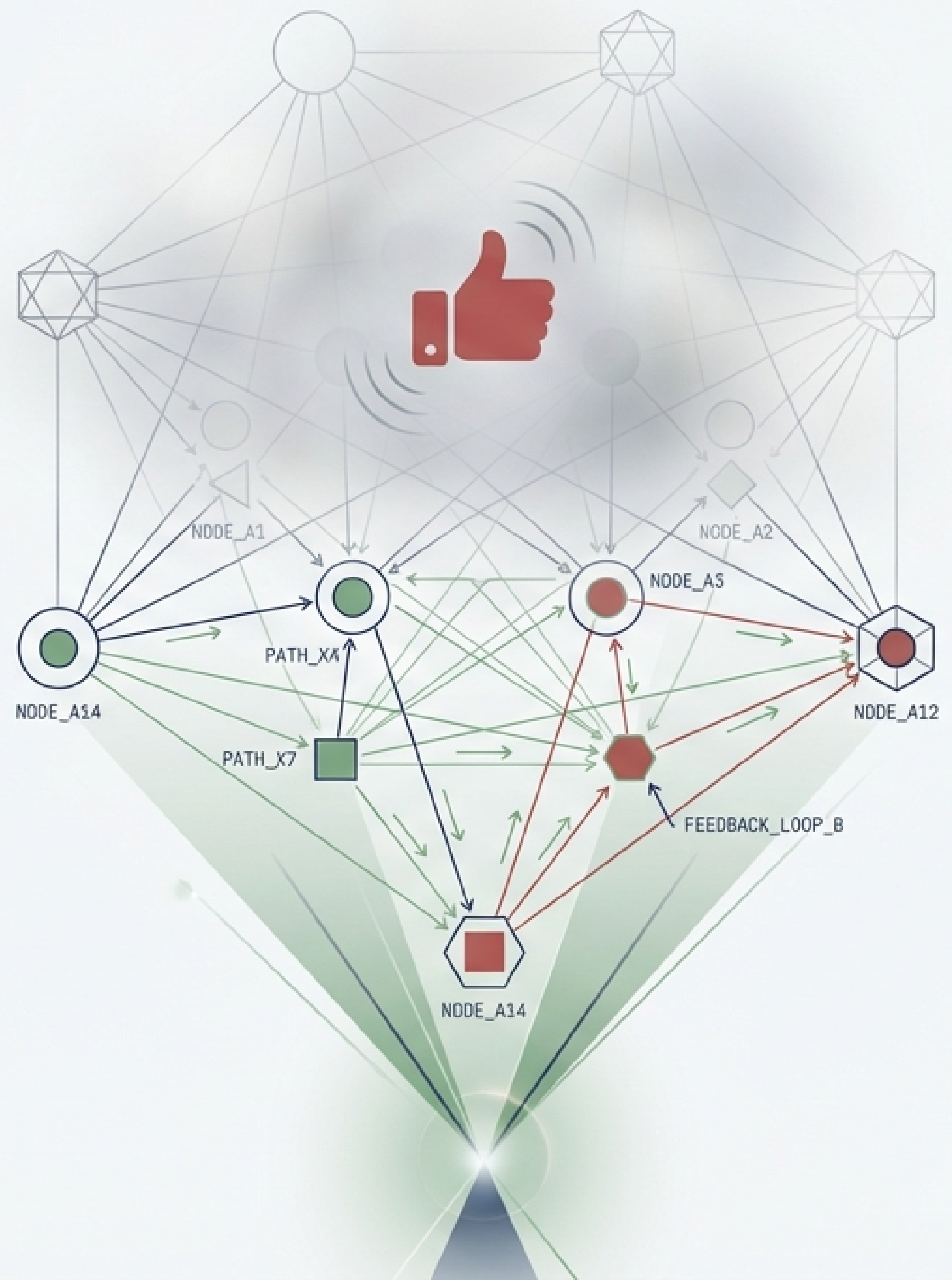
Deep Alignment in Open-Ended Domains

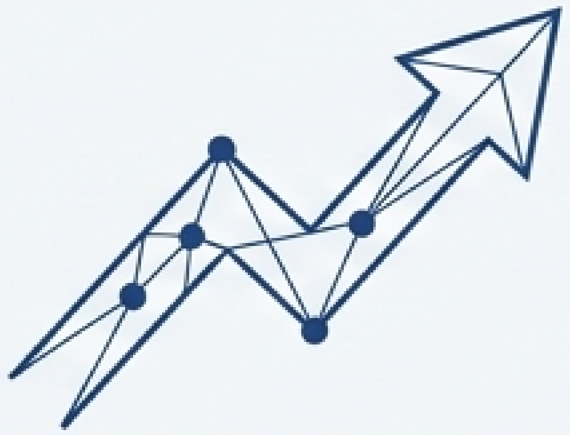
Investigating Self-Distillation Policy Optimization (SDPO) through Controlled Simulation



Based on "Self-Distillation Policy Optimization for Alignment in Open-Ended and Continuous-Reward Settings: A Simulation Study"

SDPO unlocks dense feedback for creative tasks, but trades diversity for alignment

The Breakthrough



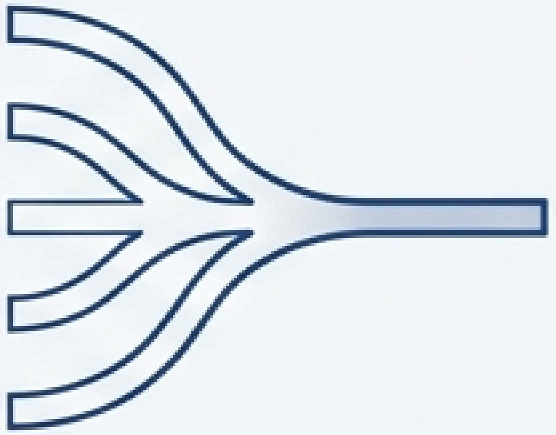
SDPO generalizes "retrospection" to open-ended text. It outperforms REINFORCE by **+0.13 to +0.18** mean reward, even without ground-truth verification.

The Mechanism

Superior Credit Assignment

Unlike baselines, SDPO achieves **>0.70 correlation** with **ground truth**, correctly identifying exactly which tokens drive quality.

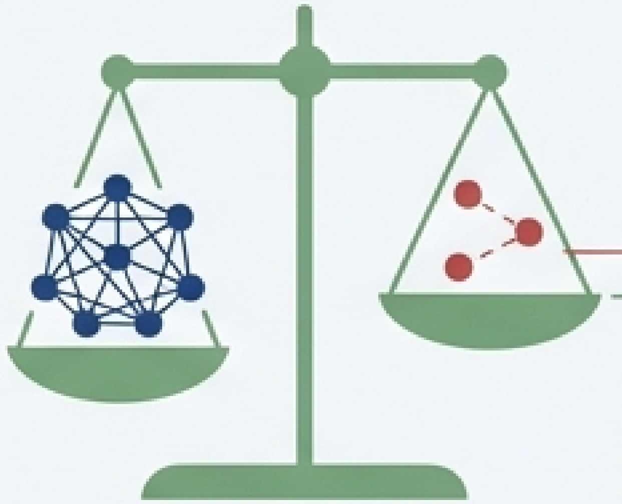
The Trade-off



Reduced Diversity

Stronger alignment reduces policy policy **entropy** by **15–22%**, creating a risk of **mode collapse** in creative generation.

