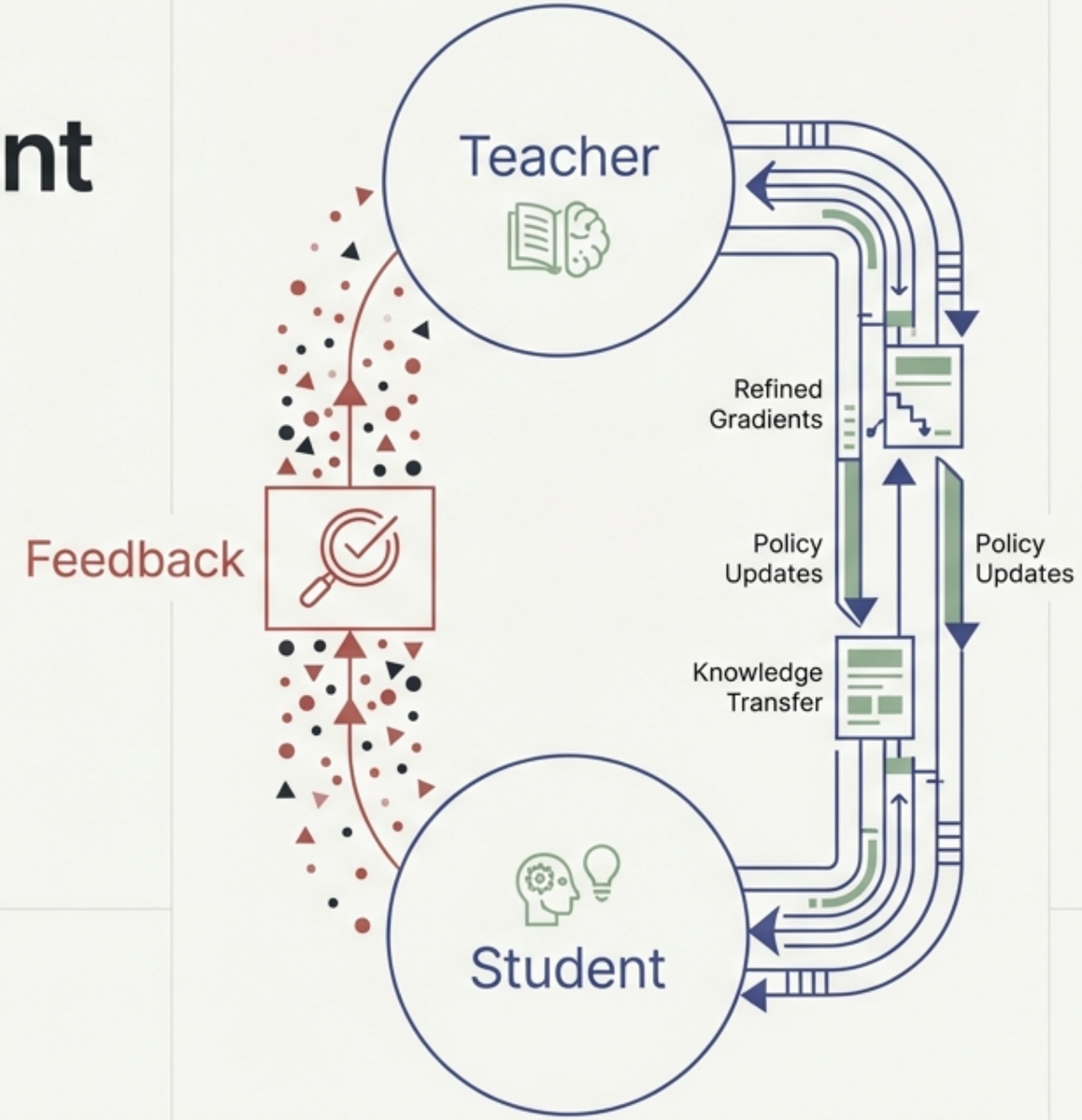
Self-Distillation Policy Optimization for Alignment in Open-Ended Settings

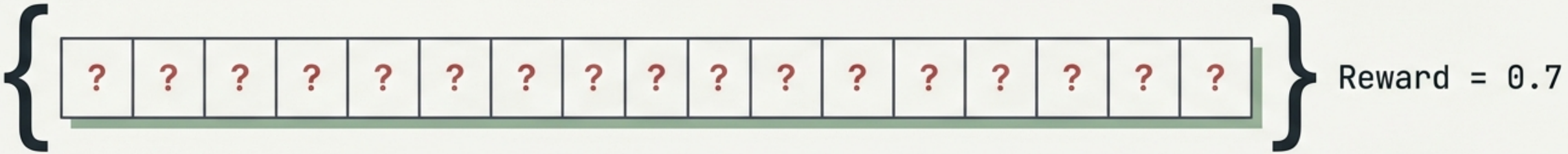
A Simulation Study on Dense Credit Assignment for Continuous-Reward Tasks



FOCUS: Open-Ended Text Generation
DOMAIN: Continuous-Reward / No Ground Truth
METHOD: Feedback-Conditioned Retrospection

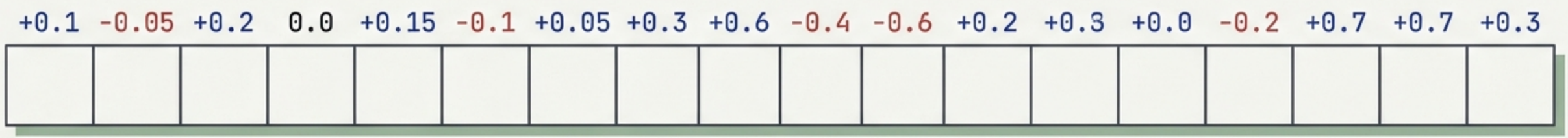
The Credit Assignment Bottleneck in Current RLHF

CURRENT PARADIGM (Sparse Reward)



PPO/DPO: Signal is diffuse.

