



FIPO: Eliciting Deep Reasoning with Future-KL Influenced Policy Optimization

The Ship of Theseus of Language Models,
Dark Secrets/ Achilles' Heel

Presenter Shuo Yang

2026.3.24



FIPO: Eliciting Deep Reasoning with Future-KL Influenced Policy Optimization



Distributional shifts
“**sparse but critical**”

Dark Secrets
“**Oops Moment**”

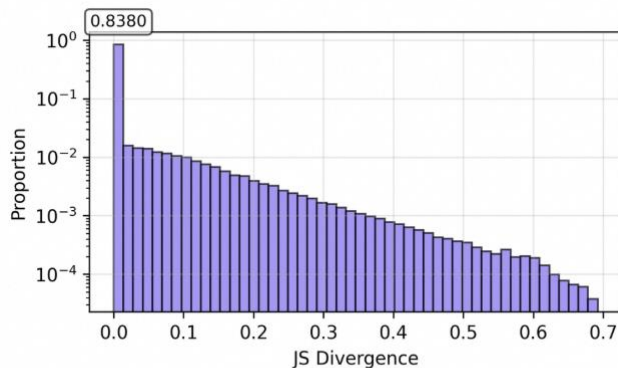
- The first pure RL recipe to surpass o1-mini
- Beat DAPO from 50.0% to a peak of 58.0%

Towards an Understanding of RLVR, Part I: The Ship of Theseus of Language Models

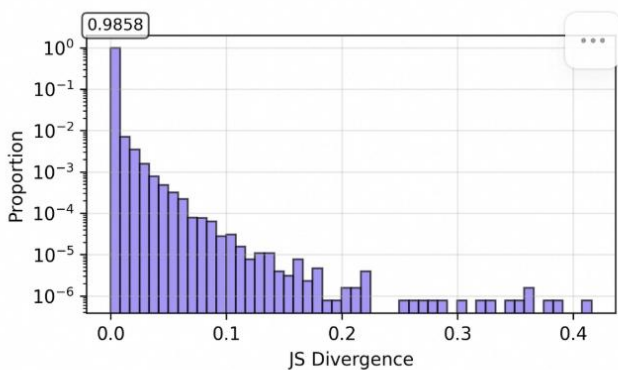
□ Motivation: how broadly are distributional shifts distributed across token positions?

$$JS_t = \frac{1}{2} D_{KL}(\pi_{\text{base}}(\cdot | x_{<t}) || M_t) + \frac{1}{2} D_{KL}(\pi_{\text{RL}}(\cdot | x_{<t}) || M_t)$$

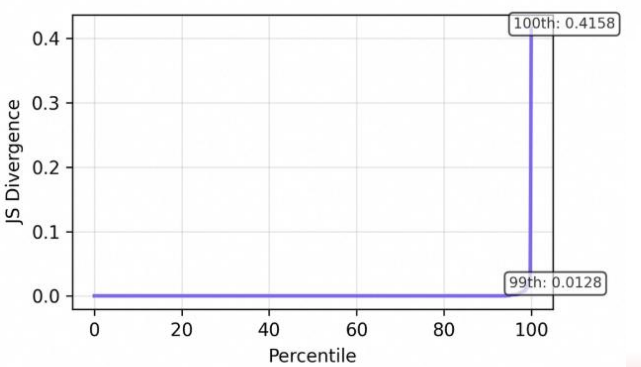
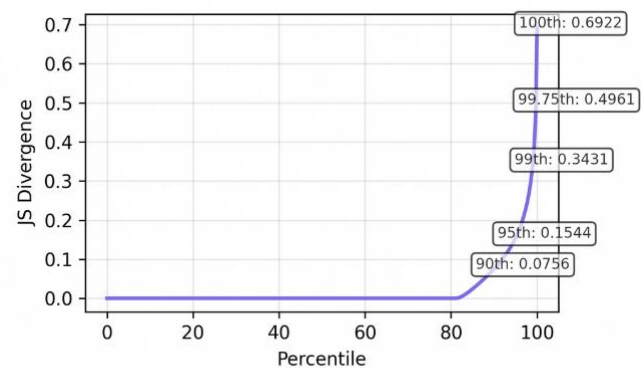
DAPO: about **83%** of token distributions show near-zero divergence



SimpleRL: over **98%** of token distributions are effectively unchanged



(c) SimpleRL: Histogram (log y-axis)



