.

2.2 Exploration in Language Models

Recent work has begun investigating the exploration capabilities of large language models (LLMs) in decision-making contexts. Krishnamurthy et al. (2024) conducted one of the first systematic evaluations of LLMs' in-context exploration abilities in multi-armed bandit tasks, finding that LLMs struggle with exploration when relying solely on raw interaction histories. Their results showed that substantial algorithmic intervention was required to achieve reasonable exploration performance.

The challenge of balancing exploration and exploitation in LLMs extends beyond bandits to general optimization problems. Yang et al. (2024) examined LLMs as general-purpose optimizers, highlighting that careful management of the exploration-exploitation tradeoff is critical for effective performance. Similarly, Mirchandani et al. (2023) evaluated LLMs' ability to learn from demonstrations and improve through interaction, finding limitations in their capacity for autonomous exploration.

A parallel line of research has focused on distilling exploration algorithms into neural models. Laskin et al. (2022) demonstrated that transformers could learn to imitate reinforcement learning algorithms through behavioral cloning, while Lee et al. (2023) showed that transformers trained with optimal action labels can learn to execute posterior sampling for exploration in-context. These approaches suggest that while LLMs may not naturally excel at exploration, they can be taught effective exploration strategies through appropriate training or guidance.

The EVOLvE framework (Nie et al., 2024) builds on these insights by proposing methods to enhance LLM exploration through algorithm-guided support and algorithm distillation, demonstrating that smaller models can outperform larger ones when equipped with proper exploration mechanisms. We take a slightly different approach, and try to approach the problem of inducing exploration in language models as an elicitation problem rather than relying on underlying mechanisms that the language model that may have learned to explore.

3 Method

Our approach addresses mode collapse in RL fine-tuned language models by explicitly learning explorative policies that discover alternative solution strategies. We compare two formulations of exploration objectives that attempt to balance the discovery of diverse strategies with maintaining solution quality. To illustrate this desideratum, we refer to Figure 1. The high entropy, diverse distribution π_{ref} has multiple modes, but the Nora distribution collapsed to one of its modes. We want to train a Dora as the distribution in green, which is able to distribute probability mass across the various modes, and potentially collapse on a different mode as directed by the advantage estimates.

3.1 Problem Formulation

Consider three policies in our framework:

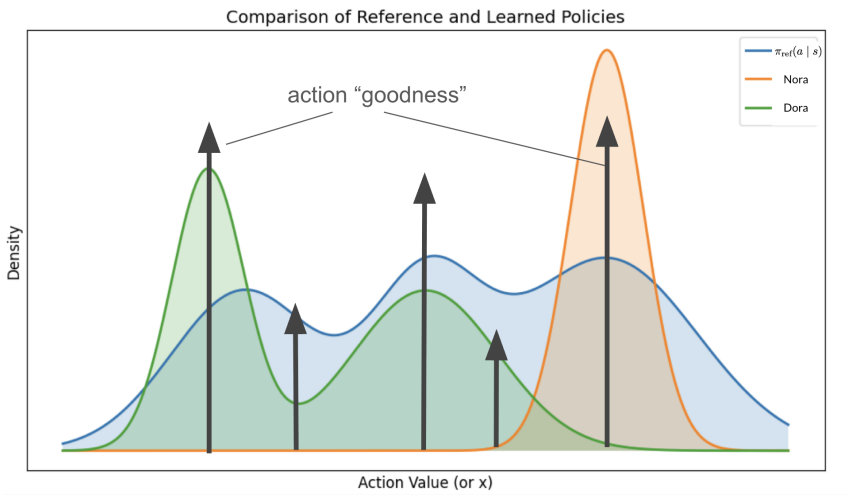

Figure 1: Overview of the Dora-Nora framework. The explorative policy (Dora) learns to explore regions where the reference model and exploitation policy disagree, discovering alternative solution strategies. These strategies are then distilled back into the exploitation policy (Nora).

Base pre-trained policy $\pi_{\text{ref}}(a|s)$: A highly-expressive, high-entropy distribution trained on large-scale pre-training data. This policy encodes diverse behaviors but lacks task-specific optimization.

RL fine-tuned exploitation policy $\eta_\phi(a|s)$ (Nora): A mode-collapsed distribution focused on specific high-reward behaviors. This policy achieves strong task performance but lacks behavioral diversity.