REQUIRED SHIFT (Dense Credit)

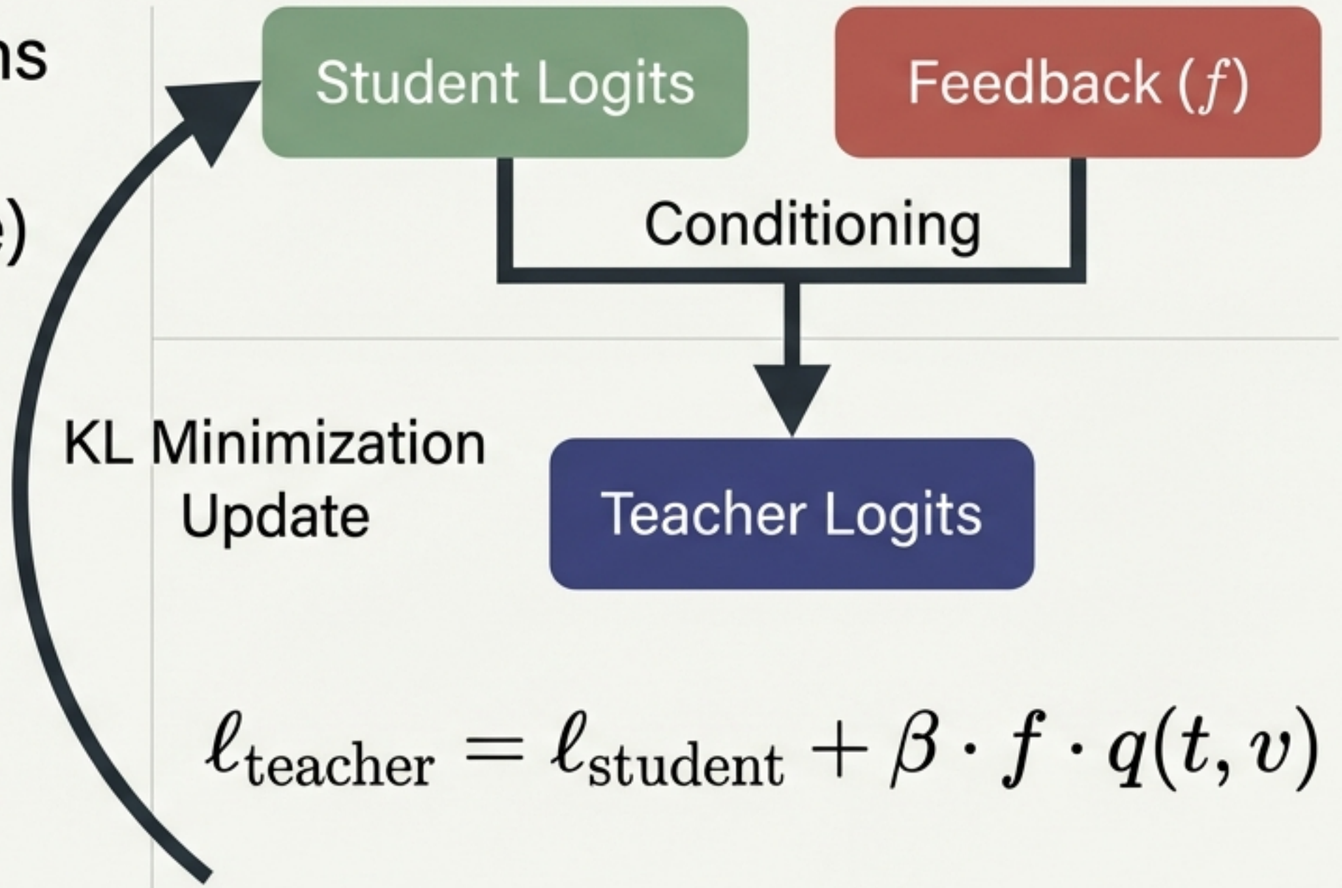


Target State: Signal is specific to the token.

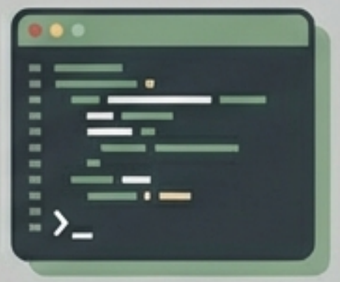
SDPO Creates Dense Signal via Feedback-Conditiorned Retrospection

The Mechanism: SDPO conditions the model on external feedback (like an error message or critique) to temporarily create a “Self-Teacher”.

The Distillation: The unconditioned "Student" policy is then trained to match this superior Teacher distribution via token-level KL divergence.



The Open Question: Can Retrospection Work Without Ground Truth?



Verifiable Tasks (Code)

- Binary Success/Fail
- Compiler = Ground Truth
- Proven Success

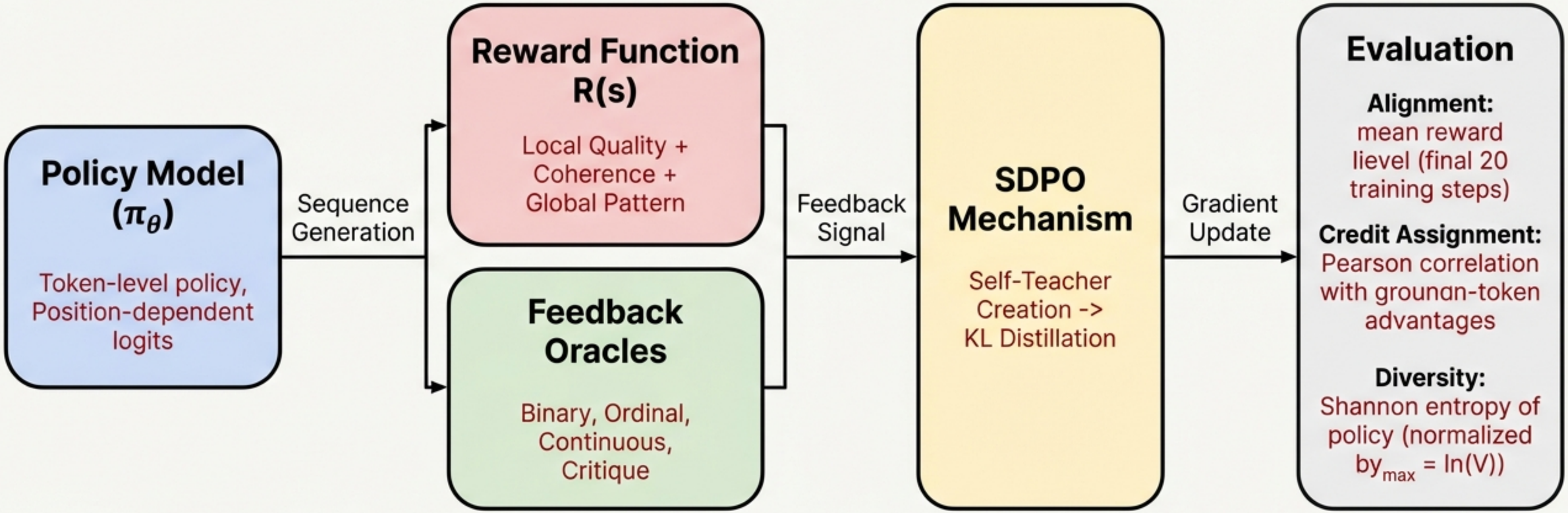
Open-Ended Generation (Language)

- Subjective Quality
- Continuous/Ordinal Rewards
- No Ground Truth Verifier

HYPOTHESIS: Can SDPO survive here?