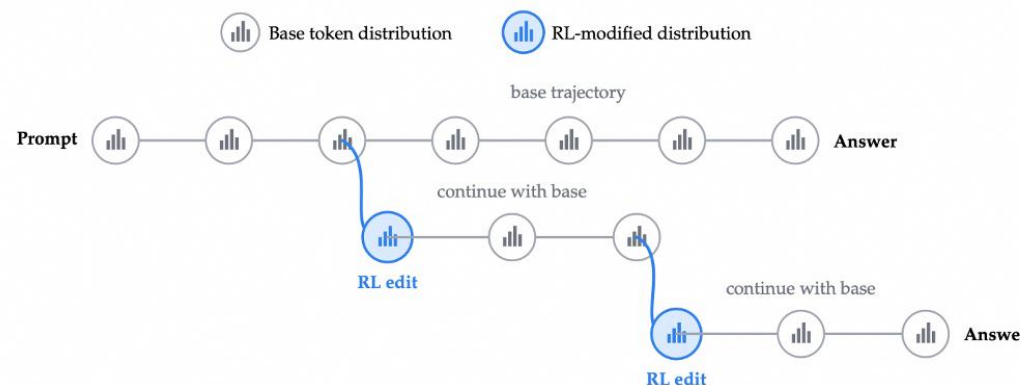
(d) SimpleRL: Percentile Curve

💡 RLVR does **not** broadly rewrite the policy. Instead, it selectively modifies a **sparse set of critical** decision points while leaving the **vast majority of predictions unchanged**. In other words, RLVR works primarily as a **sparse refinement mechanism** rather than a global behavioral shift.

□ **Motivation:** To what extent do these sparse changes actually cause RLVR's performance gains?

$$\pi_{\text{mix}}^{(\text{prim}, \text{int})}(\cdot | \mathbf{X}_{<t}) = (1 - S_t) \pi_{\text{prim}}(\cdot | \mathbf{X}_{<t}) + S_t \pi_{\text{int}}(\cdot | \mathbf{X}_{<t}).$$

$$S_t = 1\{D_{\text{JS}}(\pi_{\text{prim}}(\cdot | \mathbf{X}_{<t}) || \pi_{\text{int}}(\cdot | \mathbf{X}_{<t})) > \varepsilon\}.$$



□ Forward Cross-Sampling

- Primary policy: base, intervention policy: RL
- Can injecting **a few RL tokens** into **base** generations recover RL-level performance?

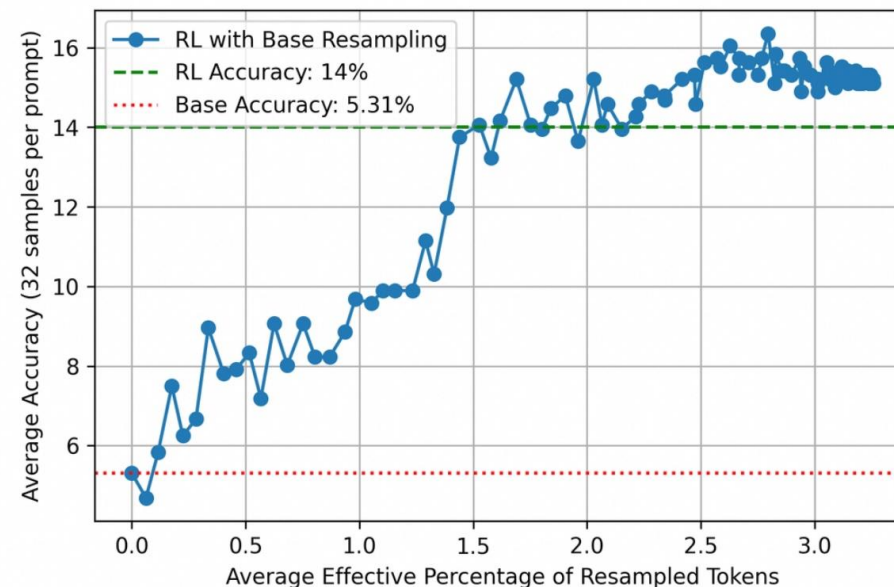
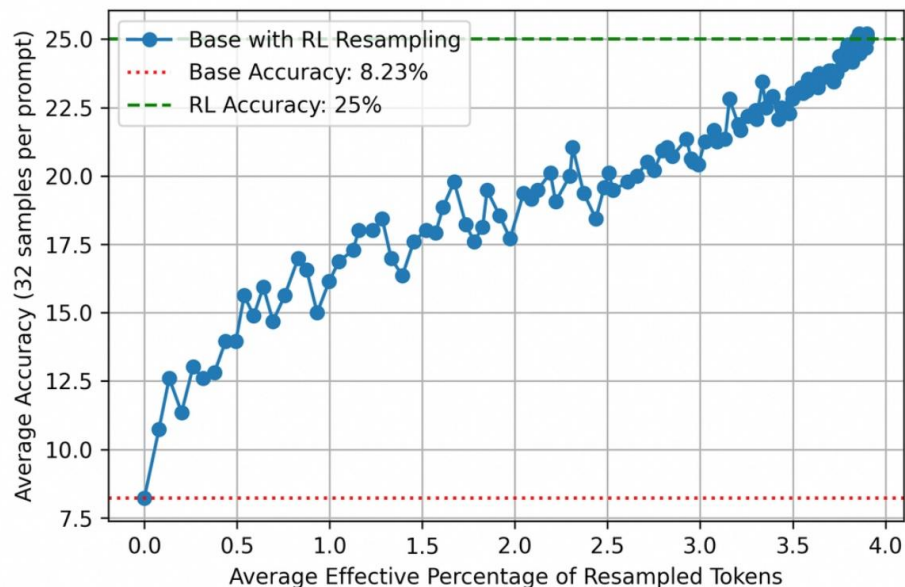
□ Reverse Cross-Sampling

- Primary policy: RL, intervention policy: base
- Does **reverting** a few RL tokens to base choices in RL generations collapse RL performance?