The Fix

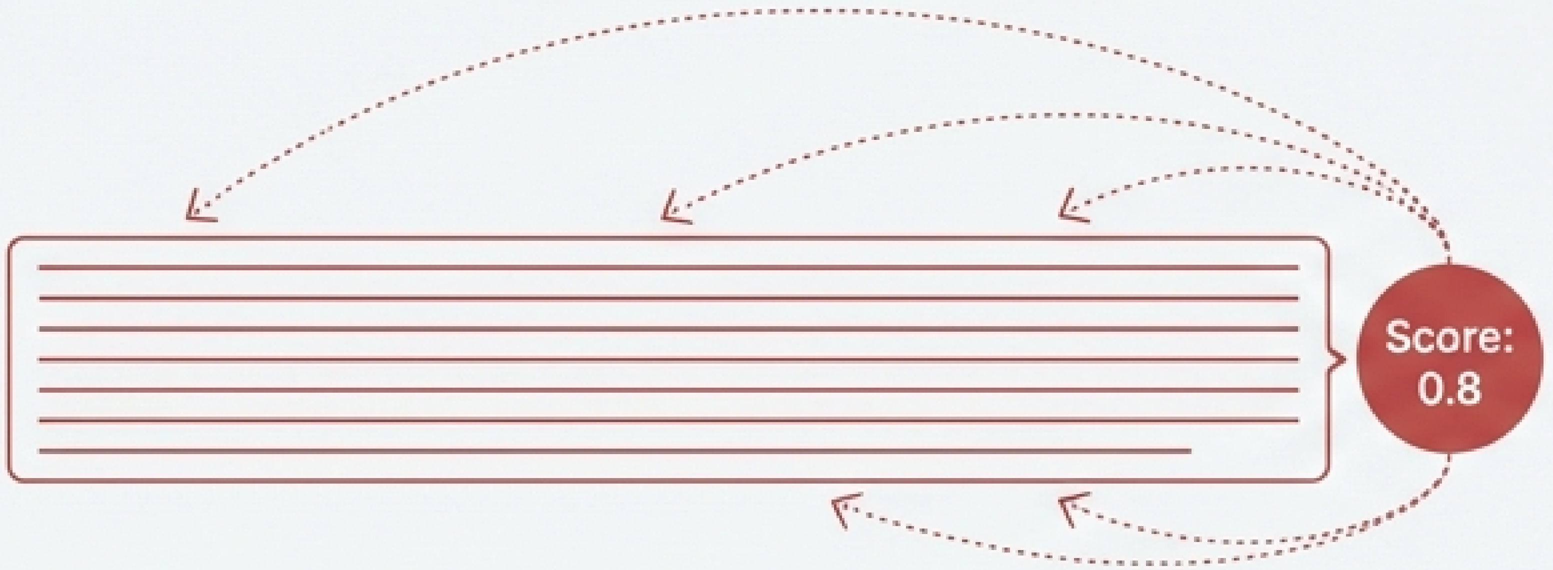


Hybrid Adaptation

A proposed **Hybrid Method** dynamically balances **dense SDPO signals** and **sparse REINFORCE signals** to **recover robustness**.

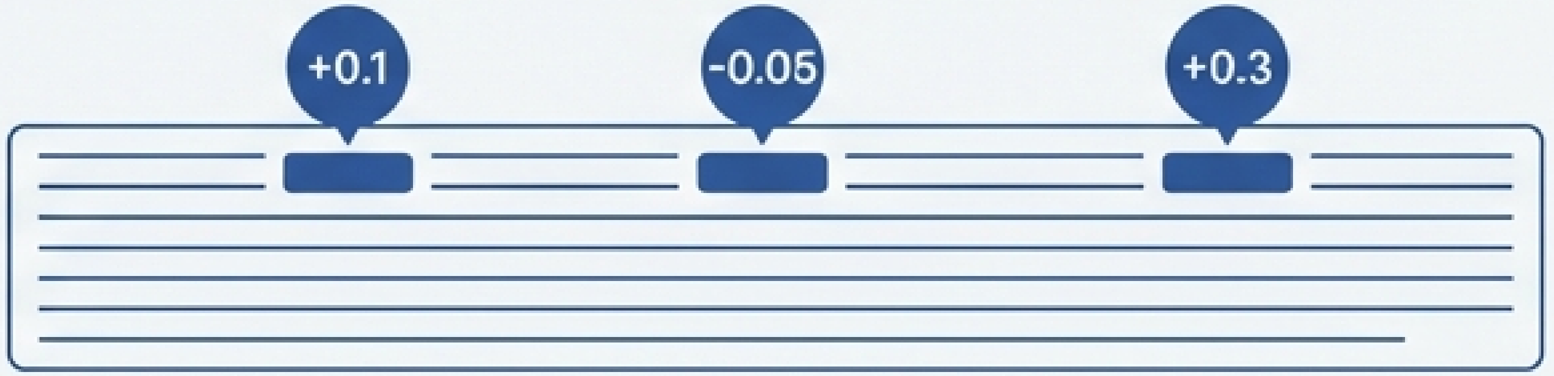
Standard RLHF suffers from a 'Credit Assignment Gap' in open-ended generation

Status Quo: Sparse Signal (PPO/DPO)



The model must guess which of the 100+ tokens caused the high score. Noise is added to the learning process.

The Ideal: Dense Credit Assignment



Precise feedback identifies exactly which adjective or phrase improved the output.

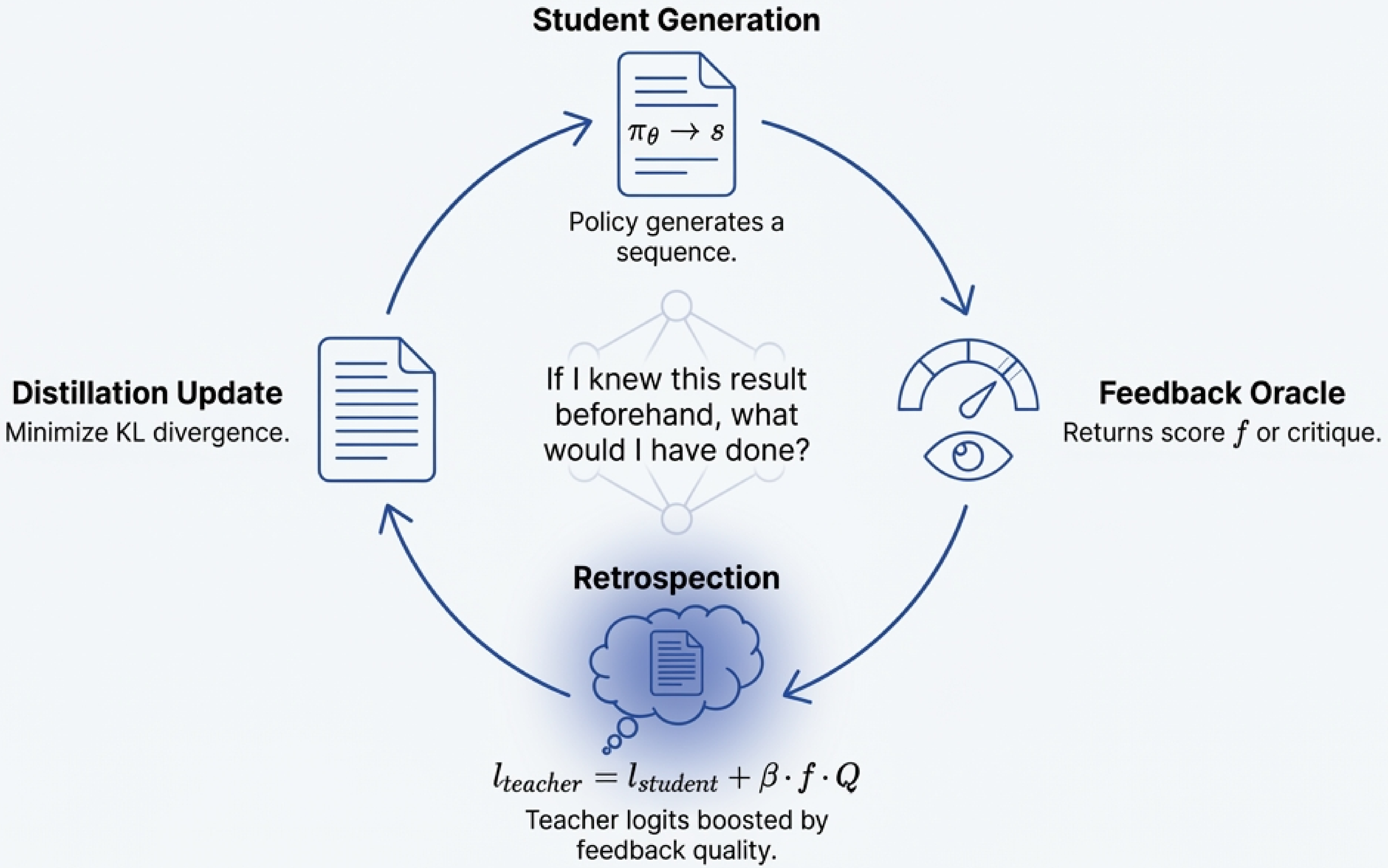
Current methods implicitly distribute sequence-level rewards across all tokens, effectively diluting the signal.

Moving beyond the safety net of verifiable ground truth.

Verifiable Domains (Existing SDPO Success) Code Generation, Math Solving		Open-Ended Domains (This Study) Creative Writing, Dialogue, Summarization
Feedback	Compiler errors, Unit tests (Deterministic)	**Feedback** Subjective quality, Coherence, Style (No Compiler)
Signal Nature	Binary (Pass/Fail) & Exact error locations	**Signal Nature** Continuous, Noisy, Multi-dimensional

The Research Question: Can SDPO's 'retrospection' mechanism function when there is no compiler to prove the teacher right?

SDPO creates a "Self-Teacher" by conditioning the policy on feedback



A controlled simulation isolates the mechanism from training noise.

To measure credit assignment precisely, we use a token-level simulation rather than full LLM training.