A Controlled Simulation Framework to Isolate the Mechanism

Modeling the full RLHF pipeline with known ground-truth parameters.



Stress-Testing Across Four Regimes of Feedback Informativeness

Low Info ⟷ **High Info**

Binary

Pass/Fail threshold at 0.5.

Ordinal

1–5 Likert scale.

Continuous

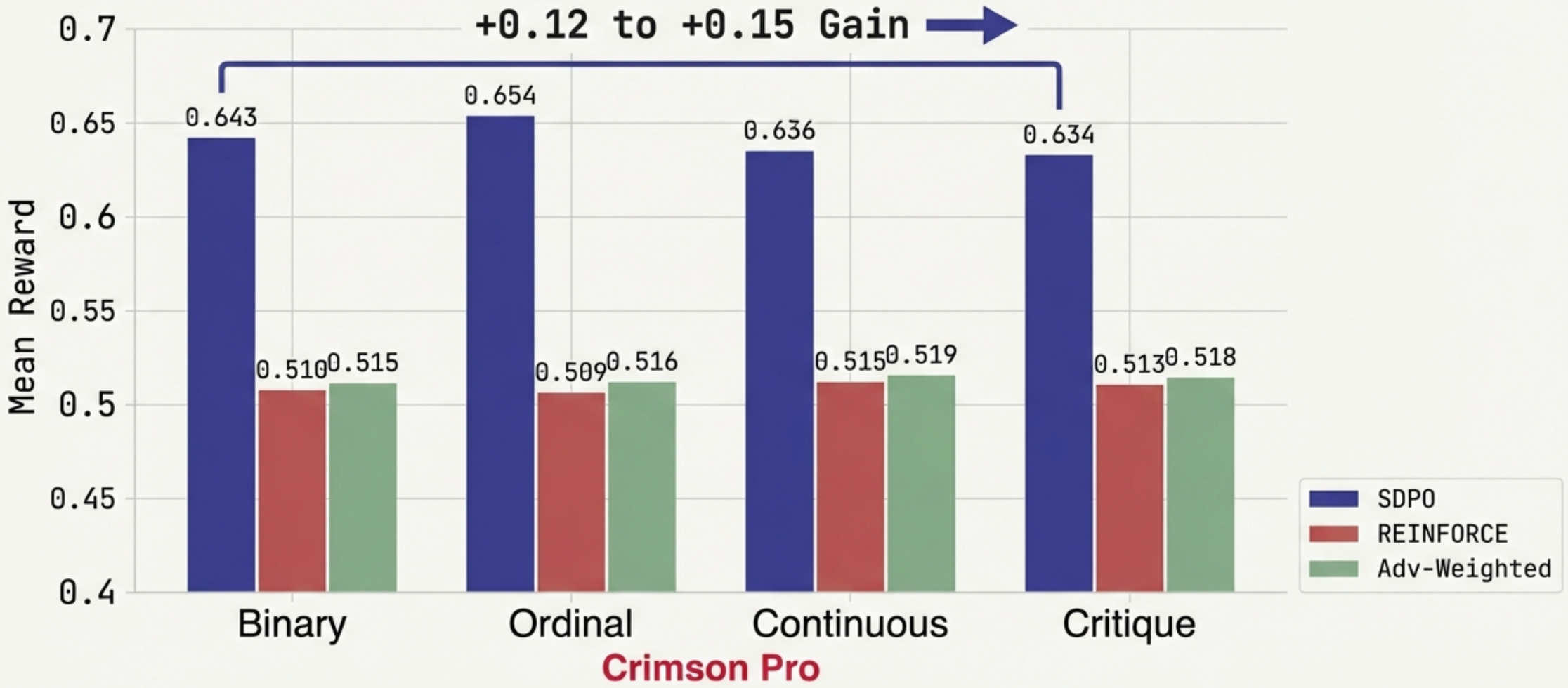
Raw scalar reward.

Critique

Continuous score + per-token hints.

All oracles tested with Gaussian noise to simulate human inconsistency.

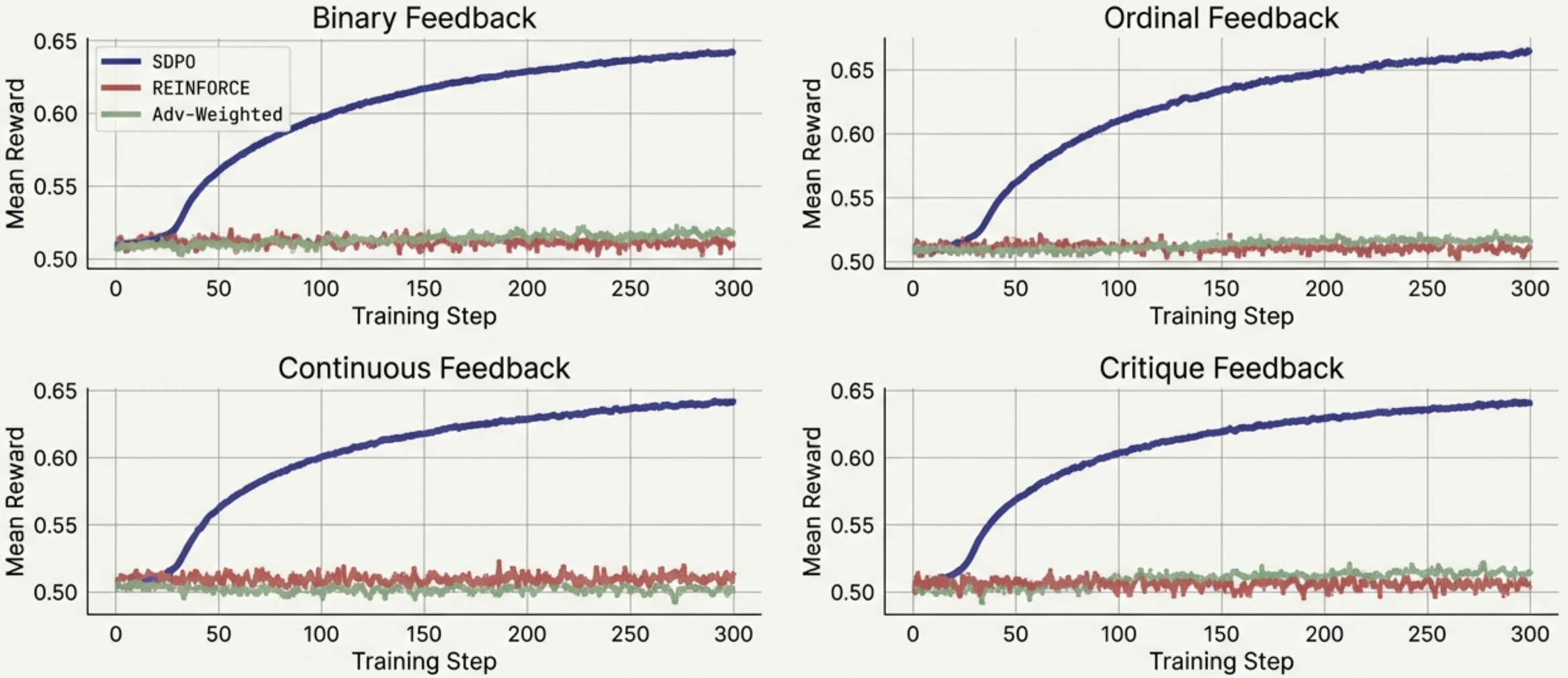
SDPO Consistently Outperforms Baselines Across All Feedback Types