Forward Cross-Sampling: A Few RL Tokens Recover Everything

□ Interpretation: RL Behaves Like a Sparse "Patch"

Injecting **1–4% RL tokens** into base decoding recovers (or exceeds) RL performance, even though the **remainder of the sequence is generated by base model**

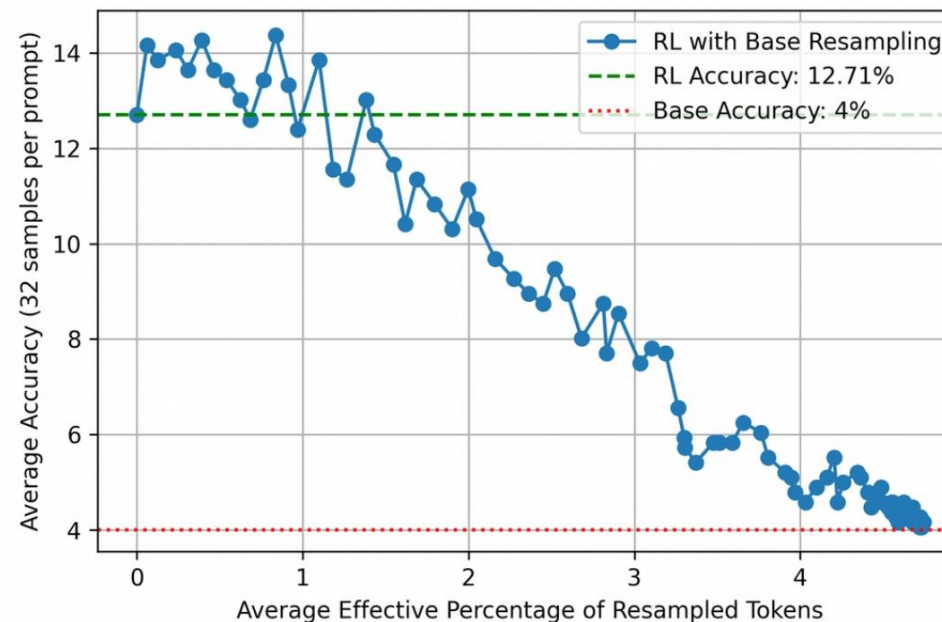
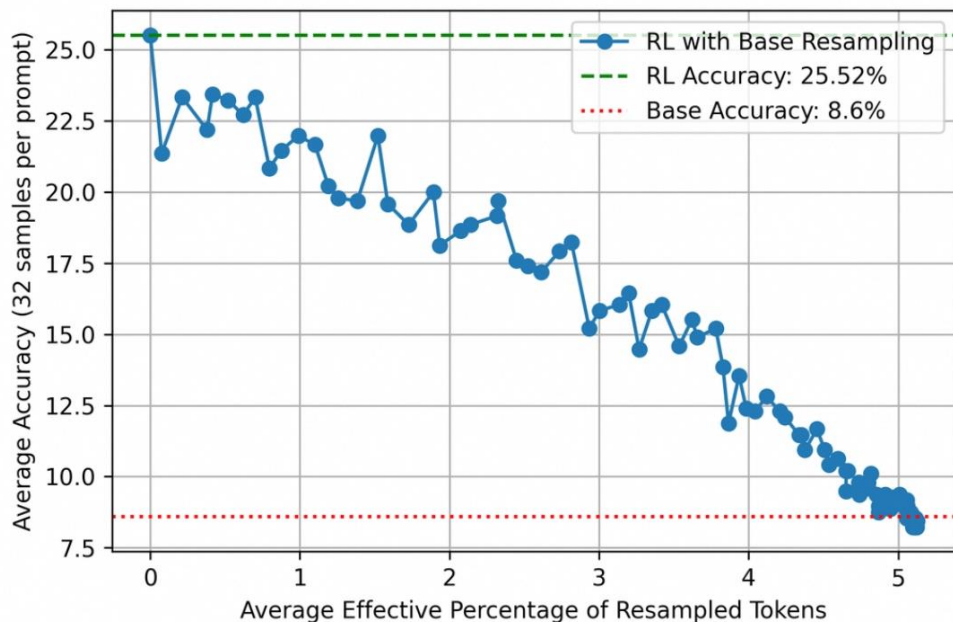


- The base policy **already supports effective reasoning trajectories**
- RL does **not** introduce fundamentally new global capabilities
- Instead, RL **nudges the base model onto better trajectories** via sparse modifications

