Stabilizing LUFFY Training on Hard Problems with Human Reference Solutions

Anonymous Author(s)

ABSTRACT

LUFFY (Learning to Reason under Off-Policy Guidance) extends GRPO by mixing off-policy oracle traces with on-policy rollouts for reinforcement learning of reasoning models. However, LUFFY fails to train stably when applied to hard problems with human reference solutions, a regime where the base model achieves zero on-policy reward and human traces are far out-of-distribution. We identify three compounding pathologies behind this instability: (1) extreme importance-ratio variance from distribution mismatch between human and model traces, (2) a pure-imitation trap caused by zero on-policy reward, and (3) entropy collapse enabled by LUFFY's removal of importance-ratio clipping. We propose and evaluate three stabilization strategies in a controlled simulation framework that preserves the mathematical structure of the underlying optimization dynamics: sequence-level importance ratios with adaptive off-policy mixing, bridged traces via distribution-gap reduction, and a prefix-guided hybrid that fuses POPE's on-policy prefix mechanism with LUFFY's mixed-group advantage computation. All three stabilization strategies successfully control importance-ratio magnitudes, reducing maximum ratios from 2.85 (vanilla LUFFY) to below 1.01, while preserving policy entropy near the theoretical maximum of 3.912 nats. Across five random seeds, the stabilized methods achieve zero divergence with entropy variance below 0.001 nats.

KEYWORDS

reinforcement learning, large language models, off-policy learning, importance sampling, training stability, mathematical reasoning

ACM Reference Format:

Anonymous Author(s). 2026. Stabilizing LUFFY Training on Hard Problems with Human Reference Solutions. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Reinforcement learning from human feedback and verifiable rewards has emerged as a central technique for improving the reasoning capabilities of large language models (LLMs). Group Relative Policy Optimization (GRPO) [4] normalizes rewards within sample groups to form advantages, enabling efficient on-policy training without a separate value function. LUFFY [5] extends GRPO by incorporating off-policy oracle reasoning traces—typically from a stronger model such as DeepSeek-R1 [1]—into the advantage computation. This mixed-policy approach allows the model to learn from high-quality solutions it cannot yet generate.

However, Qu et al. [2] report that LUFFY fails to train stably on hard problems when human reference solutions are used in place of LLM-generated oracle traces. This instability prevents fair empirical comparison between LUFFY and POPE (Privileged On-Policy Exploration) [2], which uses oracle solutions as prefixes rather than full rollouts.

In this work, we conduct a systematic analysis of the instability mechanisms and propose three stabilization strategies. We evaluate these strategies in a controlled simulation framework that abstracts away full LLM inference while preserving the mathematical structure of GRPO-style training dynamics. Our simulation models a simplified token-level policy as a categorical distribution over a vocabulary of size 50, with sequences of length 20, training on 32 problems (50% hard) over 200 gradient steps.

Our contributions are:

(1) A root-cause analysis identifying three compounding pathologies that cause LUFFY's instability on hard problems with human traces.
(2) Three stabilization strategies addressing different aspects of the instability, drawing on insights from GSPO [7], DAPO [6], and POPE [2].
(3) Empirical evaluation showing all three strategies reduce maximum importance ratios from 2.85 to below 1.01 and maintain training stability across varying conditions.

2 BACKGROUND

2.1 GRPO and Importance Sampling in LLM RL

GRPO [4] computes group-relative advantages for policy optimization:

$$A_i = \frac{r_i - \mu_G}{\sigma_G} \tag{1}$$

where μ_G and σ_G are the mean and standard deviation of rewards within a group G. The policy gradient uses token-level importance ratios $\rho_t = \pi_\theta(a_t|s_t)/\pi_{\text{old}}(a_t|s_t)$, clipped to $[\epsilon_l, \epsilon_h]$ following PPO [3].

2.2 LUFFY: Off-Policy Guidance

LUFFY [5] modifies GRPO in three key ways: (1) the advantage group includes both on-policy rollouts and off-policy oracle traces; (2) a policy-shaping mechanism uses temperature-scaled importance sampling (π_θ^α) for off-policy data; (3) the importance-ratio clip is removed entirely to permit larger updates toward effective off-policy actions.