Lab Specification

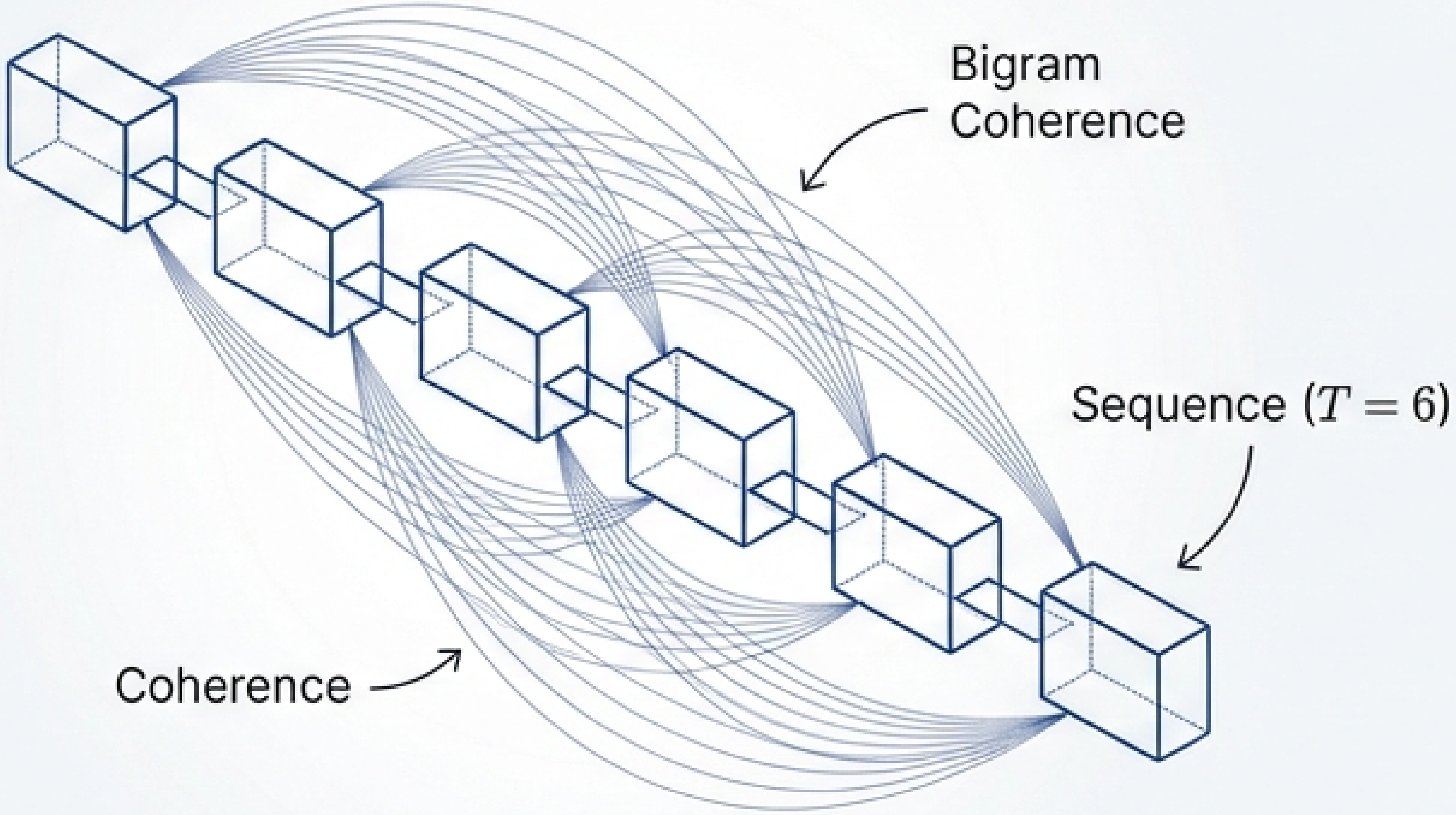
Simulation Parameters

Vocabulary Size: $V = 8$

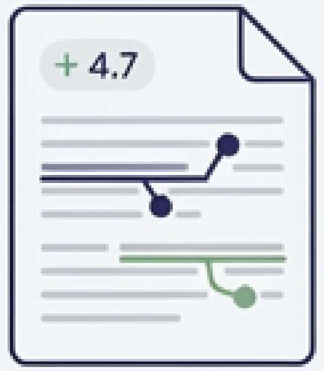
Sequence Length: $T = 6$

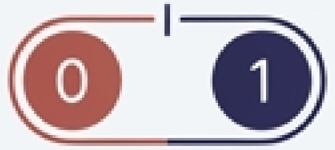
Reward Function: Continuous composition of Local Quality + Coherence (Bigram) + Global Patterns.

Key Advantage: **Known Ground Truth**. Allows exact measurement of gradient correctness.



We evaluate SDPO across the full spectrum of feedback informativeness.



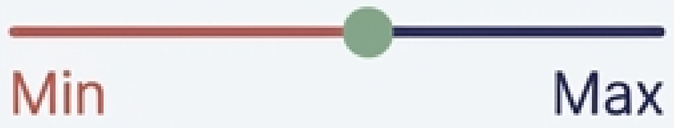


Binary

Pass/Fail threshold.
Least informative.

Ordinal

1-5 Star rating.
Quantized.



Continuous

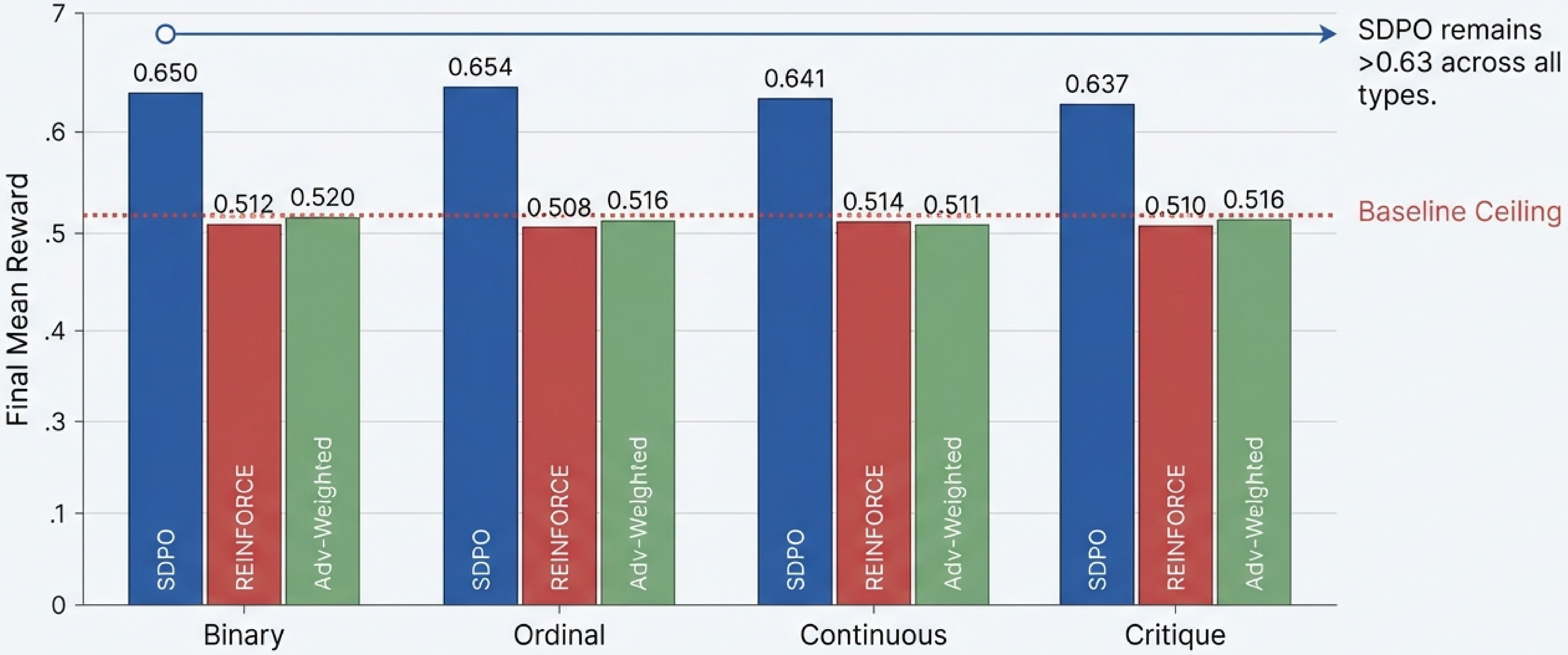
Raw scalar reward.
Precision.