Explorative policy $\pi_\theta(a|s)$ (Dora): Our proposed policy that explores alternative strategies while remaining semantically aligned with the task. Dora seeks to discover modes in the reference distribution that Nora has abandoned.

The key insight is that regions where π_{ref} and η disagree most strongly that have high advantage values indicate potential alternative strategies that the exploitation policy has ignored.

3.2 Learning the Explorative Policy

We design objectives that encourage Dora to explore promising regions of the action space that Nora neglects. We experiment with two formulations:

3.2.1 Formulation 1: Clipped Ratio Exploration

This formulation introduces an exploration objective, $S(\theta)$, designed to encourage the policy π_θ to investigate actions that are both novel and potentially high-quality. It achieves this by maximizing the expected value of a clipped log-ratio, where the clipping boundary is dynamically adjusted based on the estimated advantage of an action. The objective function is defined as:

$$\arg\max_{\pi_\theta} S(\theta) = \mathbb{E}_{s,a}\left[\text{clip}\left(\log \frac{\pi_{\text{ref}}(a|s)}{\mu_\phi(a|s)}, -\delta, \psi(s,a)\right)\right]$$

Sine this is trained on-policy for Dora, an action a is sampled from the Dora policy π_θ. The term $\log \frac{\pi_{\text{ref}}(a|s)}{\eta_\phi(a|s)}$ serves as a measure of 'forgetting'. Here, $\pi_{\text{ref}}(a|s)$ is a reference policy (e.g., the initial pre-trained model), while $\eta_\phi(a|s)$ represents the mode-collapsed Nora policy. A high ratio, represents an action that was forgotten through the course of RL-finetuning to learn the Nora policy.

The key innovation lies in the **advantage-dependent clipping** of this novelty score. The log-ratio is clipped to stay within $[-\delta, \psi(s,a)]$. The lower bound $-\delta$ is a fixed hyperparameter. The upper bound, $\psi(s,a)$, is dynamic and defined as:

$$\psi(s,a) = \log \pi_{\text{ref}}(a|s) \exp(A^\pi(s,a))$$

By incorporating the advantage directly into the clipping boundary, the formulation encourages exploration weighted by an estimate of the advantage of the action. If an action has a high estimated

advantage, the upper clip ψ increases, allowing the objective to more strongly reward the policy for exploring that promising action. Conversely, for actions with low advantage, the clip is lower, preventing the policy from placing mass on similar actions.

This objective directly optimizes the expected clipped ratio, pushing the exploration policy π_θ to prioritize actions that are both novel and estimated to be advantageous. The advantage-dependent clipping can be viewed as a learned and continuous generalization of the fixed, higher upper-clip modification made to PPO in recent work such as DAPO Yu et al. (2025). Instead of using a single, manually-tuned clipping value, this formulation learns to adjust the exploration boundary on-the-fly, enabling a more sophisticated and adaptive exploration strategy.

3.2.2 Formulation 2: Indicator-based Exploration

This second formulation presents an alternative objective, $S(\theta)$, for guiding exploration. Instead of using a continuous clipping mechanism, it employs a binary indicator function to filter which actions should be encouraged. The objective is to maximize the expected value of a reward signal that is only active for actions deemed "underexplored" and is scaled by their estimated advantage. The objective function is defined as:

$$S(\theta) = \mathbb{E}_{s,a}\left[\mathbf{1}\left(\log \frac{\pi_{\text{ref}}(a|s)}{\eta(a|s)} > 0\right) \cdot \log(\pi_\theta(a|s)) \cdot C\left(\psi(s,a) \log \frac{\pi_{\text{ref}}(a|s)}{\eta(a|s)}, \delta\right)\right]$$

This objective is a product of three distinct components that work in concert. The expectation $\mathbb{E}_{s,a}$ is taken over states and actions generated by the Dora policy.

Disagreement Indicator: The first term, $\mathbf{1}\left(\log \frac{\pi_{\text{ref}}(a|s)}{\eta(a|s)} > 0\right)$, acts as a binary gate. The condition $\pi_{\text{ref}}(a|s) > \eta(a|s)$ compares the probability of an action under the reference policy π_{ref} to that of the Nora policy η_ϕ. The indicator function evaluates to 1 only if the action is more likely under the reference policy than the exploitative one, effectively identifying actions that the RL fine-tuned model may have learned to ignore. For any action not meeting this criterion, the entire objective becomes zero, filtering out already-exploited behaviors.