Crimson Pro

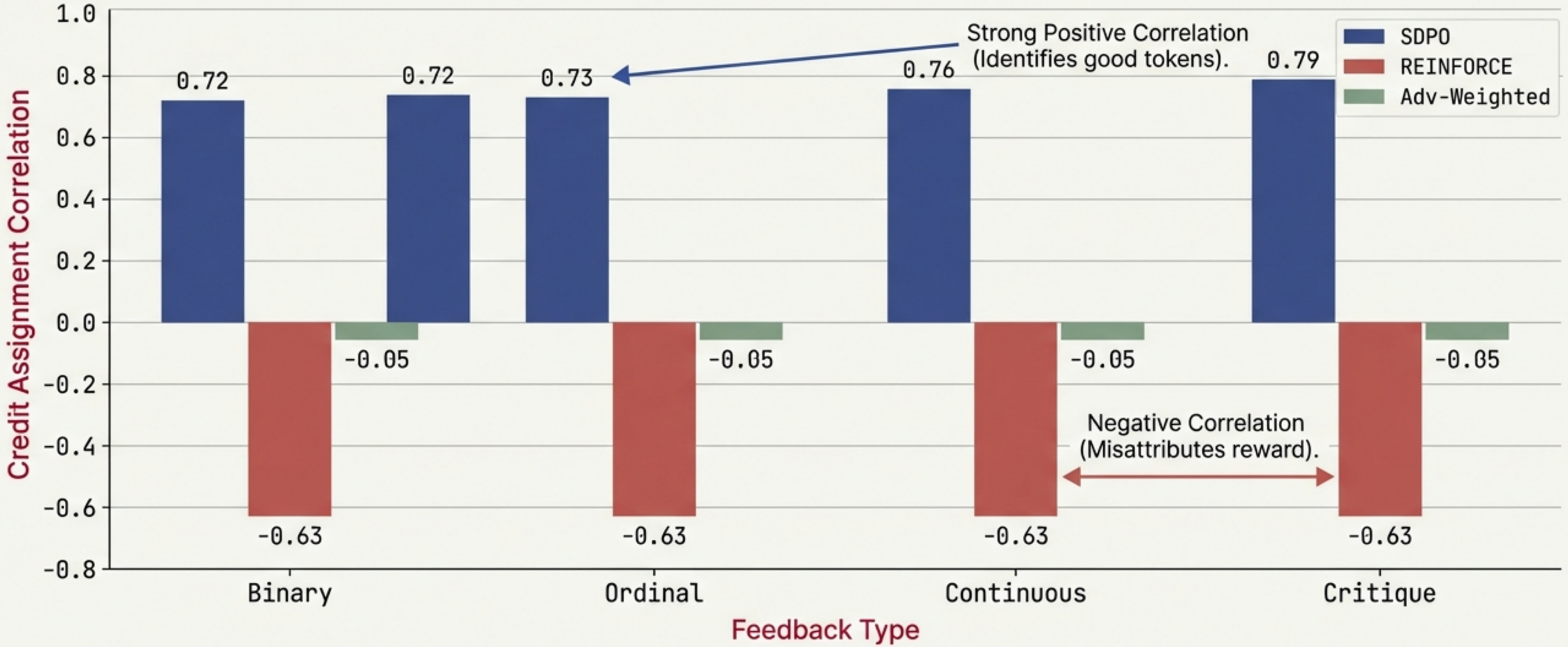
Modern Scientific Editorial

Rapid Convergence and Sustained Advantage

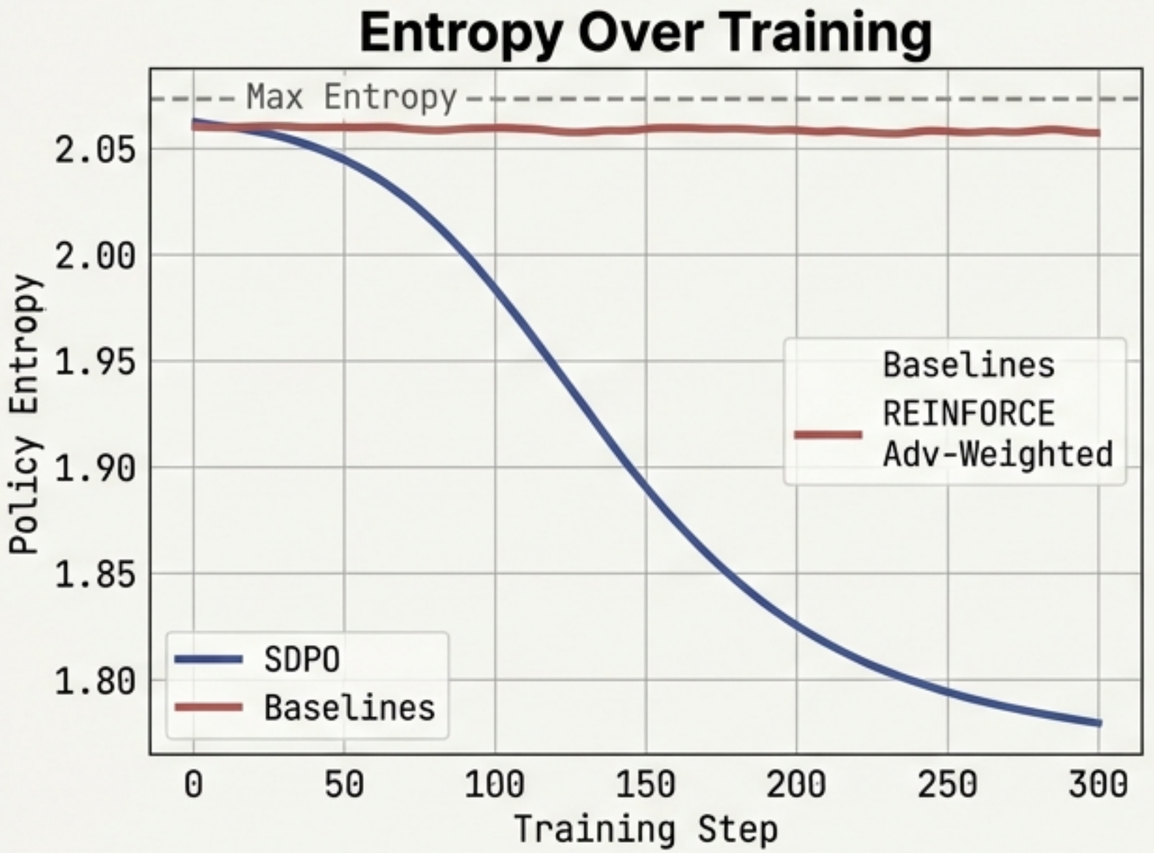


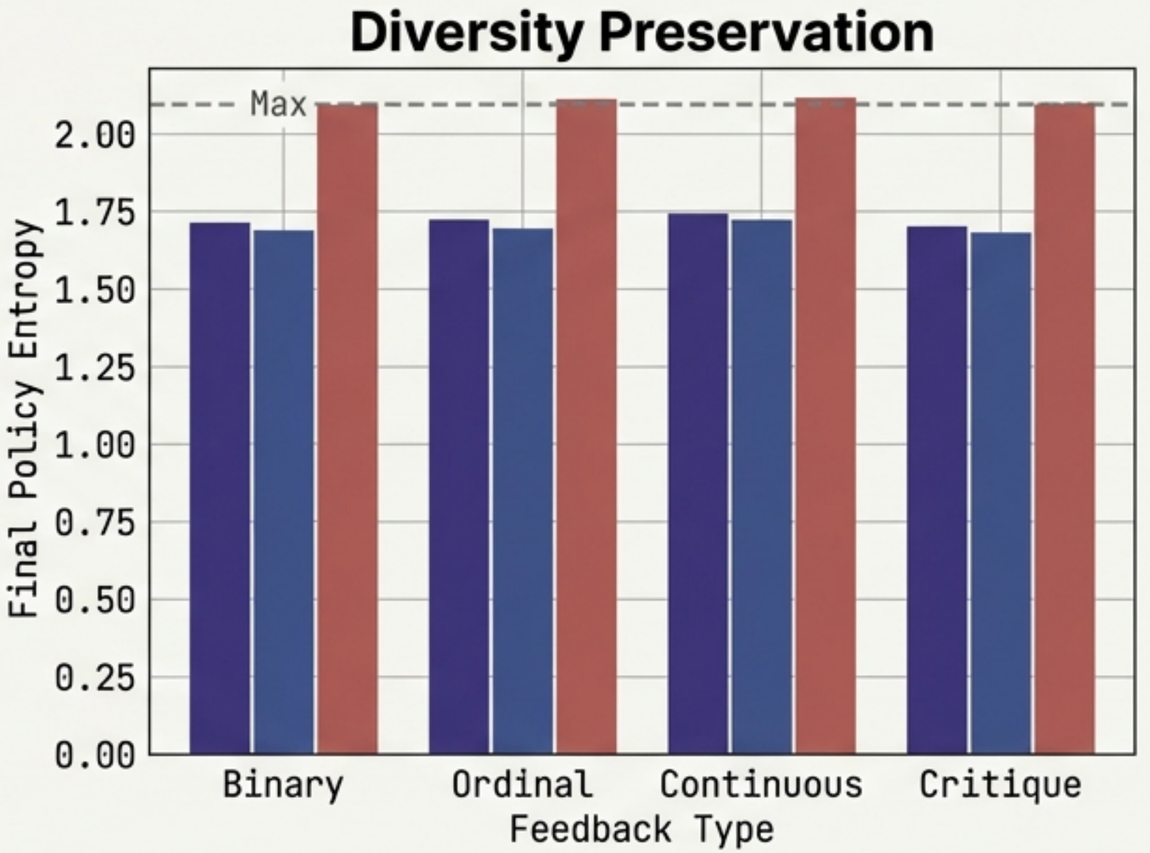