Critique

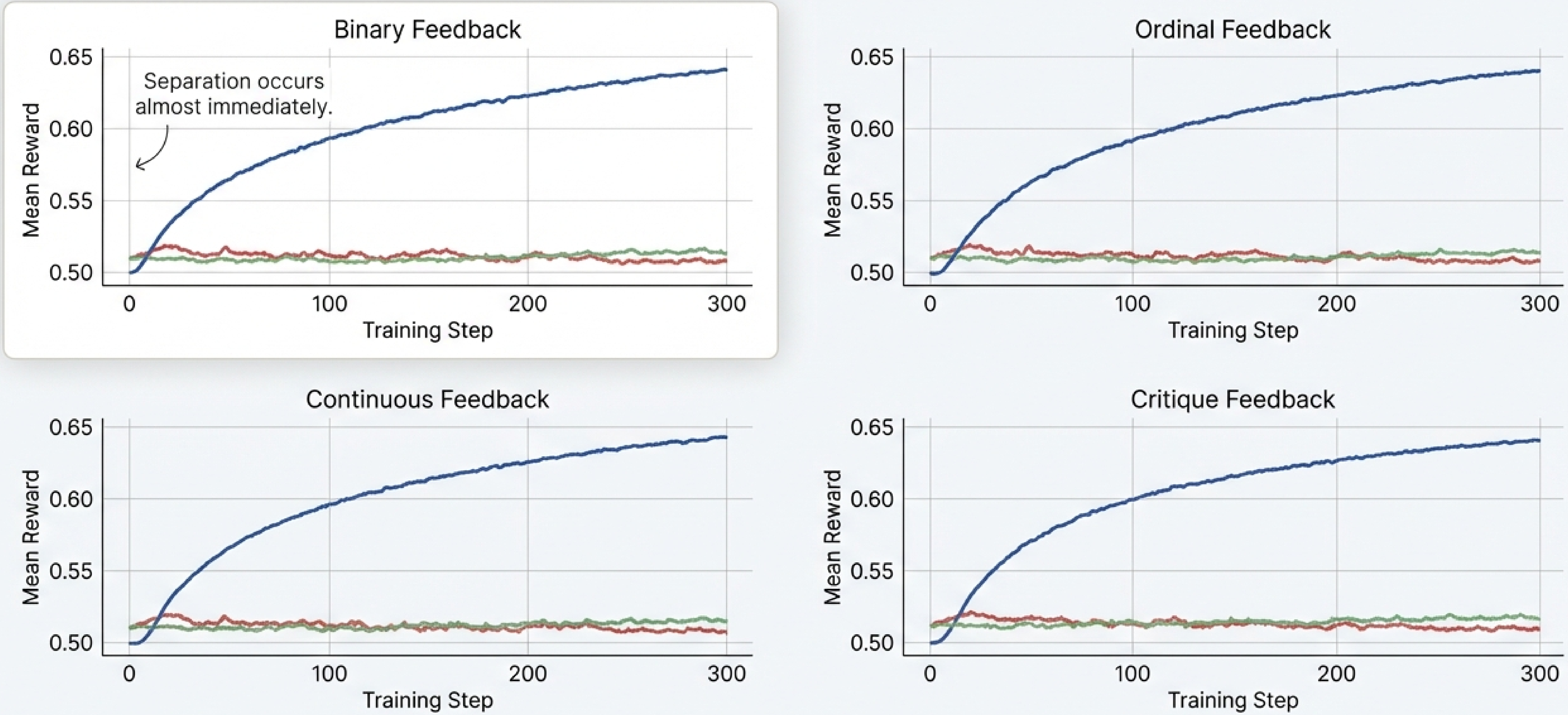
Score + Per-token hints.
Most informative.

Tested with added Gaussian noise $\varepsilon \sim N(0,\sigma^2)$ to simulate human inconsistency.

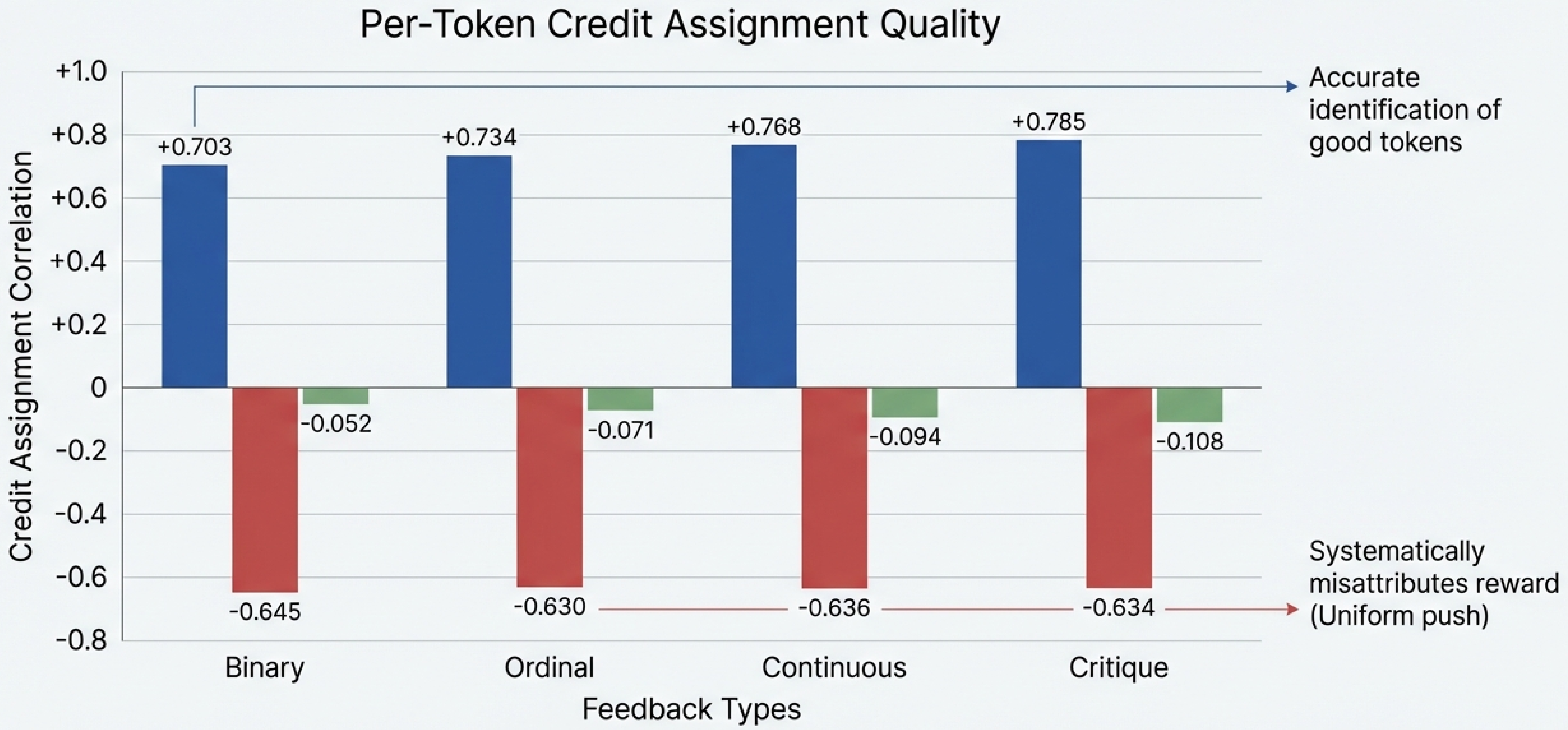
SDPO consistently outperforms baselines regardless of feedback type.