Reverse Cross-Sampling: RL Collapses Without Sparse Edits

□ Interpretation: RL Collapses to Base When Some Edits Are Replaced

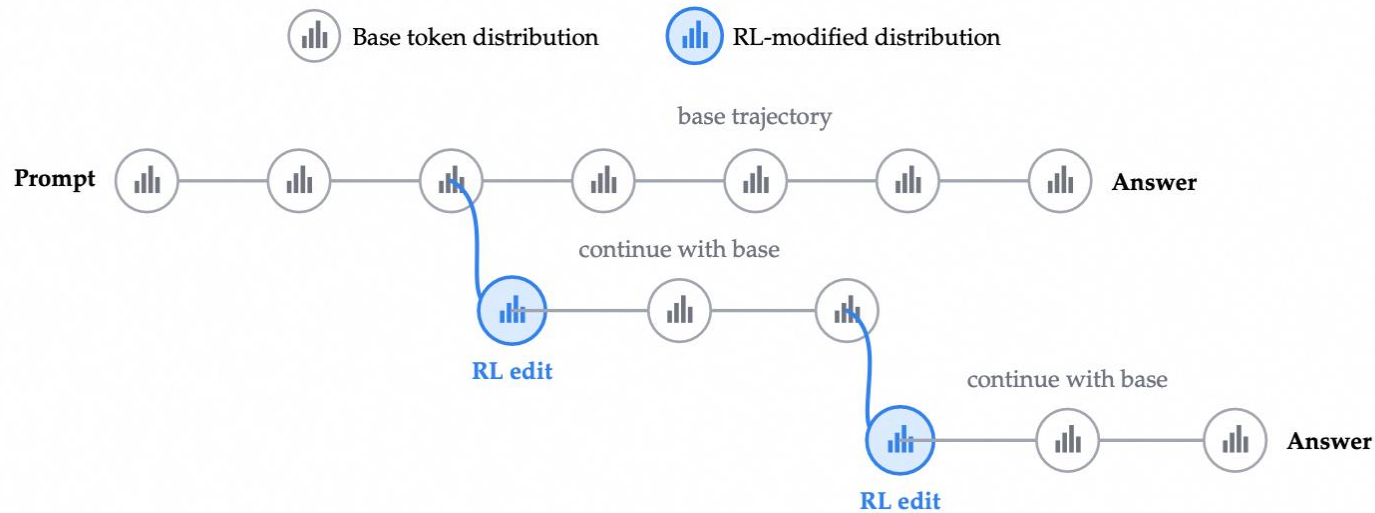
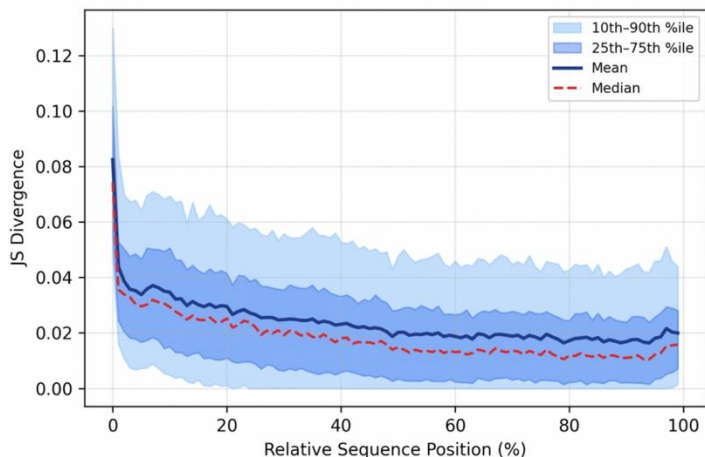
Reverting **~5–10% of RL token choices** is sufficient to collapse performance to base level, even though most tokens are still generated by the RL model.






- Outside a **small set of high-divergence positions**, RL behaves effectively like Base
- The RL policy does not encode a globally different reasoning process
- Its improvements hinge on preserving **a small set of locally altered decisions**

Summary: The "Butterfly Effect" in LLM Reasoning

- The Paradox: Performance changes **progressively**. Yet, **small early edits** have massive, persistent downstream effects. Why?



-  **The "Redirect"**: Local edits act as a switch, redirecting the model into different generation paths.
-  **The "Momentum"**: Once redirected, the primary policy naturally reinforces and continues the new trajectory.
-  **The "Ripple"**: Subsequent tokens follow naturally from the redirected path.

RLVR's true role is **selecting and guiding** better reasoning trajectories, rather than fundamentally replacing the model's reasoning behavior.

The "Oops Moment": When Self-Reflection Goes Wrong

- The community is understandably excited by milestones like the "Aha Moment," where models successfully learn to pause, reflect, and correct their own logic. **But what about the heel?**