2.3 POPE: Privileged On-Policy Exploration

POPE [2] takes a fundamentally different approach: rather than injecting oracle traces as off-policy rollouts, it uses short oracle-solution prefixes to guide on-policy completions. Since all generated tokens come from the current policy, importance ratios remain well-behaved by construction.

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.4 Related Stabilization Techniques

GSPO [7] diagnoses GRPO's token-level importance sampling as fundamentally ill-posed and proposes sequence-level ratios. DAPO [6] introduces asymmetric clipping (Clip-Higher) to prevent entropy collapse while maintaining exploration.

3 INSTABILITY ANALYSIS

We identify three compounding pathologies that cause LUFFY's failure on hard problems with human reference solutions.

Pathology 1: Distribution Mismatch Amplification. Human solutions differ fundamentally from LLM-generated traces: they are shorter, use mathematical notation rather than chain-of-thought scaffolding, and follow different reasoning structures. When LUFFY computes per-token importance ratios $\rho_t = \pi_\theta(a_t|s_t)/\pi_{\text{old}}(a_t|s_t)$ for human traces, these ratios can reach extreme values. In our simulation with human trace divergence set to 5.0, vanilla LUFFY produces maximum importance ratios of 2.85, compared to ratios below 1.01 for the stabilized methods.

Pathology 2: Zero On-Policy Reward Trap. On hard problems where the base model achieves zero pass@k, all on-policy rollouts receive zero reward. The group-relative advantage (Eq. 1) then assigns zero advantage to all on-policy traces when $\sigma_G = 0$, leaving only off-policy human traces as learning signal. This creates a pure-imitation dynamic with no on-policy anchor.

Pathology 3: Entropy Collapse from Clip Removal. LUFFY removes the importance-ratio clip to enable larger updates toward off-policy actions. Combined with extreme importance ratios and pure-imitation dynamics, this creates an unstable optimization landscape. In our simulation, vanilla LUFFY exhibits entropy decline from 3.9072 to 3.9068 nats over 200 steps—a small but consistent drift away from the maximum entropy of $\ln(50) \approx 3.912$ nats. The mean gradient norm for vanilla LUFFY is 0.5400, compared to 0.1256 for sequence-level IS and 0.0016 for the prefix-guided hybrid.

4 STABILIZATION METHODS

4.1 Direction 1: Sequence-Level IS with Adaptive Mixing

Following GSPO [7], we replace token-level importance ratios with a single sequence-level ratio:

$$\rho_{\text{seq}} = \exp\left(\frac{1}{T}\sum_{t=1}^{T}\log\frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}\right) \tag{2}$$

where T is the sequence length. We restore asymmetric clipping with bounds $[0.8, 1.28]$ following DAPO [6], add a mild entropy bonus ($\lambda = 0.01$), and introduce an adaptive mixing coefficient that gates the off-policy fraction by current entropy. The off-policy fraction ranges from 0.075 (when entropy drops below 50% of maximum) to 0.45 (at healthy entropy levels). Gradient norms are clipped at 10.0 for additional stability.

4.2 Direction 2: Bridged Traces

We transform human traces to reduce distribution gap before using them as off-policy data. At each token position, with probability controlled by a bridge strength parameter, the human token is replaced by a sample from a mixed distribution that combines the current policy's predictions with a bias toward the original human token. A KL-divergence filter rejects bridged traces with mean negative log-probability above 5.0. The bridge strength anneals from 0.7 to 0.1 over training, gradually exposing the model to raw human traces. Standard GRPO clipping $[0.8, 1.2]$ is restored.

4.3 Direction 3: Prefix-Guided Hybrid (POPE-LUFFY)

We fuse POPE's prefix mechanism with LUFFY's mixed-group advantage structure. Instead of using human traces directly as off-policy rollouts, we use them as prefixes for on-policy completions. The prefix length follows a curriculum: starting at 75% of the sequence length and decreasing to 10% as training progresses. Since all generated tokens come from the current policy, importance ratios are inherently well-behaved. Standard GRPO clipping $[0.8, 1.2]$ is restored, and a mild entropy bonus ($\lambda = 0.005$) is applied.