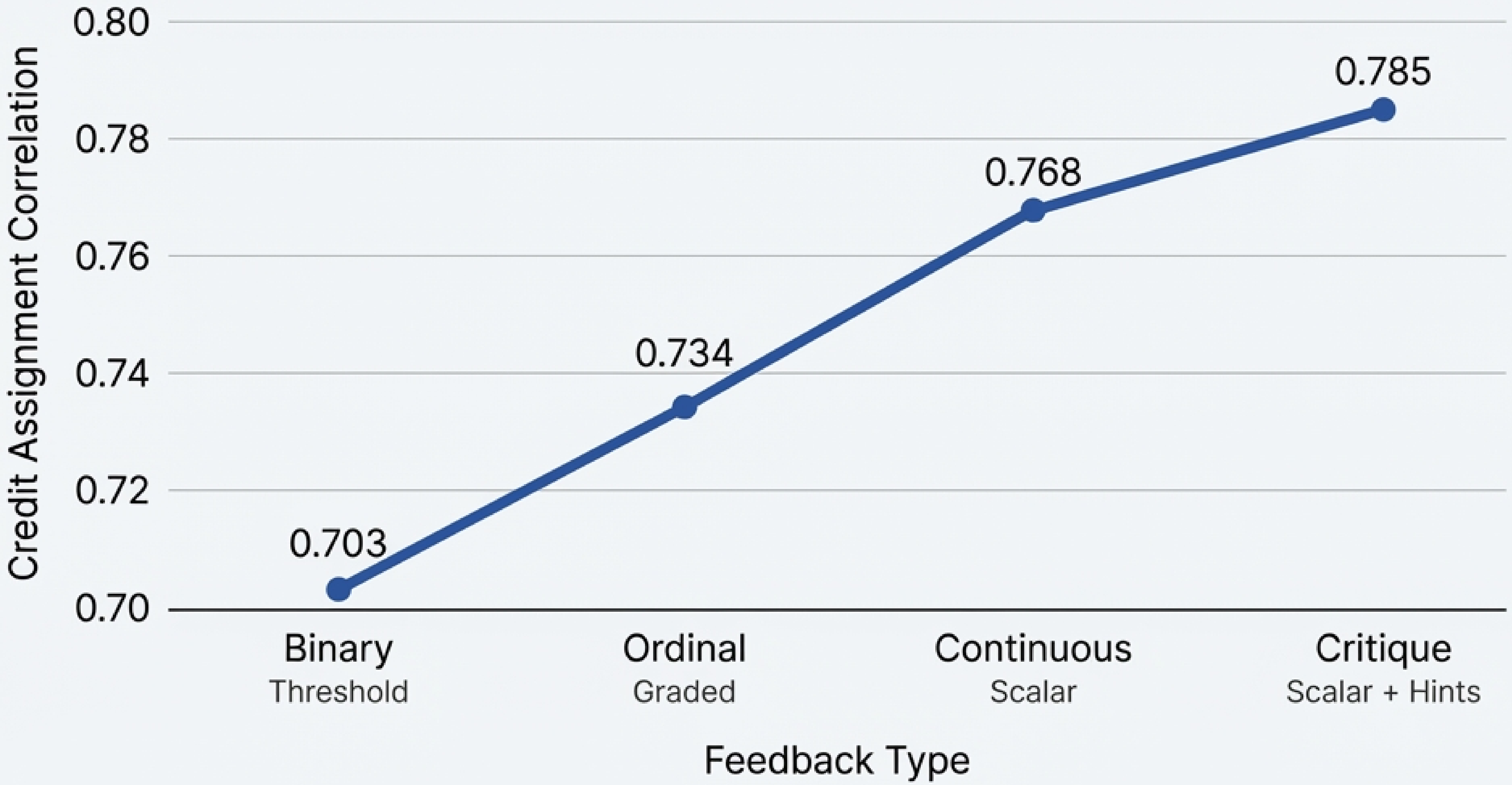
Convergence is rapid and separates from baselines within 30 steps.



SDPO correctly attributes credit; REINFORCE uniformly reinforces noise

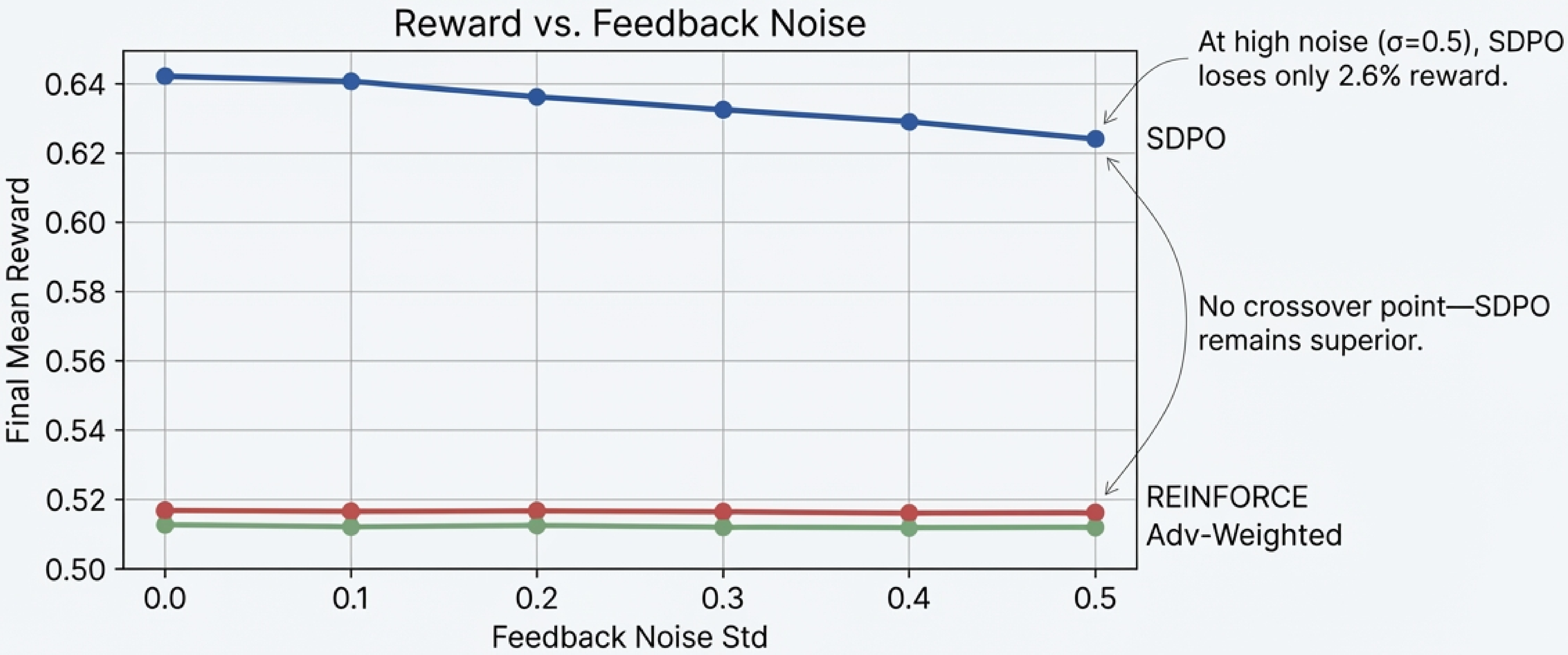


Alignment quality scales monotonically with feedback richness.