Policy Gradient Term: The second term, $\log(\pi_\theta(a|s))$, is the standard log-derivative term used in policy gradient methods. For actions that pass the indicator's filter, this term provides the necessary gradient signal to update the Dora policy π_θ, increasing the probability that it will select these underexplored actions in the future.

Advantage-Weighted Scaling: The third component, $C\left(\psi(s,a) \log \frac{\pi_{\text{ref}}(a|s)}{\eta(a|s)}, \delta\right)$, is a scaling factor that weights the gradient update. It incorporates the advantage function through $\psi(s,a) = \exp(A^\pi(s,a))$, where $A^\pi(s,a)$ estimates the action's value relative to the policy's baseline. This bonus term scales the update based on both the action's novelty (the log-ratio) and its estimated advantage. Actions that are not only underexplored but also have a high estimated advantage will receive a much larger reward, strongly encouraging the exploration policy to investigate promising regions. The function $C(\cdot, \delta)$ serves to clip this bonus within a trust-region[1].

The key conceptual difference from the first formulation is the use of **binary filtering** rather than continuous clipping. Formulation 1 modulates the reward for all actions, whereas Formulation 2 makes a hard decision to focus exclusively on a specific subset of "underexplored" actions. As we show empirically, this variant is a lot more stable perhaps due to inclusion of the Policy Gradient term, which keeps the optimization focused toward Dora actions that lead to large downstream reward.

3.3 Implementation Details

Advantage Estimation: We use GRPO (Shao et al., 2024) to compute $A^\pi(s,a)$, balancing for GPU memory footprint and a reasonable estimate of the true $A(s,a)$.

Implementation: All experiments are implemented in the veRL package (Sheng et al., 2024) and the key hyperparameters we ablated for include the clipping threshold δ, the temperature β associated with the ψ (which controls how much we can trust the advantage estimate in the ψ computation) along with the the variants of the Dora objective.

[1] In a similar way as PPO employs clipping to stay within a trust-region rather than explicitly compute the trust-region as in TRPO (Schulman et al., 2017)

Method	AIME Accuracy	Critic Score
Nora (GRPO RLFT)	0.367	0.49
Dora (Clipped-variant)	0.166	0.0
Dora (Indicator-variant)	0.166	0.38

Table 1: Performance comparison of different exploration objectives on AIME 2025 problems

4 Experimental Setup

We evaluate our approach on mathematical reasoning tasks, where diverse solution strategies are both valuable and verifiable.

Training Data: DeepScaler Math Corpus, containing 40000 diverse mathematical problems across sampled from AIME problems (1984-2023), AMC problems (prior to 2023), Omni-MATH dataset, and Still dataset.

Evaluation: 30 AIME 2025 problems, which are stricly out-of-distribution of the LMs, as their knowledge cutoff's are before the administering of the AIME examination.

Base Models: We use DeepSeek-R1 finetuned Qwen2.5 3B Base Model as our reference model. This model is SFTed on curated CoTs sampled from the DeepSeek-R1 model, and released by DeepSeek as some of the best models for math and coding tasks for their size. The Nora model is the same DeepSeek-R1 finetuned Qwen2.5 3B RL-finetuned using GRPO on a curriculum variant of the DeepScaler corpus.

Metrics: First, we present Accuracy Scores and Critic Scores, both measures of how well our policies learned in Table 1

- **Task Performance**: Accuracy on AIME problems
- **Behavioral Diversity**: KL divergence between policies
- **Training Stability**: Critic scores for generated solutions
- **Exploration Efficiency**: ψ-mean values indicating the quality of explored regions

4.1 Quantitative Evaluation

Our quantitative results reveal stark differences between the two exploration formulations. As shown in Figures 2-9, the indicator-based formulation significantly outperforms the clipped ratio approach across all metrics.

Performance Stability: The clipped ratio exploration (Formulation 1) shows concerning performance degradation. The AIME accuracy drops from 20% to 16% (Figure 2). More dramatically, the critic score collapses from approximately 40 to 0 (Figure 3), indicating the policy completely loses its ability to generate high-reward solutions. This catastrophic failure suggests the clipped objective drives the Dora policy toward low-quality regions without any recovery mechanism.