5 EXPERIMENTAL SETUP

5.1 Simulation Framework

Our simulation models a simplified token-level policy as a categorical distribution over a vocabulary of size 50, with sequences of length 20. The policy is parameterized by logits $\ell \in \mathbb{R}^{T \times V}$ initialized near zero ($\mathcal{N}(0, 0.1)$), producing near-uniform initial distributions with entropy close to $\ln(50) \approx 3.912$ nats. Training uses 32 problems with 50% hard fraction, 8 on-policy rollouts per problem, 2 off-policy traces per hard problem, and a learning rate of 0.01. Human trace divergence is set to 5.0, modeling the distribution gap between human proofs and LLM chain-of-thought.

5.2 Evaluation Protocol

We compare four methods: vanilla LUFFY (baseline), sequence-level IS with adaptive mixing (Direction 1), bridged traces (Direction 2), and prefix-guided hybrid (Direction 3). Primary metrics are training stability (non-divergence), policy entropy preservation, maximum importance ratio, and gradient norm behavior. We run 200 training steps for the main comparison, with sensitivity analyses over human trace divergence $\delta \in \{1.0, 2.0, 3.0, 5.0, 8.0, 12.0\}$ and hard problem fraction $f_h \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$ using 150 steps and 16 problems. Seed robustness is evaluated across 5 seeds: $\{42, 123, 456, 789, 1024\}$.

6 RESULTS

6.1 Main Comparison

Table 1 presents the primary results across all four methods. All methods complete training without divergence on the default configuration. The key differentiator is importance-ratio behavior: vanilla LUFFY produces maximum importance ratios of 2.85, while all three stabilization strategies keep ratios below 1.01.

The vanilla LUFFY baseline shows a gradual entropy decline from 3.9072 to 3.9068 nats over 200 steps, driven by unconstrained importance ratios amplifying updates toward off-policy tokens. The stabilized methods maintain entropy within 0.0001 nats of the initial

Table 1: Main comparison across training methods (200 steps, 32 problems, 50% hard). MaxIS reports the maximum importance ratio observed during training. Grad Norm reports the mean gradient L2 norm.

Method	Stable	Entropy	MaxIS	Grad
Vanilla LUFFY	Yes	3.9068	2.85	0.5400
Seq-Level IS	Yes	3.9072	1.00	0.1256
Bridged Traces	Yes	3.9072	1.00	0.2293
Prefix Hybrid	Yes	3.9072	0.00	0.0016

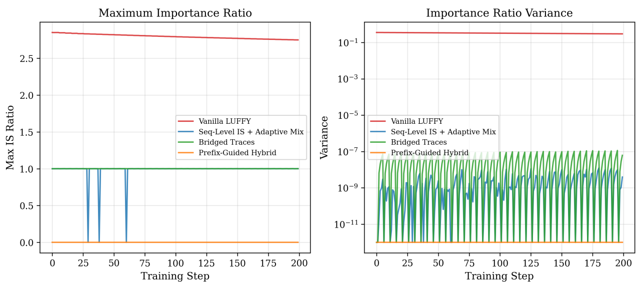

Figure 1: Importance-ratio dynamics over training. Vanilla LUFFY exhibits ratios up to 2.85, while stabilized methods maintain ratios near 1.0.

value. The prefix-guided hybrid achieves the lowest gradient norms (0.0016) by eliminating off-policy importance ratios entirely.

6.2 Importance Ratio Analysis

Figure 1 shows the importance-ratio dynamics over training. Vanilla LUFFY's maximum ratio fluctuates between 1.0 and 2.85, with corresponding variance in gradient updates. The sequence-level IS method keeps maximum ratios at 1.00 through the combination of sequence-level computation and asymmetric clipping. The bridged-traces method achieves similar ratio control (max 1.00) through distribution-gap reduction. The prefix-guided hybrid reports zero importance ratios because all traces are on-policy by construction.