Early separation of the SDPO line demonstrates rapid learning, maintaining a performance advantage across all feedback regimes.

The Mechanism Verified: Superior Credit Assignment



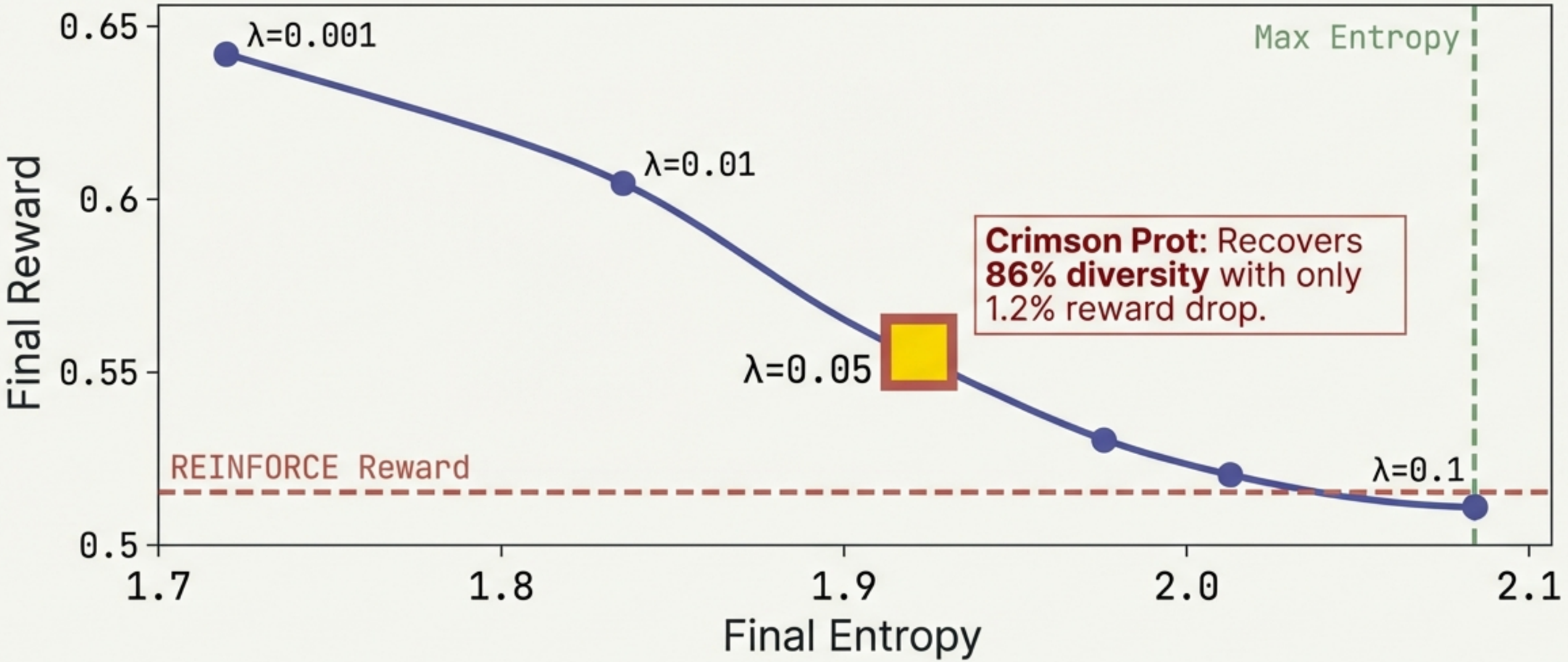
The Trade-off: SDPO Reduces Policy Diversity