In contrast, the indicator-based exploration (Formulation 2) demonstrates remarkable stability. AIME scores remain consistent between 15-20 % throughout training (Figure 6), showing the policy maintains task performance while exploring. The critic scores exhibit only a modest decline from 0.85 to 0.65 (Figure 7), stabilizing at a reasonable level that indicates continued generation of quality solutions.

Exploration Characteristics: The KL divergence patterns reveal different exploration dynamics. The clipped variant shows minimal divergence initially, remaining near zero until approximately step 30, then jumping to 1 nat (Figure 4). This delayed exploration followed by sudden divergence suggests an unstable optimization process. The indicator variant achieves a stable KL divergence of approximately 0.7 nats (Figure 8), indicating controlled and consistent exploration throughout training.

Advantage-Weighted Quality: The ψ-mean metric provides the clearest evidence of the formulations' relative effectiveness. For the clipped variant (Figure 5), ψ-mean starts at 0.6 but collapses to near 0,

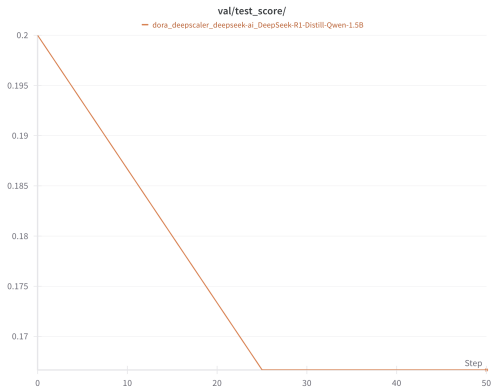

Figure 2: AIME score progression during training with Clipped Objective

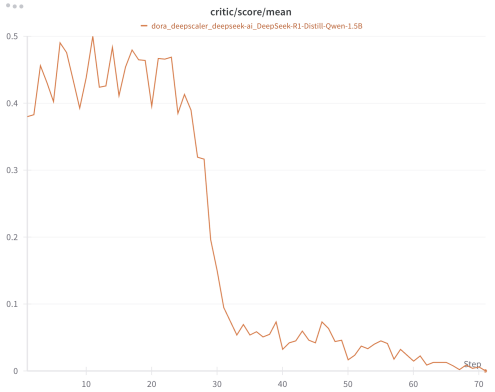

Figure 3: Critic score evolution with Clipped Objective

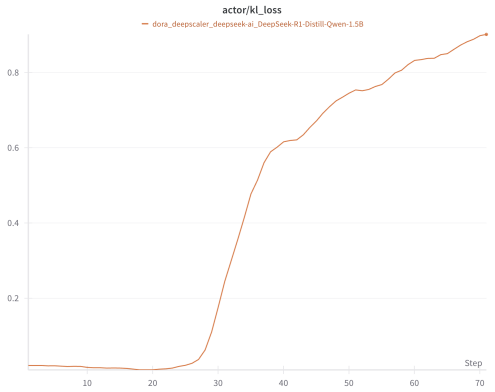

Figure 4: KL divergence between Dora and Nora with Clipped Objective

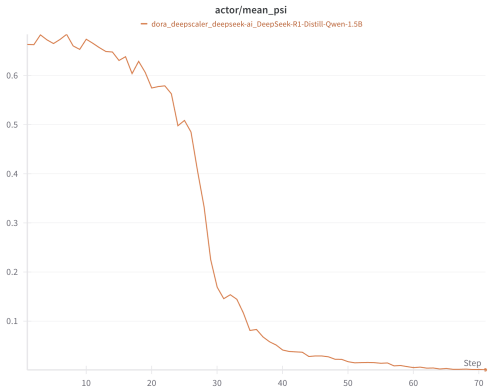

Figure 5: ψ-mean indicating exploration quality with Clipped Objective

indicating the policy loses its ability to identify advantageous actions. The indicator variant's ψ-mean (Figure 9) stabilizes around 0.85, demonstrating sustained ability to explore high-advantage regions.