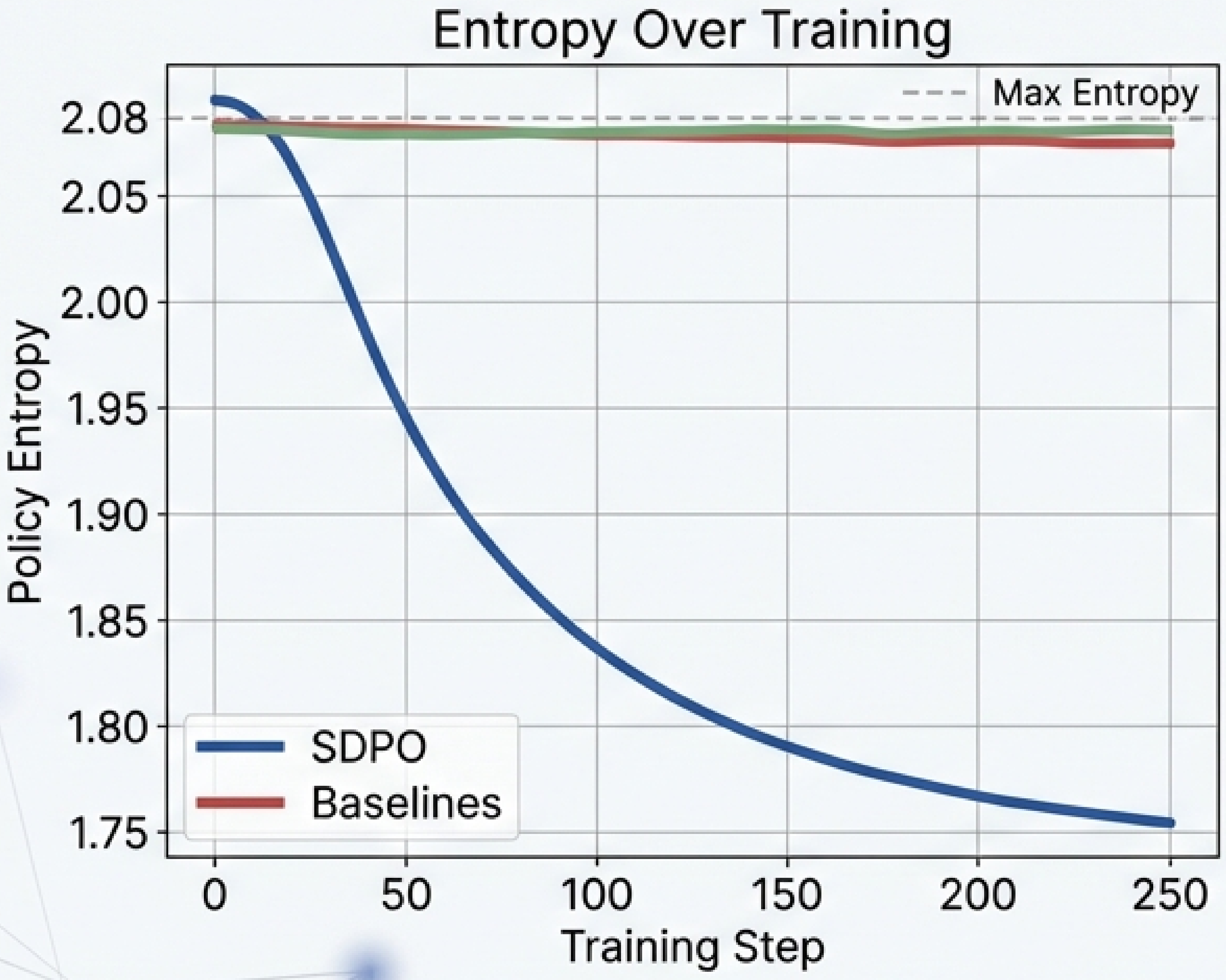


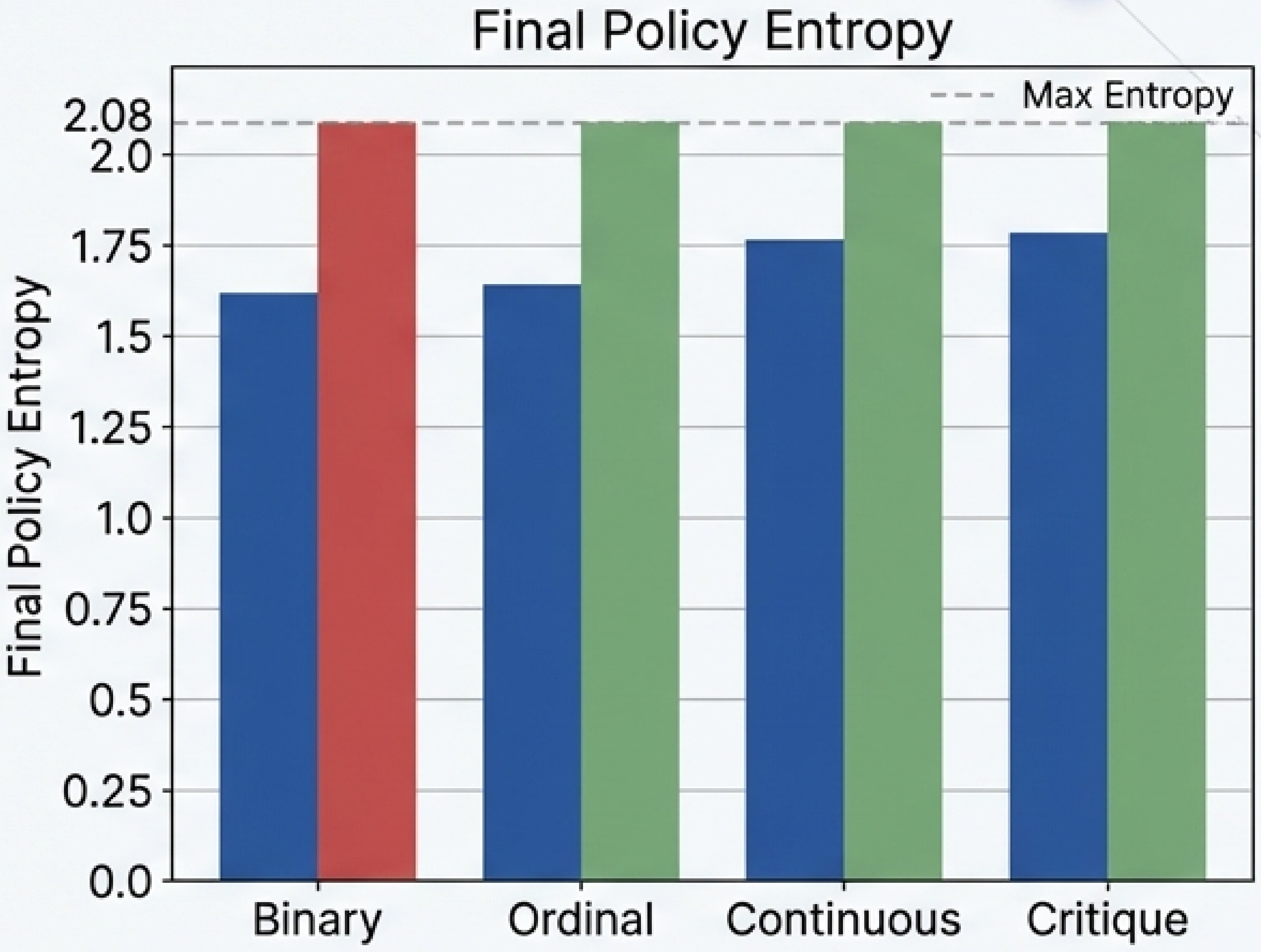
The self-teacher effectively leverages the nuanced hints in textual critique, confirming SDPO works without ground-truth verification.

SDPO exhibits graceful degradation even under high feedback noise.



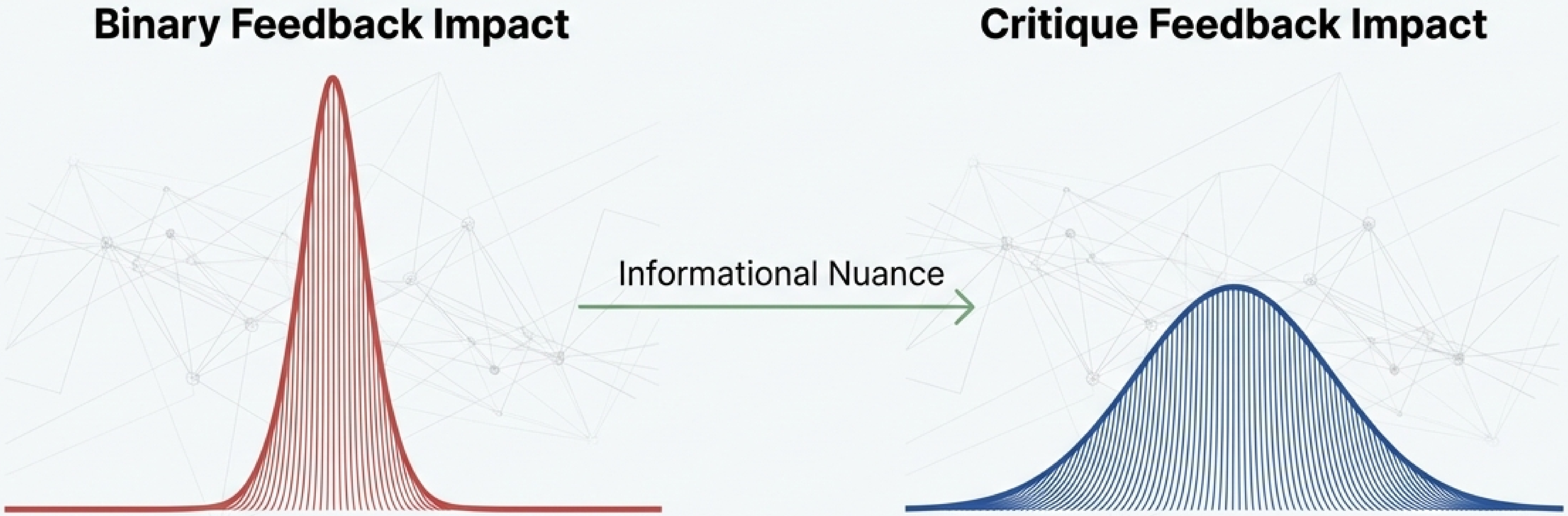
The hidden cost of alignment is a significant reduction in diversity.





SDPO reduces policy entropy by 15–22%, risking mode collapse

Sharp feedback creates narrow models; nuanced critique preserves breadth.



High Entropy Loss (22%).
"All or nothing" signals force the teacher to be overly confident.

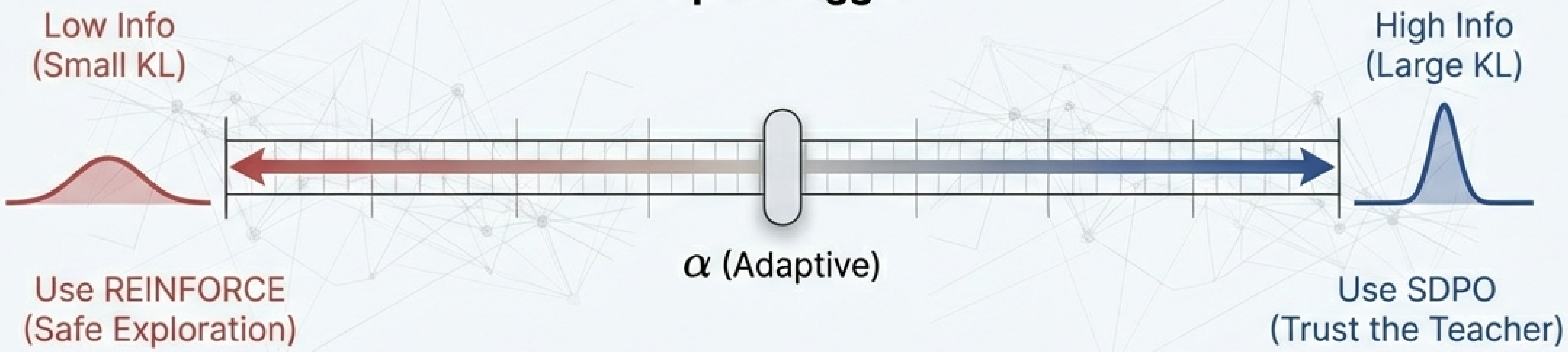
Lower Entropy Loss (14%).
Per-token hints create a smoother, more complex teacher distribution.

