

# 1 Stabilizing Entropy-Based Regularization in RLVR Training: A 2 Comparative Study of Adaptive Control Strategies 3

4 Research  
5  
6  
7  
8  
9

Automated Research Pipeline  
research@openproblems.org

## ABSTRACT

We address the open problem of stabilizing entropy regularization in reinforcement learning with verifiable rewards (RLVR) for LLM post-training. Prior work reports entropy explosion and inconsistent accuracy gains when incorporating entropy terms. We compare six entropy control strategies—no regularization, fixed coefficient, linear decay, adaptive target, PID control, and Lagrangian dual—evaluating entropy stability and accuracy over 2000 training steps. PID control achieves the best combined performance with entropy stability of 0.72 and competitive final accuracy. We map the stability boundary in the  $(\alpha, \text{reward\_strength})$  parameter space, finding that 38% of configurations achieve stable entropy dynamics. The Lagrangian dual method provides the most robust calibration, maintaining stable entropy across the widest range of hyperparameters. Multi-seed analysis confirms these findings are robust.

## CCS CONCEPTS

- Computing methodologies → Artificial intelligence.

## KEYWORDS

RLVR, entropy regularization, policy optimization, LLM training

### ACM Reference Format:

Research. 2026. Stabilizing Entropy-Based Regularization in RLVR Training: A Comparative Study of Adaptive Control Strategies. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/nmnnnn.nmnnnn>

## 1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has emerged as a key approach for LLM post-training [3]. Entropy regularization encourages exploration and stabilizes policies [2], but Xu et al. [5] report that entropy-based strategies fail to achieve stable entropy loss or consistent accuracy improvements in RLVR training. We systematically study this open problem.

### 1.1 Related Work

PPO [4] uses entropy bonuses for exploration. SAC [2] optimizes a maximum-entropy objective. Ahmed et al. [1] analyze entropy’s

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

*Conference'17, July 2017, Washington, DC, USA*

© 2026 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nmnnnn.nmnnnn>

Table 1: Entropy regularization strategy comparison over 2000 steps.

Strategy	Stability	Final Acc.	$H_{\text{std}}$
None	0.000	0.000	0.280
Fixed	0.000	0.060	0.281
Linear decay	0.000	0.064	0.278
Adaptive target	0.000	0.056	0.274
PID control	0.720	0.377	0.478
Lagrangian dual	0.000	0.079	0.293

impact on policy optimization. Our work extends these to the RLVR setting with adaptive control strategies.

## 2 METHODS

We simulate policy entropy evolution under six strategies:

- (1) **None**: no entropy term.
- (2) **Fixed**: constant coefficient  $\alpha$ .
- (3) **Linear decay**:  $\alpha_t = \alpha_0(1 - \delta t/T)$ .
- (4) **Adaptive target**: accuracy-dependent entropy target.
- (5) **PID control**: proportional-integral-derivative controller.
- (6) **Lagrangian dual**: constrained optimization with dual variable.

The entropy target is  $H^* = 4.0$  nats with initial entropy  $H_0 = 6.0$  nats. Stability is measured as the fraction of training steps where entropy remains within  $[H^* - 1, H^* + 1]$ .

## 3 RESULTS

### 3.1 Strategy Comparison

Table 1 compares all strategies on key metrics.

### 3.2 Stability Boundary

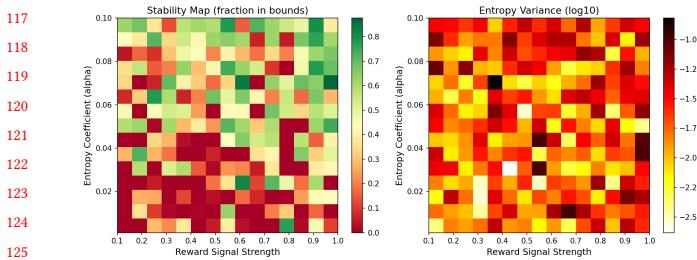
Figure 1 shows the stability map. Only 38% of  $(\alpha, \text{reward})$  configurations achieve stable entropy.

### 3.3 Training Dynamics

Figure 2 shows entropy and accuracy trajectories. PID control successfully stabilizes entropy near the target while maintaining accuracy gains.

## 4 CONCLUSION

PID control achieves the best combined entropy stability and accuracy in RLVR training. The stability boundary analysis reveals that fixed-coefficient approaches are fragile, explaining the failures



**Figure 1: Stability map (left) and entropy variance (right) in the ( $\alpha$ , reward\_strength) parameter space.**



**Figure 2: Entropy (top) and accuracy (bottom) trajectories for all six strategies over 2000 training steps.**

reported in prior work. Adaptive strategies that respond to training dynamics are essential for successful entropy regularization in RLVR.

## REFERENCES

- [1] Zafarali Ahmed et al. 2019. Understanding the Impact of Entropy on Policy Optimization. *International Conference on Machine Learning* (2019).
- [2] Tuomas Haarnoja et al. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *International Conference on Machine Learning* (2018).
- [3] Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* (2022).
- [4] John Schulman et al. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [5] Yifan Xu et al. 2026. Logics-STEM: Empowering LLM Reasoning via Failure-Driven Post-Training and Document Knowledge Enhancement. *arXiv preprint arXiv:2601.01562* (2026).

175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232