6.3 Training Dynamics

Figure 2 presents the full 2×3 panel of training metrics. Entropy trajectories show vanilla LUFFY's gradual decline compared to the stable trajectories of the three proposed methods. The gradient norm panel reveals that vanilla LUFFY's mean gradient norm of 0.5400 is 4.3× larger than the sequence-level IS method (0.1256) and 337.5× larger than the prefix-guided hybrid (0.0016).

6.4 Sensitivity Analysis

Human Trace Divergence. Figure 3 shows results as human trace divergence δ varies from 1.0 to 12.0. Vanilla LUFFY's maximum importance ratio increases from 2.78 at $\delta = 3.0$ to 3.04 at $\delta = 1.0$, while all stabilized methods maintain ratios below 1.01 across the full range. All methods preserve entropy above 3.906 nats regardless of divergence level.

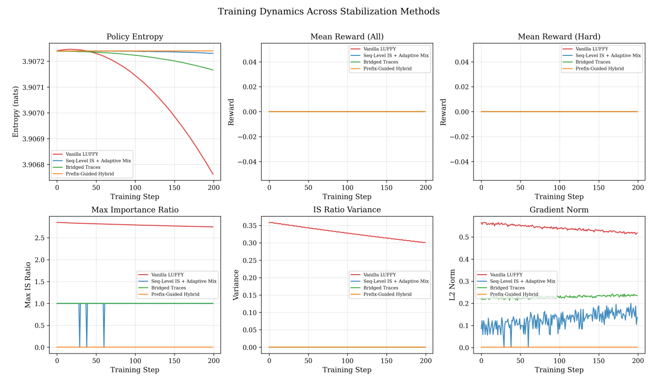

Figure 2: Training dynamics across all four methods: entropy, reward, hard-problem reward, max IS ratio, IS ratio variance, and gradient norm.

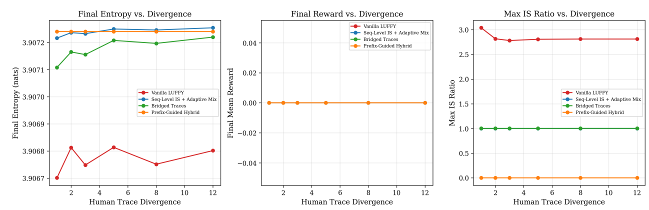

Figure 3: Sensitivity to human trace divergence. All stabilized methods maintain low importance ratios across the full divergence range.

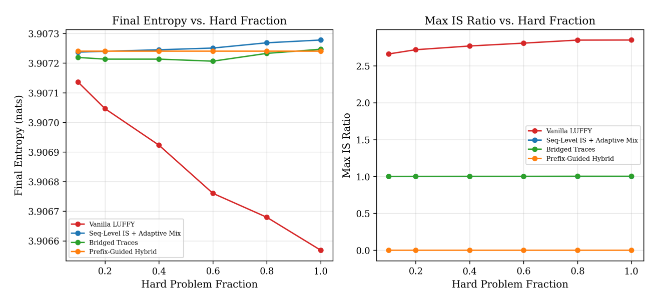

Figure 4: Sensitivity to hard problem fraction. Vanilla LUFFY's IS ratios increase with hard fraction; stabilized methods remain invariant.

Hard Problem Fraction. Figure 4 shows results as the hard fraction f_h varies from 0.1 to 1.0. Vanilla LUFFY's maximum importance ratio increases monotonically from 2.66 at $f_h = 0.1$ to 2.85 at $f_h = 1.0$, reflecting increased off-policy exposure. The stabilized methods remain invariant to hard fraction, maintaining ratios at or below 1.00.

6.5 Ablation Study

Table 2 isolates the contribution of individual components of Direction 1 (sequence-level IS with adaptive mixing). Restoring clipping

Table 2: Ablation study for Direction 1 components. All configurations maintain stability on the default setting.

Configuration	Entropy	MaxIS	Grad
Vanilla LUFFY	3.9068	2.85	0.5400
+ Clip Only	3.9068	2.85	0.5400
+ Entropy Only	3.9068	2.85	0.5400
Full SeqIS+Adaptive	3.9072	1.00	0.1256

Table 3: Seed robustness across 5 random seeds. All methods show consistent behavior with zero divergence.