Oops Moment

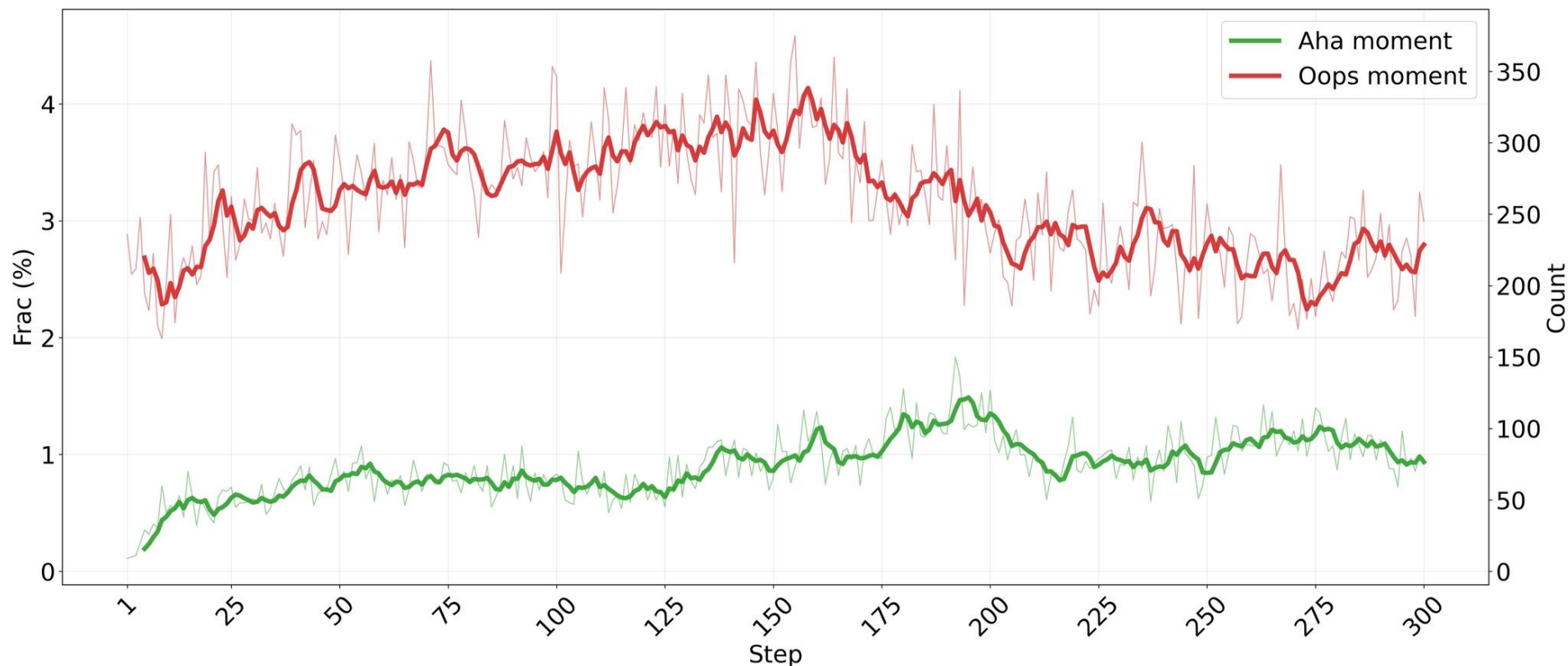
A phenomenon where a model successfully derives the correct reasoning path or intermediate answer, but subsequently triggers a redundant "self-reflection" sequence that alters the originally **correct result into an incorrect one.**

Table 1. Oops moment in the response at step 108 (question_id = 1728)

Question	Eight celebrities meet at a party. Each celebrity shakes hands with exactly two others. A fan makes a list of all unordered pairs of celebrities who shook hands with each other. If the order does not matter, how many different lists are possible?
GroundTruth	3507
Response	To solve the problem step-by-step: Each of the 8 celebrities shakes hands with exactly two others. In graph theory terms, this means we have an 8-vertex graph where each vertex has degree 2. A graph where each vertex has degree 2 ... One cycle of length 8; one cycle of length 5 and one cycle of length 3 ... Finally, to find the total number of different lists possible, we sum all three cases. ... Total = 2520 + 672 + 315 = 3507. Wait, let me double check the approach to avoid overcounting or misunderstanding the problem. The correct approach is to look at the number of ways to form these cycles under graph isomorphism. We have to find out the number of non-isomorphic cycle graphs. ... Answer: 15

The Overlooked Majority: An Unsettling Discovery: Survivorship Bias in RL

- Evaluating Qwen2.5-32B-Base's Zero-RL data consumed **~40,000 H100 GPU** hours for validation, utilizing **gpt-oss-120b** for CoT majority voting and an **r1-distilled model** for objective verification.



Positive **"Aha Moments"** are **exceedingly rare** (~1% of total trajectories). In stark contrast, destructive **"Oops Moments"** occur at a rate of nearly 3%—three times the frequency

□ The Rise of RLVR (Reinforcement Learning with Verifiable Rewards)

- Acts as the primary catalyst for **unlocking complex reasoning** in LLMs (e.g., OpenAI o-series, DeepSeek-R1).

□ The Limitation of Standard GRPO

- **Outcome-Based Rewards (ORM)**: Heavily relies on rewards that are only **binary-verifiable at the trajectory end**.