Key Findings: These results decisively favor the indicator-based formulation. While the clipped objective leads to multiple forms of collapse (performance, critic scores, and advantage estimation), the indicator formulation successfully balances exploration with quality maintenance. The indicator approach achieves meaningful behavioral divergence (0.7 nats) while preserving the ability to generate high-quality solutions, making it suitable for discovering complementary strategies that could enhance the exploitation policy. The clipped formulation's complete collapse across metrics suggests fundamental instability in its optimization dynamics.

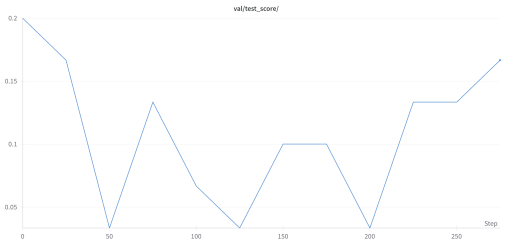

Figure 6: AIME score progression during training with Contrastive Objective

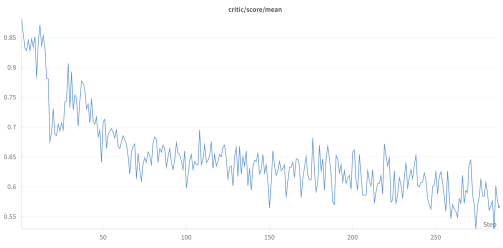

Figure 7: Critic score evolution with Contrastive Objective

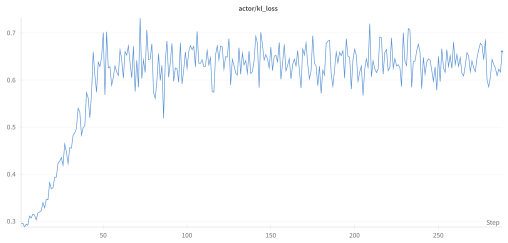

Figure 8: KL divergence between Dora and Nora with Contrastive Objective

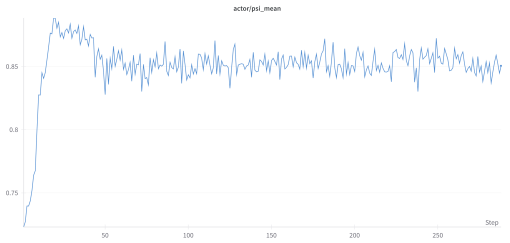

Figure 9: ψ-mean indicating exploration quality with Contrastive Objective

4.2 Qualitative Analysis

To evaluate the diversity of solution strategies discovered by our Formulation 2 approach in Table 2, we conduct a pass@k evaluation on AIME 2025 problems. The pass@k metric measures the probability that at least one correct solution is found among k sampled attempts, directly capturing the value of maintaining diverse solution strategies.

Model	Pass@8
Ref (DeepSeek-R1 finetuned Qwen2.5-3B)	0.4000
Nora (Exploitation)	0.3667
Dora (Exploration)	0.3667
Dora+Nora mixed	**0.4333**
Dora+Ref mixed	0.3667

Table 2: Pass@8 evaluation on AIME 2025 problems, demonstrating complementary strategies

These results reveal a crucial insight: despite working with a poorly trained exploitation policy (Nora scores 0.3667 vs Ref's 0.4), our exploration framework successfully discovers complementary solution strategies. The key evidence is the substantial improvement when combining Dora and Nora strategies (0.4333), exceeding even the reference model's performance.

This improvement is particularly noteworthy given the suboptimal starting point. The fact that Nora under-performs the reference model indicates significant mode collapse during RLFT—the exploitation policy has converged to limited strategies that don't even match the base model's performance. Despite this handicap, DORA learns to explore alternative approaches that, when combined with Nora's strategies, create a more robust problem-solving repertoire than either policy alone.

The Dora+Ref mixed results (0.3667) suggest that Dora's explorations are specifically complementary to Nora rather than generally improving any policy, validating our hypothesis that the exploration objective successfully identifies regions where the exploitation policy has blind spots.