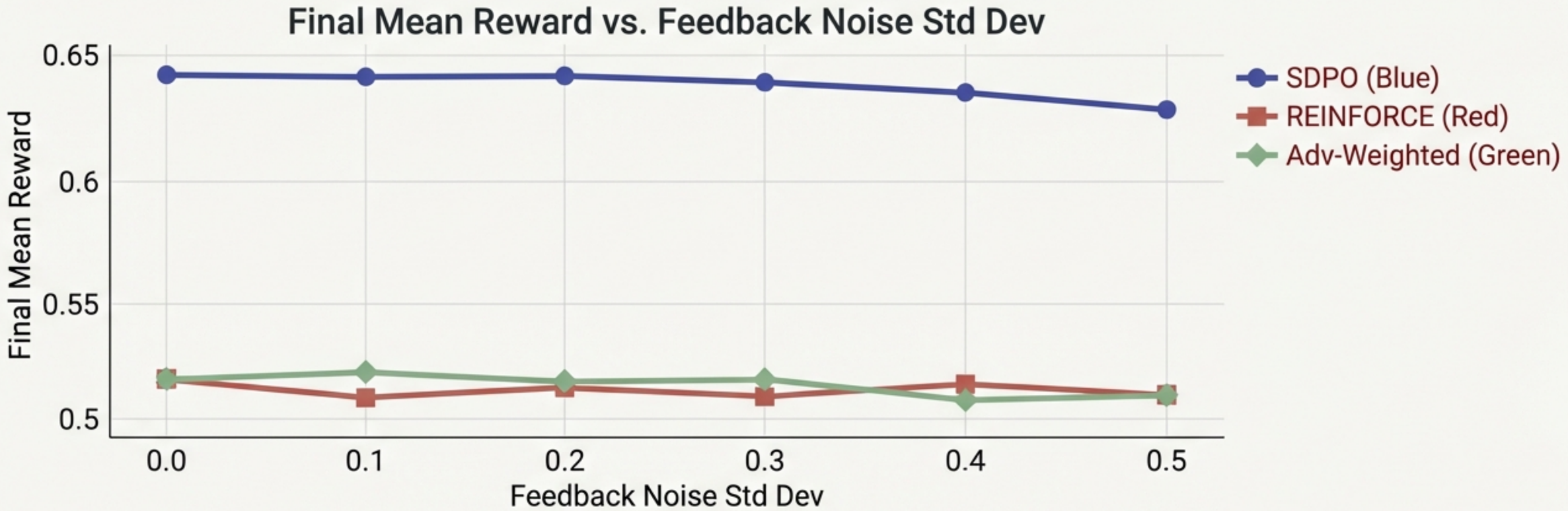


Insight: The ‘Self-Teacher’ distribution is sharp, causing the student to collapse towards the mode, reducing variety by ~15-20%.

Tuning the Pareto Frontier Recovers Diversity



Graceful Degradation Under Feedback Noise



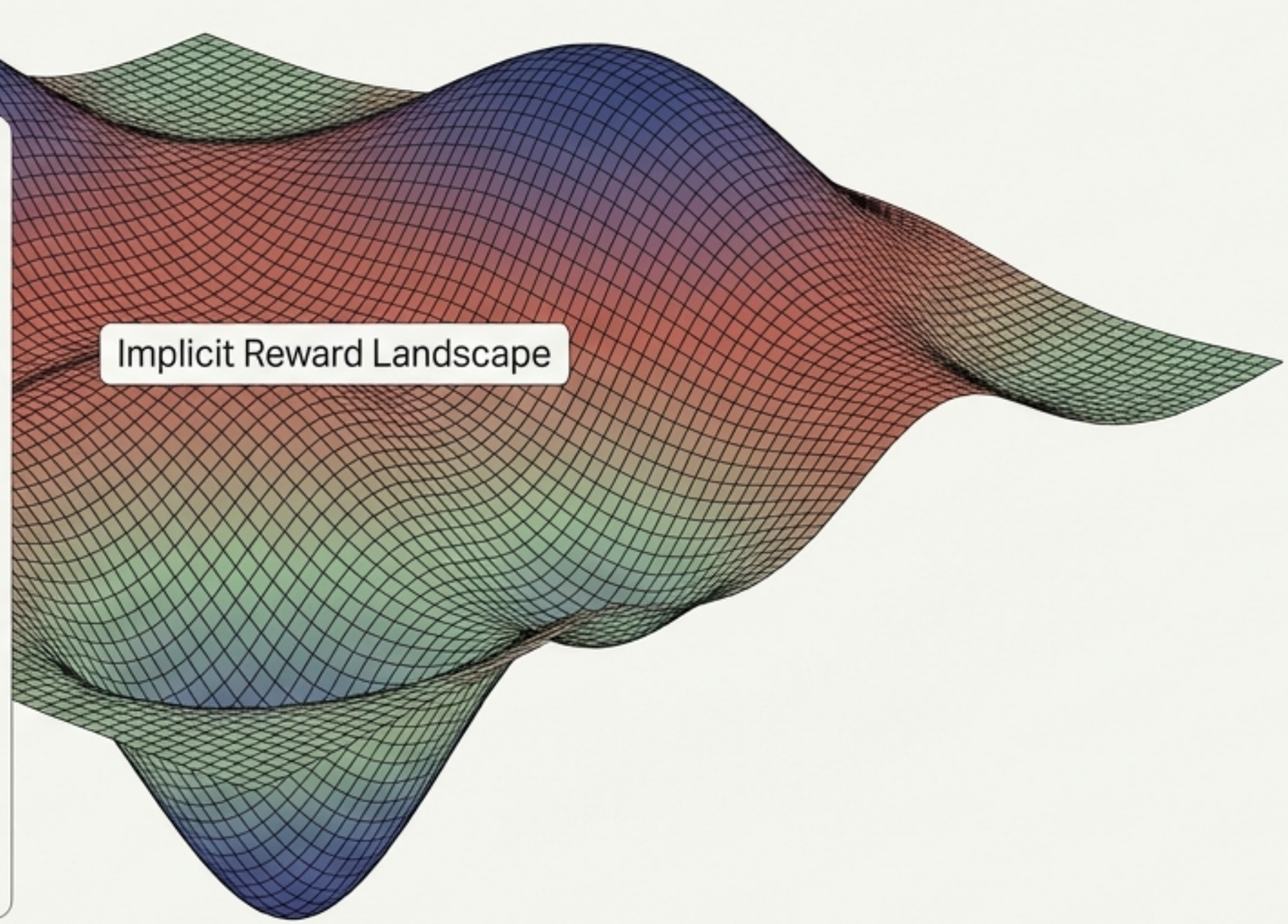
No Crossover Point. Even with 50% noise, SDPO remains superior to baselines.