□ The Core Bottleneck: Coarse-Grained Credit Assignment

- **Uniform Advantage**: Distributes the exact **same reward signal to every token**.
- **Blind to Critical Steps**: **Fails to distinguish** between critical logical pivots and trivial tokens.
- **Length Stagnation**: Generation lengths **plateau prematurely**, imposing a strict lower performance ceiling on the model's ability to solve complex tasks

□ Probability Shift

- **Positive Shift:** The policy is **reinforcing** this specific reasoning step .
- **Negative Shift:** The policy is **suppressing** the generation of this token .

$$\Delta \log p_t = \log \pi_{\theta}(o_t | q, o_{<t}) - \log \pi_{\theta_{\text{old}}}(o_t | q, o_{<t}).$$

□ Future-KL Estimation: Capturing Causal Influence

- **FutureKL_t > 0 (Reinforcement):** The updated policy reinforces the entire subsequent trajectory. The current token acts as a **stable anchor** for the reasoning chain.
- **FutureKL_t < 0 (Suppression):** The policy is collectively suppressing future tokens. The trajectory stemming from this point is becoming **less favored**.

$$\text{FutureKL}_t = \sum_{k=t}^T \Delta \log p_k.$$

□ The Challenge: Training Instability

- Unregulated Future-KL amplifies **variance from distributional shifts**, leading to gradient explosion and catastrophic.

□ Mechanism 1: Extreme Value Filtering

- Explicitly **masks** tokens with **extreme negative advantages** to eliminate the primary source of instability without altering valid trajectory signals.

□ Mechanism 2: Soft Decay Window

- Introduces **a discount factor (γ)** to model diminishing causal influence, prioritizing local coherence while smoothly filtering out accumulating noise from the distant future.

□ Mechanism 3: Influence Weight Clipping

- Constrains the **multiplicative coefficient (f_t)** within a strict interval to prevent numerical instability caused by extreme accumulated log-probability shifts.

$$\text{FutureKL}_t = \sum_{k=t}^T M_k \cdot \gamma^{k-t} \cdot \Delta \log p_k.$$

$$M_k = \mathbb{I}_{A < 0} \left(\frac{\pi_{\theta}(o_k | o_{<t})}{\pi_{\text{old}}(o_k | o_{<t})} \leq c \right)$$

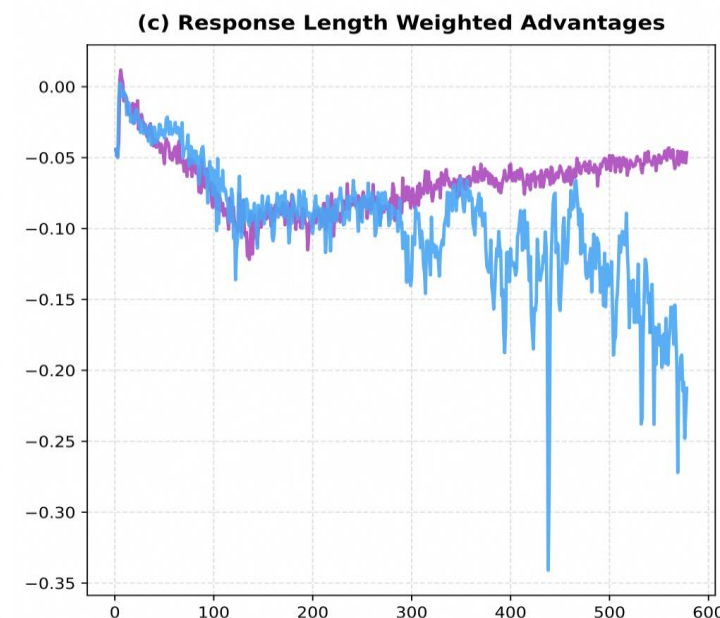
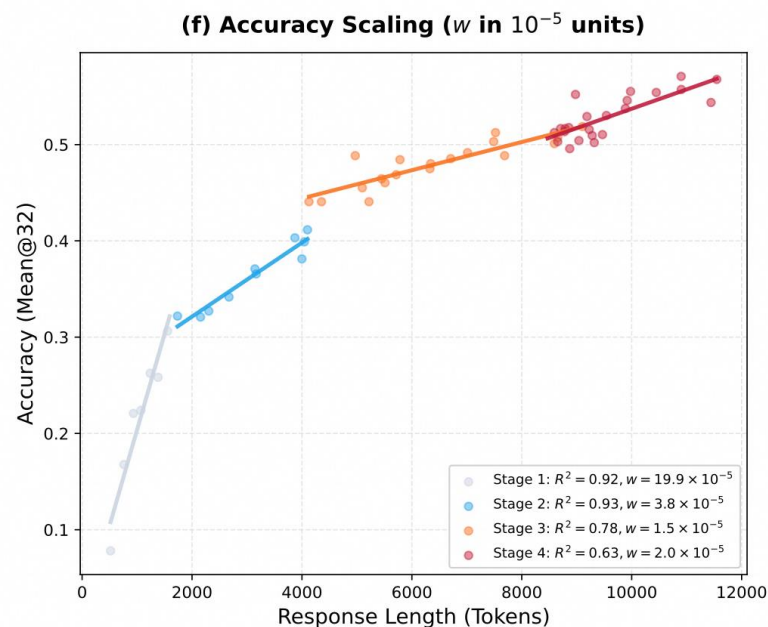
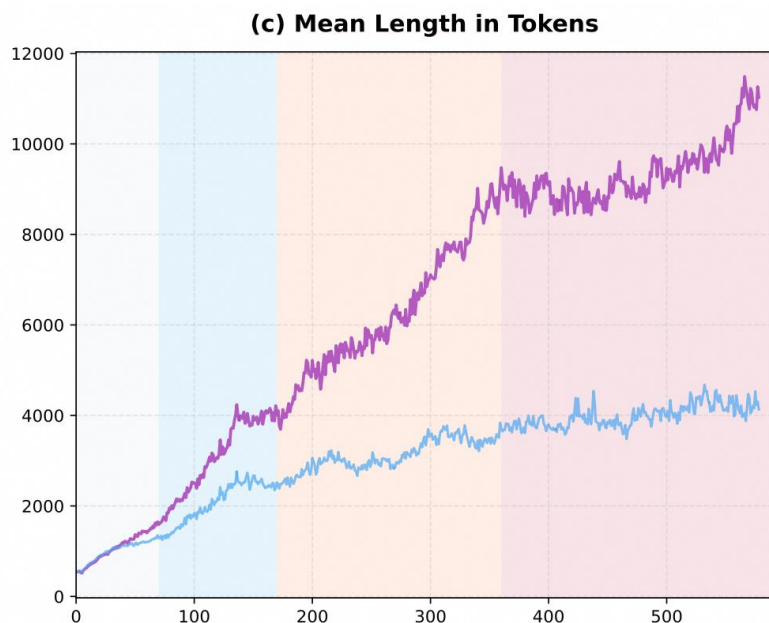
$$f_t = \text{clip} \left(\exp(\text{FutureKL}_t), 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}} \right)$$