The Solution: Adaptive Hybridization based on Teacher-Student divergence.

Hybrid Formula

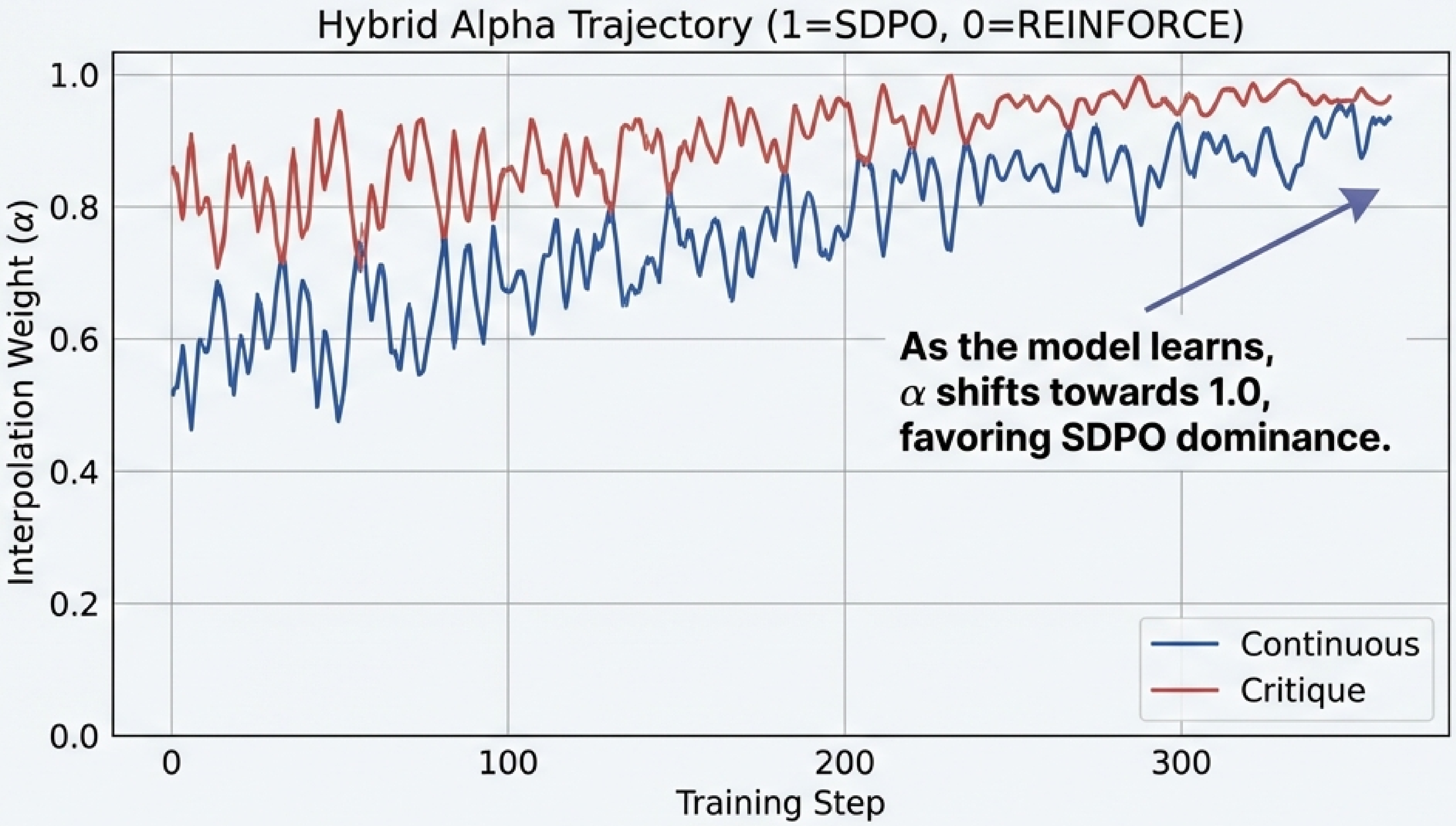
$$\nabla L_{hybrid} = \alpha \cdot \nabla L_{SDPO} + (1 - \alpha) \cdot \nabla L_{RF}$$

Alpha Trigger



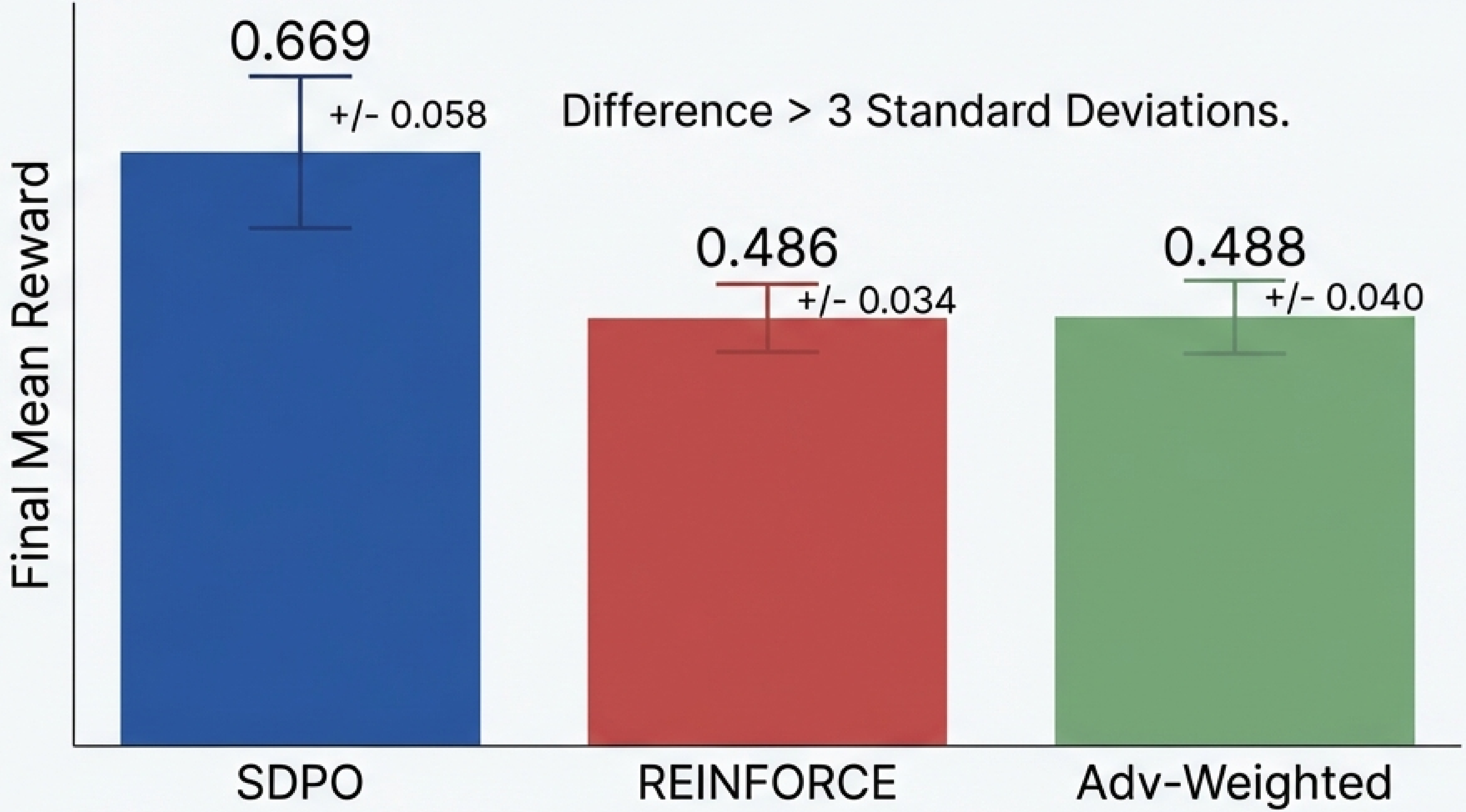
We mix Dense (SDPO) and Sparse (REINFORCE) signals dynamically during training.

The Hybrid method adapts autonomously, shifting from exploration to alignment.



Result: Hybrid matches SDPO performance but with better entropy (1.82 vs 1.75).