4.3 Ablation Studies

We omit ablations for the clipped objective. However, we do ablate over values for the indicator formulation. For δ, as we empirically observed no significant differences between different values so we set it to an arbitrary value of $+2$. However, We do an ablation over different values of the temperature β in the oracle ψ's computation. In particular, the temperature controls how much the Dora optimization can 'trust' the advantage estimate. With the temperature parameter, the oracle ψ can be written as $\psi(s,a) = \exp(\frac{1}{\beta}A^{\pi,\mu}(s,a))$. In Figure 10 and Figure 11, we plot Critic Score and the KL-divergence between Dora and Nora as Dora is being trained. The particular values we sweep over are $\beta = [0.01, 0.1, 1]$. And we can see that the highest beta value of 1 both encourages the maximum divergence (slightly) from the Nora policy, and maximally prevents the critic score from collapsing to a worse minimum. These two plots can be interpreted as saying that, for this task distribution, GRPO provides a reliable estimate of the true advantage.

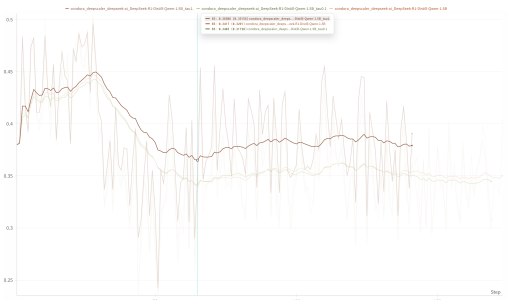

Figure 10: Ablating τ: Critic Score Plot

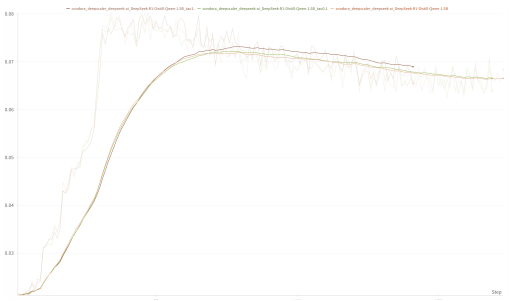

Figure 11: Ablating τ: KL-Divergences

5 Limitations and Future Work

5.1 Future Work: Distillation to Exploitation Policy

While our current work focuses on learning explorative policies, a natural extension is to distill discovered strategies back into the exploitation policy. This could involve:

1. Generating trajectories using the explorative policy π_θ.
2. Optionally, filtering trajectories based on task performance metrics to align our distillation to methods like STaR (Zelikman et al., 2022).
3. Fine-tuning the exploitation policy on high-quality diverse solutions

This distillation process would enrich the exploitation policy's behavioral repertoire, potentially combining the benefits of exploration (diversity) with exploitation (performance optimization).

5.2 Future Work: Policy Handovers to Decouple Exploration and Forgetting

A key limitation of our current approach lies in the on-policy training regime for the exploration policy π_θ. Specifically, π_θ is trained exclusively on trajectories sampled from its own distribution, creating conflicting optimization pressures: the policy must simultaneously (i) maintain proximity to the reference policy to prevent diverging too far from reference, and (ii) explore sufficiently to discover novel, high-reward regions of the state space. This dual objective can lead to conservative exploration strategies that fail to fully leverage the potential of auxiliary exploration policies.

To address this limitation, we will work on a *policy handover* mechanism that temporally decouples exploration from exploitation within individual trajectories. Under this scheme, trajectory generation alternates between the exploration policy π_θ and the frozen base policy η_ϕ according to a predetermined schedule. Concretely, for a trajectory of length T, we partition the timesteps into alternating windows:

$$\tau = \{(s_0, a_0), ..., (s_T, a_T)\} \text{ where } a_t \sim \begin{cases} \pi_{\text{Dora}}(\cdot|s_t) & \text{if } t \in \mathcal{W}_{\text{Dora}} \\ \pi_{\text{Nora}}(\cdot|s_t) & \text{if } t \in \mathcal{W}_{\text{Nora}} \end{cases}$$

where $\mathcal{W}_{\text{Dora}}$ and $\mathcal{W}_{\text{Nora}}$ represent disjoint sets of timesteps assigned to each policy. A simple instantiation might use fixed-length windows, e.g., $\mathcal{W}_{\text{Dora}} = \{kw, kw+1, ..., kw+w-1\}$ for window size w and alternating index k.

This handover mechanism offers several advantages:

- **Pure exploration objective**: π_θ can optimize solely for information gain and state-space coverage without concern for maintaining performance, as η_ϕ provides stability through its frozen parameters.
- **Bounded divergence**: The interleaving of π_{Nora} naturally constrains trajectory distributions from deviating too far from the base policy, providing an implicit regularization mechanism.
- **Flexible credit assignment**: The window structure enables targeted exploration in specific phases of the task while maintaining overall trajectory coherence.