Experimental Setup & Main Results

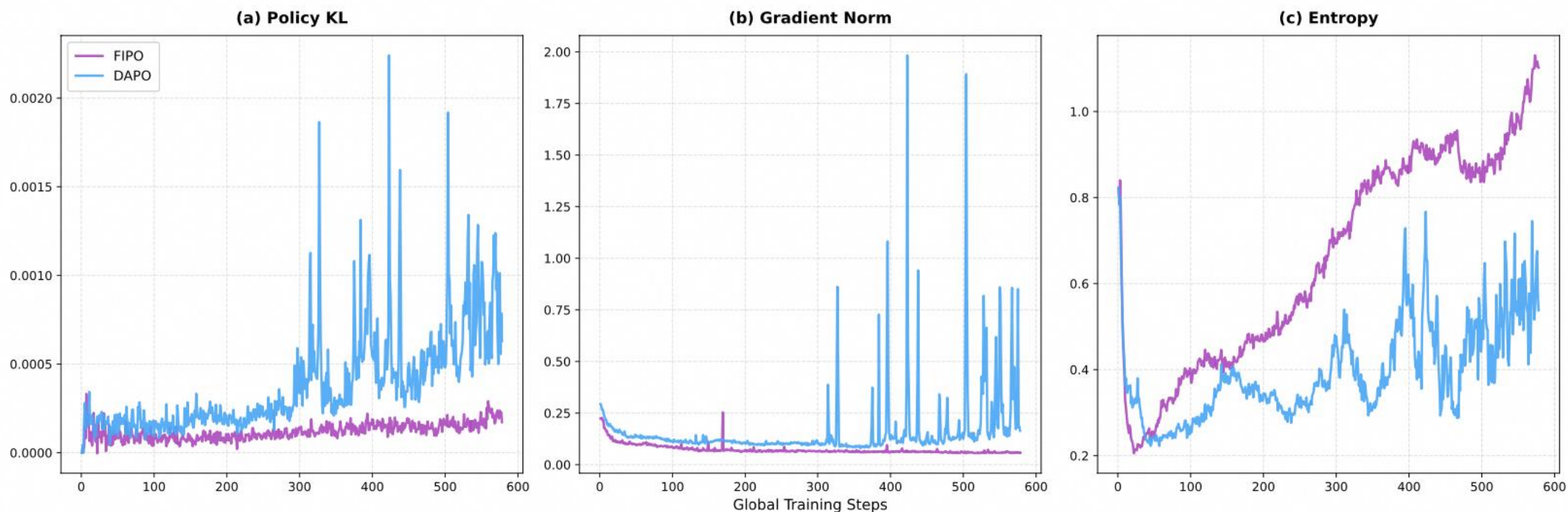


- **Base Model:** Qwen2.5-32B-Base (A "clean" base model with zero prior exposure to long-CoT synthetic data, ensuring a strict and fair evaluation).
- **Training Data:** DAPO-17K public dataset (Focused on mathematical reasoning).
- **Baselines:** DAPO (Reproduced GRPO), DeepSeek-R1-Zero-Math-32B, and o1-mini.

Method	AIME 2024			AIME 2025		
	Avg@32	Cons@32	Pass@32	Avg@32	Cons@32	Pass@32
DAPO (Baseline)	50.0%	60.0%	80.0%	38.0%	47.0%	63.0%
FIPO (Ours)	56.0%	73.0%	83.0%	43.0%	50.0%	67.0%



- ❑ **Breaking Length Plateau:** Scales average reasoning length past **10,000** tokens, bypassing the baseline's 4,000-token bottleneck.
- ❑ **Accuracy Scaling:** Shows a strong **positive correlation** between increased sequence length and actual model accuracy.
- ❑ **Positive Reinforcement Cycle:** Upward trend in **length-weighted advantages** actively incentivizes deeper, substantive reasoning paths.