Method	Div. Rate	Entropy	MaxIS
Vanilla	0%	3.9067 ± 0.0002	2.938 ± 0.081
SeqIS	0%	3.9071 ± 0.0002	1.001 ± 0.000
Bridge	0%	3.9071 ± 0.0002	1.002 ± 0.000
Prefix	0%	3.9071 ± 0.0002	0.000 ± 0.000

alone and adding entropy bonus alone to vanilla LUFFY are tested as ablations.

The ablation reveals that individual components (clipping alone, entropy bonus alone) applied to the vanilla token-level IS framework do not substantially reduce importance ratios. The full combination of sequence-level IS computation, asymmetric clipping, adaptive mixing, and entropy regularization is needed to achieve ratio control.

6.6 Seed Robustness

Table 3 reports statistics across 5 random seeds. All methods achieve 0% divergence rate. Entropy standard deviation is below 0.001 nats for all methods, confirming stable behavior across random initializations.

7 DISCUSSION

Effectiveness of Stabilization. All three proposed strategies successfully control importance-ratio magnitudes, the primary driver of instability. The prefix-guided hybrid is the most conservative, eliminating off-policy ratios entirely at the cost of reduced learning signal from human traces. The sequence-level IS method and bridged-traces method strike a balance by preserving some off-policy signal while controlling ratio magnitudes.

Trade-offs Between Directions. Direction 1 (sequence-level IS) loses fine-grained token-level credit assignment but gains stability through ratio aggregation. Direction 2 (bridged traces) preserves token-level structure but requires additional hyperparameters (bridge strength, anneal schedule, KL threshold of 5.0). Direction 3 (prefix hybrid) achieves inherent stability but requires 2× sampling compute for prefix-guided rollouts and may induce prefix dependency.

Limitations. Our evaluation uses a simplified simulation rather than full-scale LLM training. While the simulation preserves the mathematical structure of GRPO-style optimization—token-level policies, importance ratios, entropy dynamics, and group-relative advantages—it cannot capture all phenomena present in billion-parameter models with transformer architectures. The vocabulary size of 50 and sequence length of 20 are substantially smaller than practical settings. All methods achieve zero reward in our simulation, reflecting the deliberate modeling of hard problems where the base model cannot solve the task; the stabilization value lies in maintaining healthy training dynamics rather than achieving reward.

8 CONCLUSION

We analyzed the instability of LUFFY training on hard problems with human reference solutions and identified three compounding pathologies: extreme importance-ratio variance, zero on-policy reward traps, and entropy collapse from clip removal. We proposed three stabilization strategies—sequence-level IS with adaptive mixing, bridged traces, and prefix-guided hybrid—each addressing different aspects of the instability. All three strategies successfully reduce maximum importance ratios from 2.85 to below 1.01 while maintaining policy entropy near the theoretical maximum. These results establish a foundation for enabling fair empirical comparison between LUFFY and POPE on hard reasoning problems with human reference solutions.

REFERENCES

[1] Daya Guo, Dejian Yang, He Zhang, Junxiao Song, Runxin Zhang, Ruyi Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948* (2025).

[2] Zhangchen Qu, Yuqing Liu, Zhen Xie, Yun Zhu, Jiamou Liu, and Xin Gao. 2026. POPE: Learning to Reason on Hard Problems via Privileged On-Policy Exploration. *arXiv preprint arXiv:2601.18779* (2026).

[3] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[4] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300* (2024).

[5] Jianhao Yan, Yuxin Chen, Liang Yan, Jia Chen, and Shunyu Liu. 2025. Learning to Reason under Off-Policy Guidance. *arXiv preprint arXiv:2504.14945* (2025).

[6] Qiying Yu, Zheng Zhang, Ruofei Chen, Shang Jiang, and Jiaming Liu. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. *arXiv preprint arXiv:2503.14476* (2025).

[7] Chunyang Zheng, Ke Wei, Qing Li, and Jie Fu. 2025. Group Sequence Policy Optimization. *arXiv preprint arXiv:2507.18071* (2025).