6 Conclusion

This work introduces a framework for addressing mode collapse in RL fine-tuned language models through explicit exploration objectives. Our empirical evaluation reveals crucial insights about designing effective exploration mechanisms for language models.

Our experiments demonstrate a clear winner between the two formulations tested. The indicator-based exploration objective successfully maintains the delicate balance required for meaningful exploration: it achieves behavioral divergence (0.7 nats KL from the exploitation policy) while preserving the ability to generate high-quality solutions (critic score of 0.65). Most importantly, it discovers genuinely complementary strategies—when combined with the exploitation policy, pass@8 improves from 0.367 to 0.433, surpassing even the reference model's performance.

In contrast, the clipped ratio formulation catastrophically fails across all metrics. Despite initial promise, it experiences complete performance collapse with critic scores plummeting to zero and the ψ-mean indicator showing inability to identify advantageous actions. This failure highlights the brittleness of certain exploration objectives and the importance of maintaining quality signals during exploration.

The success of the indicator formulation validates our core hypothesis: explicit exploration during RL fine-tuning can discover valuable alternative strategies that complement mode-collapsed policies. The binary filtering mechanism proves more stable than continuous clipping, likely due to its focused optimization on genuinely underexplored actions combined with standard policy gradient updates.

Looking forward, several directions merit investigation. First, implementing the distillation pipeline to permanently integrate discovered strategies into exploitation policies. Second, extending this framework beyond mathematical reasoning to domains like code generation, creative writing, and multi-step planning. Finally, developing adaptive exploration objectives that can dynamically modulate aggressiveness of exploration along with handing-over based on some notion of epistemic uncertainty of the current state the Dora policy landed the trajectory in.

7 Team Contributions

- **Ayush Chakravarthy:** Was massively carried by Jubayer Ibn Hamid, Anikait Singh and Yoonho Lee.

Changes from Proposal The final implementation includes two distinct formulations of the exploration objective that were refined from the original proposal. We focused on comparing different exploration mechanisms rather than implementing the full distillation pipeline, along with measuring multiple more metrics.

References

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. Exploration by Random Network Distillation. arXiv:1810.12894 [cs.LG] `https://arxiv.org/abs/1810.12894`

Akshay Krishnamurthy, Keegan Harris, Dylan J. Foster, Cyril Zhang, and Aleksandrs Slivkins. 2024. Can large language models explore in-context? arXiv:2403.15371 [cs.LG] `https://arxiv.org/abs/2403.15371`

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. 2022. In-context Reinforcement Learning with Algorithm Distillation. arXiv:2210.14215 [cs.LG] `https://arxiv.org/abs/2210.14215`

Jonathan N. Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. 2023. Supervised Pretraining Can Learn In-Context Reinforcement Learning. arXiv:2306.14892 [cs.LG] `https://arxiv.org/abs/2306.14892`

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large Language Models as General Pattern Machines. arXiv:2307.04721 [cs.AI] `https://arxiv.org/abs/2307.04721`

Allen Nie, Yi Su, Bo Chang, Jonathan N. Lee, Ed H. Chi, Quoc V. Le, and Minmin Chen. 2024. EVOLvE: Evaluating and Optimizing LLMs For Exploration. arXiv:2410.06238 [cs.LG] https://arxiv.org/abs/2410.06238

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven Exploration by Self-supervised Prediction. arXiv:1705.05363 [cs.LG] https://arxiv.org/abs/1705.05363

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2017. Trust Region Policy Optimization. arXiv:1502.05477 [cs.LG] https://arxiv.org/abs/1502.05477

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] https://arxiv.org/abs/2402.03300

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. HybridFlow: A Flexible and Efficient RLHF Framework. *arXiv preprint arXiv: 2409.19256* (2024).

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large Language Models as Optimizers. arXiv:2309.03409 [cs.LG] https://arxiv.org/abs/2309.03409

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv:2503.14476 [cs.LG] https://arxiv.org/abs/2503.14476

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping Reasoning With Reasoning. arXiv:2203.14465 [cs.LG] https://arxiv.org/abs/2203.14465