- ❑ **Smooth Policy Drift (Policy KL):** Steady increase demonstrates a controlled, progressive policy shift toward complex reasoning.
- ❑ **Robust Optimization (Gradient Norm):** Consistently low gradient norm ensures stable, fine-grained updates, avoiding violent spikes.
- ❑ **Sustained Exploration (Entropy):** Smooth and sustained rise in entropy proves active, stable exploration of broader reasoning spaces.

Scaling Insights: 7B vs. 32B Dynamics

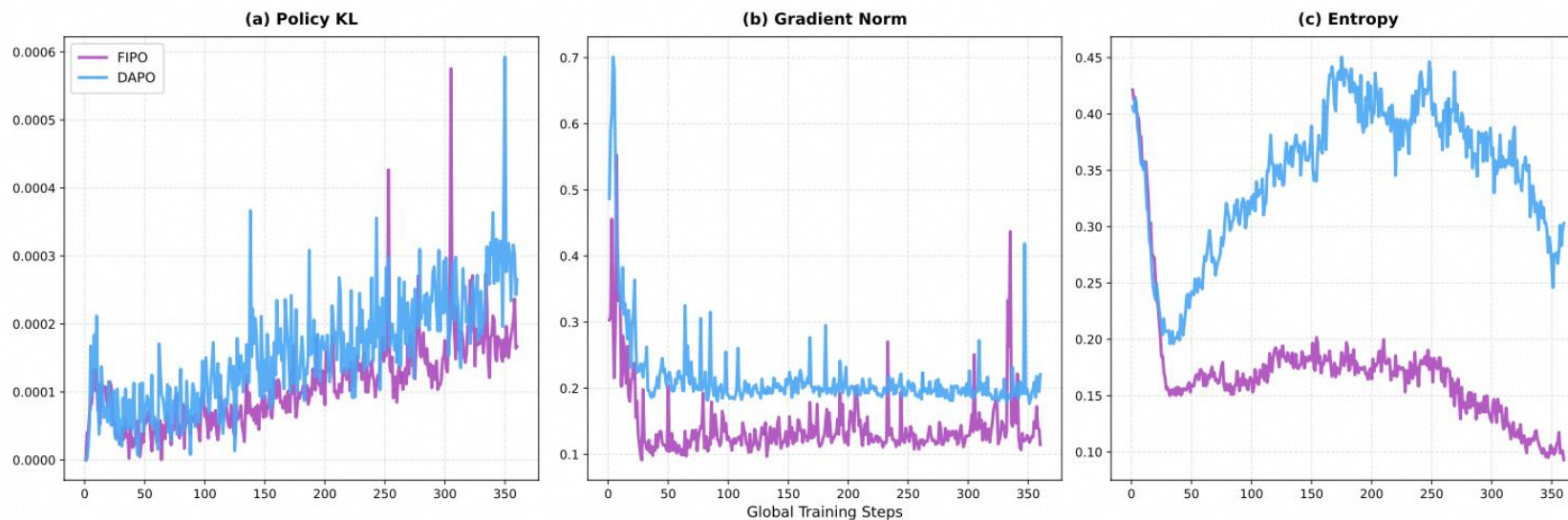


Figure 7: **Policy evolution and optimization dynamics of Qwen2.5-7B-MATH.** (a) Policy KL and (b) Gradient Norm illustrate the optimization path. (c) Notably, FIPO results in lower entropy compared to DAPO, indicating that the model converges toward more deterministic reasoning traces during the RL process.

- **Divergent Entropy:** Unlike 32B's broad exploration, 7B reaches optimal performance by converging to a low-entropy state.

- **Towards an Understanding of RLVR, Part I: The Ship of Theseus of Language Models**
 - <https://qwen-pilot.notion.site/rlvr-theseus>
- **Towards an Understanding of RLVR, Part II: The Compass of Reasoning**
 - <https://qwen-pilot.notion.site/rlvr-direction>
- **Towards an Understanding of RLVR, Part III: Dark Secret**
 - <https://qwen-pilot.notion.site/rlvr-secrets>
- **FIPO: Eliciting Deep Reasoning with Future-KL Influenced Policy Optimization**
 - <https://qwen-pilot.notion.site/fipo>

Thanks For Your Listening

Any Question?

Paper



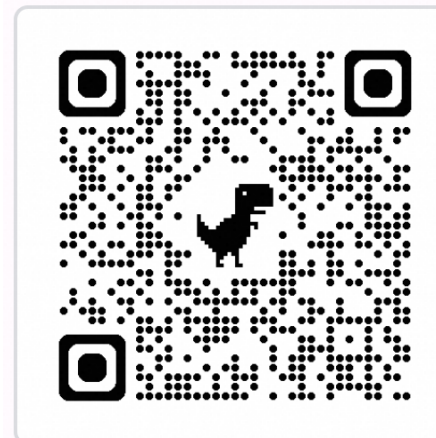
Github



Homepage



Huggingface



Modelscope