Theoretical Grounding: Implicit Rewards and SDFT

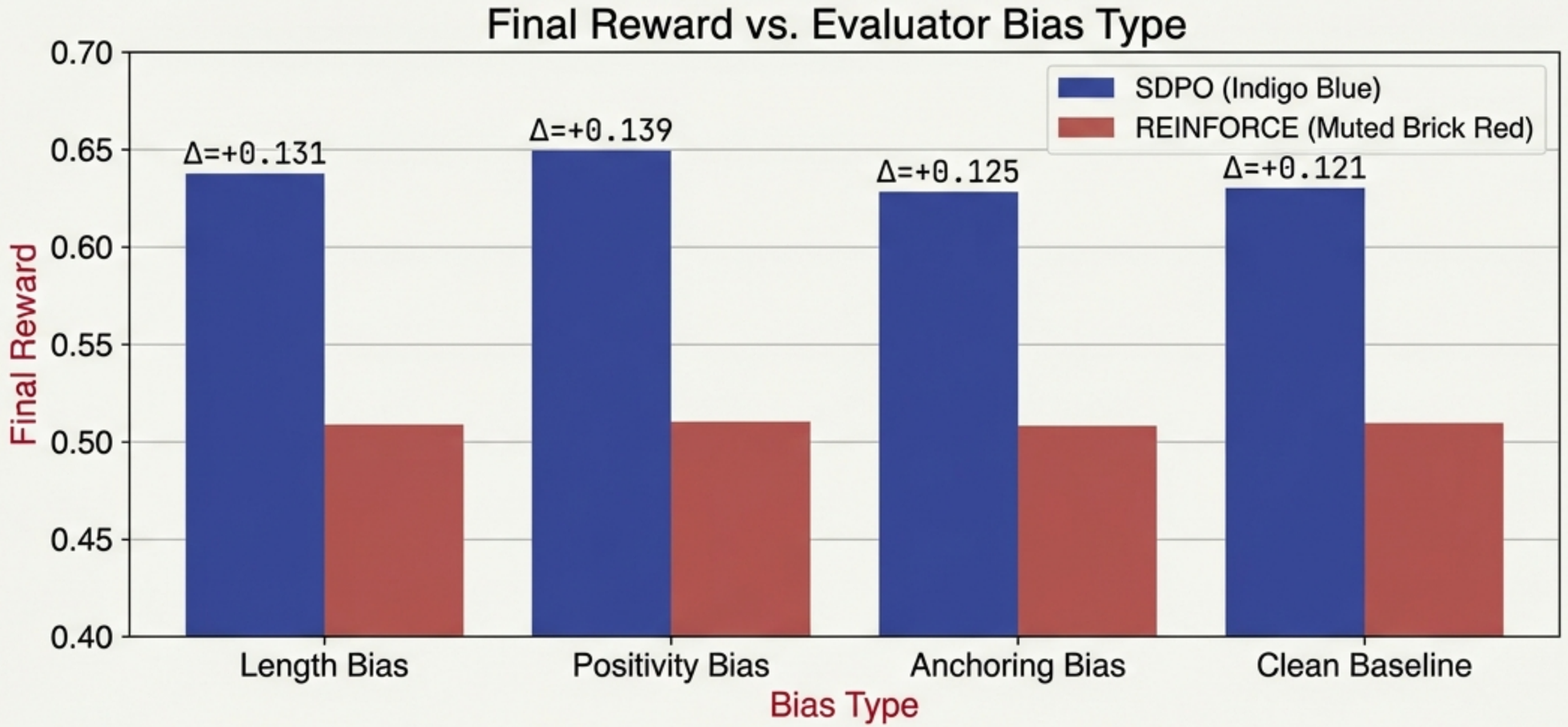
SDPO creates a smooth reward landscape defined by:

$$r(y, x, c) = \log \pi(y \mid x, c) - \log \pi_k(y \mid x)$$

- Conditioning on continuous feedback "c" creates a smooth function.
- Noise perturbs the landscape but does not destroy the gradient direction.
- Result: Graceful degradation instead of collapse.