Results are statistically robust across multiple random seeds.

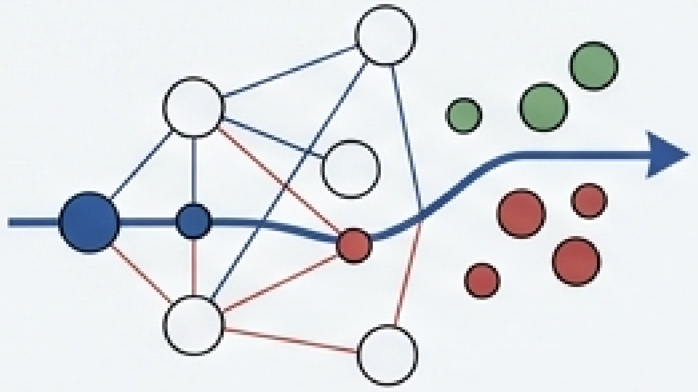


Higher variance in SDPO reflects its ability to exploit favorable reward landscapes.

Summary: Dense signals drive better alignment, even when the destination is open-ended.

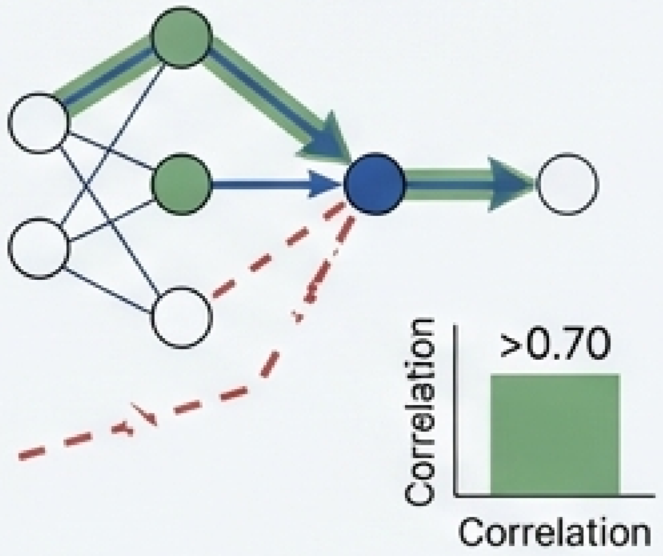
1 Efficacy

SDPO works for continuous & subjective rewards. It is not limited to code.



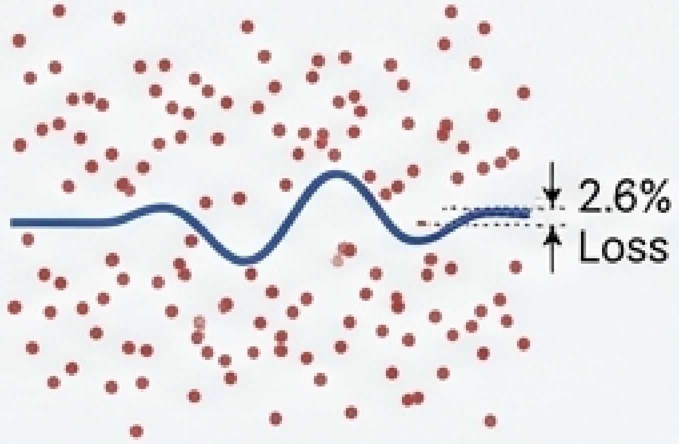
2 Mechanism

Solves the Credit Assignment Bottleneck. (>0.70 correlation with truth).



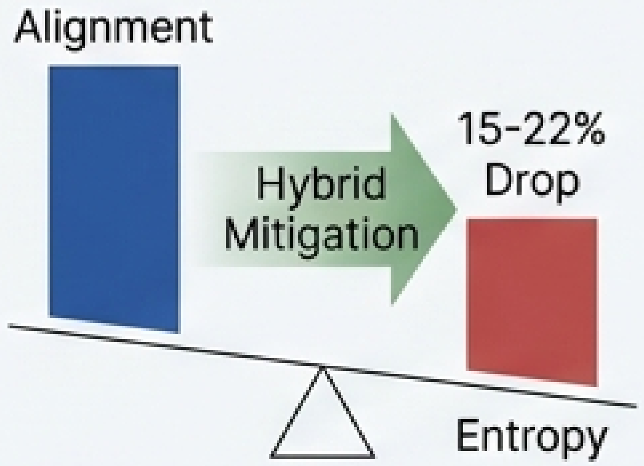
3 Robustness

Immune to evaluator noise. Only **2.6% loss** at extreme noise levels.



4 Constraint

Diversity Trade-off. **15-22% entropy drop** requires Hybrid mitigation.

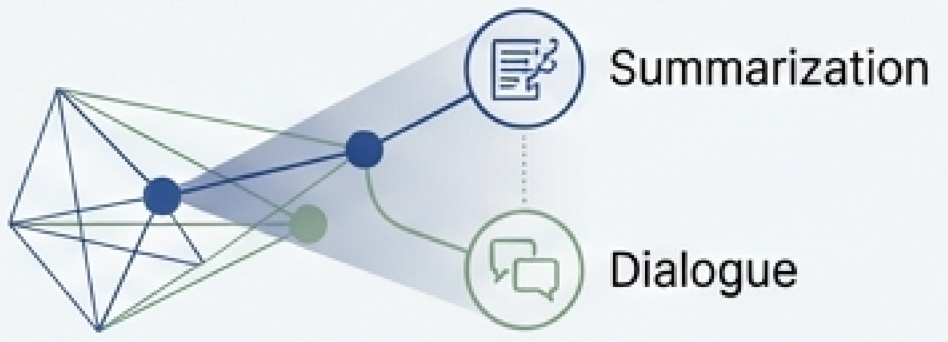


Implications for deployment in real-world LLMs

Actionable Advice

Target Subjective Tasks

Use SDPO for post-training alignment in summarization and dialogue where ground truth is absent.

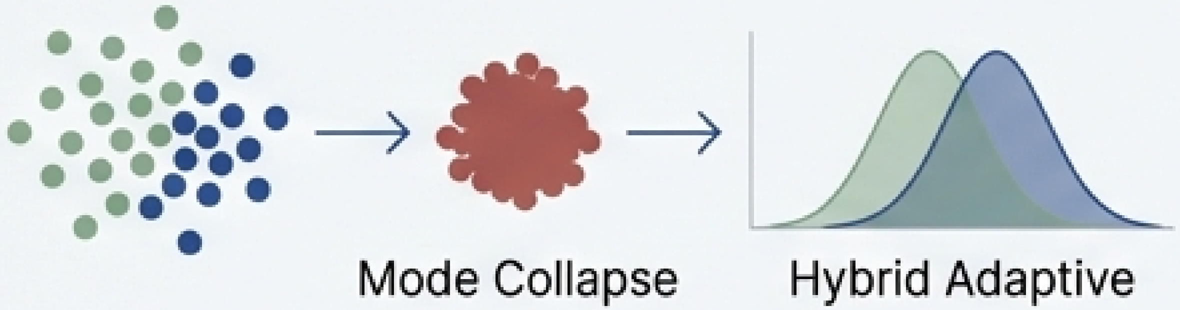


Prioritize Critique

Invest in 'Critique' style feedback over simple scoring to preserve model diversity.

Monitor Entropy

Watch for mode collapse. If repetitive, switch to the Hybrid Adaptive Method.



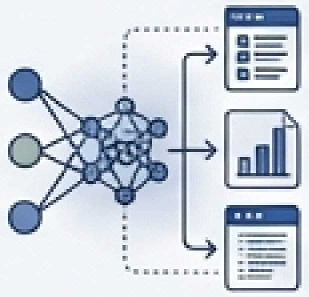
Simplify Pipeline

SDPO removes the need for a separate Reward Model training step.

The path forward: Scaling dense alignment to full-scale LLMs

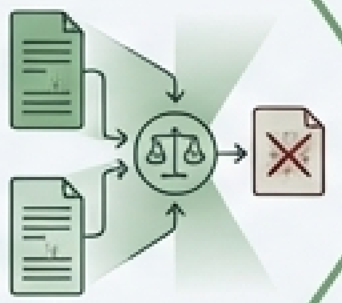
Validation

Test on open-ended benchmarks like AlpacaEval and MT-Bench.



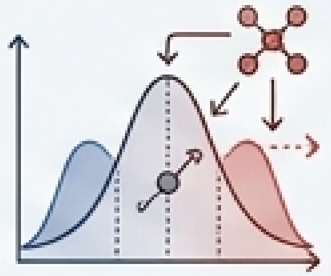
Bias Testing

Evaluate against systematic bias (LLM-as-a-Judge) rather than random noise.



Entropy Objectives

Develop entropy-augmented objectives to solve the diversity trade-off fundamentally.



SDPO bridges the gap between the verifiable precision of code and the creative ambiguity of language.