# Research Summary: Optimization Dynamics and Stability Analysis in Group-Based Reward Policy Optimization (GRPO)

**Author:** Zhizheng Zhao

## 1. Abstract

This research investigates the limitations of **Group-Based Reward Policy Optimization (GRPO)** in Large Language Models (LLMs), specifically targeting sample efficiency and exploration in sparse-reward environments. The study is structured into two primary thrusts: **(1) Enhancing Advantage Granularity** (transitioning from sample-level to token-level credit assignment) and **(2) Mitigating Distribution Mismatch** (addressing gradient vanishing in hard prompts). Through a series of five controlled experiments, this work analyzes the trade-offs between bias, variance, and exploration stability in RLHF.

## 2. Thrust I: Enhancing Advantage Granularity

**Problem Identification:** GRPO assigns a uniform scalar advantage ($A_{sample}$) to all tokens in a response. This coarse granularity fails to distinguish "critical reasoning steps" from "hallucinations" within the same sample.

### 2.1 Investigation: Learned Token-Level Critic

- **Hypothesis:** A parameterized critic model $\phi$ can learn to assign a bias $b_\phi(t)$ to individual tokens, decomposing the sample-level reward into fine-grained contributions.
- **Methodology:** The advantage function was modified to:

  $A_{token}(t) = A_{sample} + b_\phi(t)$

  The critic was trained via an auxiliary loss minimizing the correlation between bias and group advantage:

  $\mathcal{L}(\phi) = -A_{sample} \cdot \sum_{t \in \tau} b_\phi(t)$
- **Analysis of Outcome:** Experiments revealed high training instability. The introduction of a token-level critic introduced significant **variance** into the gradient estimation. The added computational complexity of the critic did not yield convergence improvements, suggesting that without dense ground-truth supervision, unsupervised credit assignment introduces more noise than signal.

### 2.2 Investigation: Probability-Based Importance Sampling

- **Hypothesis:** High-reward samples containing low-probability tokens ("Rare Gems") provide high-value learning signals that are often washed out by standard gradients.
- **Methodology:** Designed a dynamic **Probability Compensation** factor $C$ to amplify gradients for improbable correct tokens:

  $C = 2 - \frac{p_t}{p_{max}}, \quad A_{token}(t) = A_{sample} \cdot C$

- **Analysis of Outcome:** This method led to **policy divergence** in early training phases. The amplification of low-probability tokens acts similarly to an aggressive importance sampling weight, which unbounded the gradient variance. This highlights the necessity of clipping or trust-region constraints even when scaling advantages.

# 3. Thrust II: Addressing Distribution Mismatch & Exploration

**Problem Identification:** In difficult tasks, the "All-Fail" phenomenon occurs where all $N$ samples in a group yield zero reward ($R = 0 \Rightarrow A = 0$). This leads to **Zero-Gradient** updates, causing the model to ignore the most challenging subset of the data distribution.

## 3.1 Investigation: Dynamic Resource Re-allocation

- **Hypothesis (Concentration of Force):** Dynamically culling "solved" prompts (where mean reward is maximized) to re-allocate sampling budget to "unsolved" prompts would break the exploration barrier.
- **Analysis of Outcome:** Results showed negligible improvement. The finding suggests that the exploration barrier in "All-Fail" groups is **exponential, not linear**. Simply increasing sampling $N$ linearly (via resource reallocation) is insufficient to cover the vast action space of reasoning paths.

## 3.2 Investigation: Negative Advantage for Failure Suppression

- **Hypothesis:** Consistently failing regions in the policy distribution represent "known bad" modes. Explicitly suppressing them forces probability mass redistribution to unexplored regions.
- **Methodology:** For groups where $\forall i, R_i = 0$, manually override the advantage:

  $$A_{sample} = -c \quad (\text{where } c > 0)$$

  New Gradient Update: $\nabla \mathcal{L} \propto c \cdot \nabla \log \pi(t)$ (Minimizing likelihood of failed paths).
- **Analysis of Outcome:** This approach demonstrated **positive potential**, showing minor uplifts in exploration. *Note:* This intuition was later validated by the broader research community (e.g., **NGRPO**), confirming that negative feedback is a viable exploration driver in sparse reward settings.

## 3.3 Investigation: Dense Reward Shaping via GT Likelihood

- **Hypothesis:** Use the likelihood of the Ground Truth (GT) answer, conditioned on the model's generated reasoning, as a dense reward signal ($R_{new} \in [0, 1]$) to replace binary zeros.

  $$R_{new} = P_\theta(\text{GT Answer} \mid \text{Generated Chain-of-Thought})$$
- **Analysis of Outcome: Failure Mode Analysis:** The model learned to optimize the *connection* between "Arbitrary Reasoning" and "Correct Answer," leading to **Logical Incoherence**. This experiment provided a critical insight: Reward Shaping must respect the causal dependency of reasoning; optimizing $P(y|x)$ without constraining the validity of $x$ leads to "Reward Hacking."

# 4. Key Research Insights

This series of explorations provided three fundamental insights into RLHF dynamics:

1. **The Granularity-Variance Trade-off:** Moving from sample-level to token-level advantage without dense supervision introduces variance that destabilizes training, outweighing the theoretical benefits of fine-grained feedback.
2. **Exploration requires Directed Guidance:** Passive methods (like resource re-allocation) fail to solve hard exploration problems. Active suppression of bad modes (Negative Advantage) is more effective.
3. **Structural Integrity in Reward Shaping:** Dense rewards derived from likelihoods can decouple reasoning from conclusions. Effective reward models must assess the *process*, not just the *conditional probability* of the outcome.