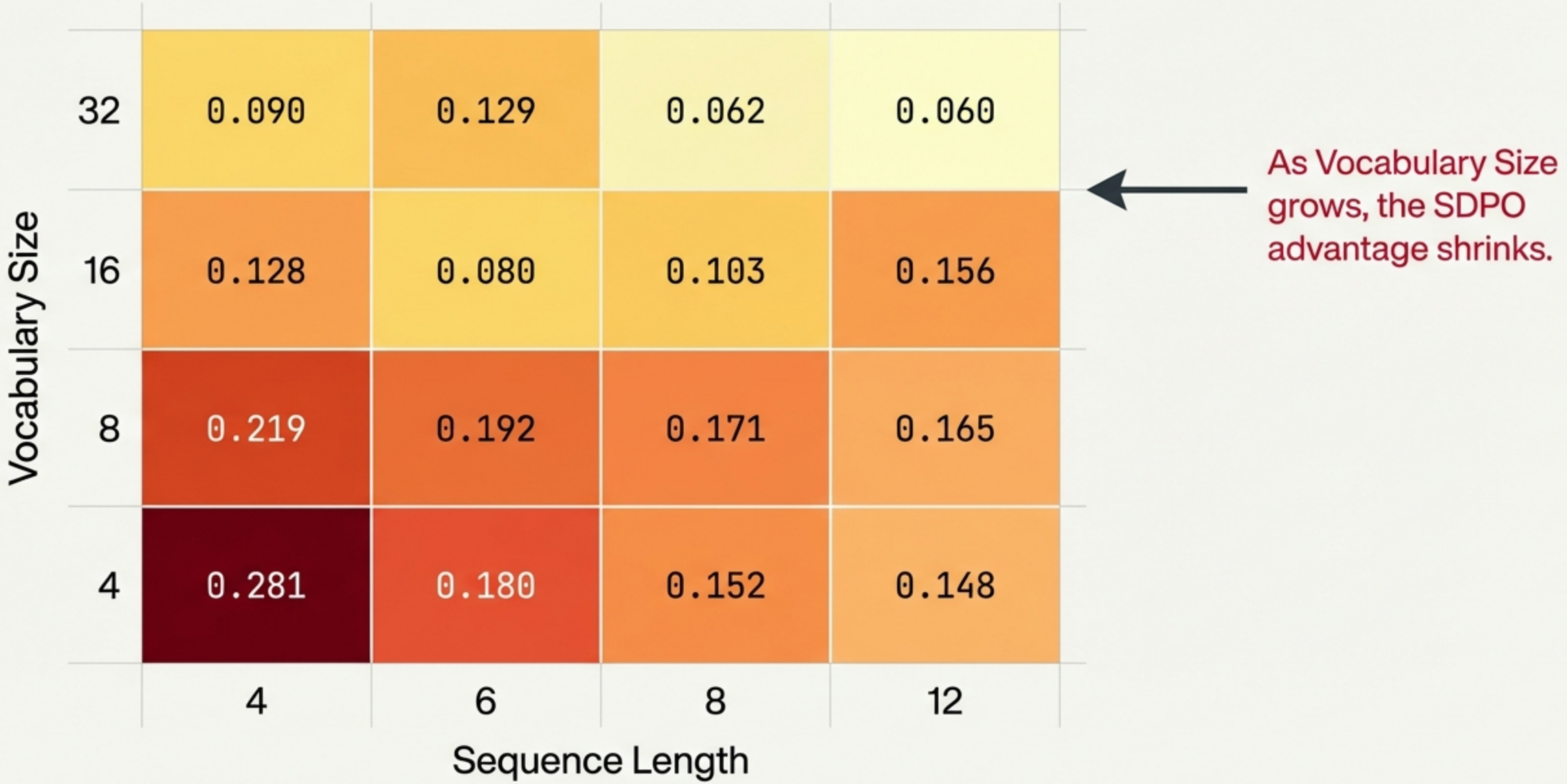


Robustness to Systematic Evaluator Biases

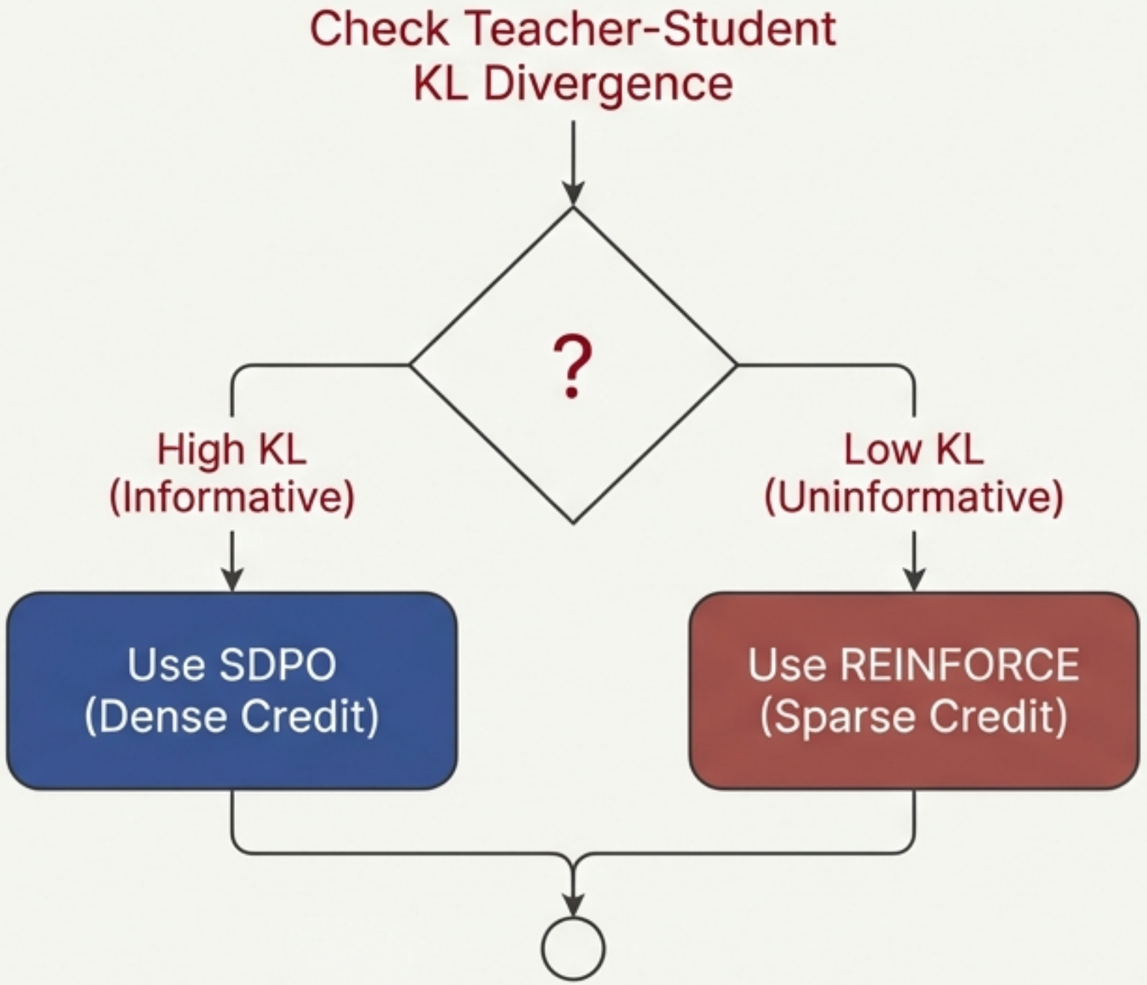


Surprising Result: SDPO's advantage increases under Positivity Bias.

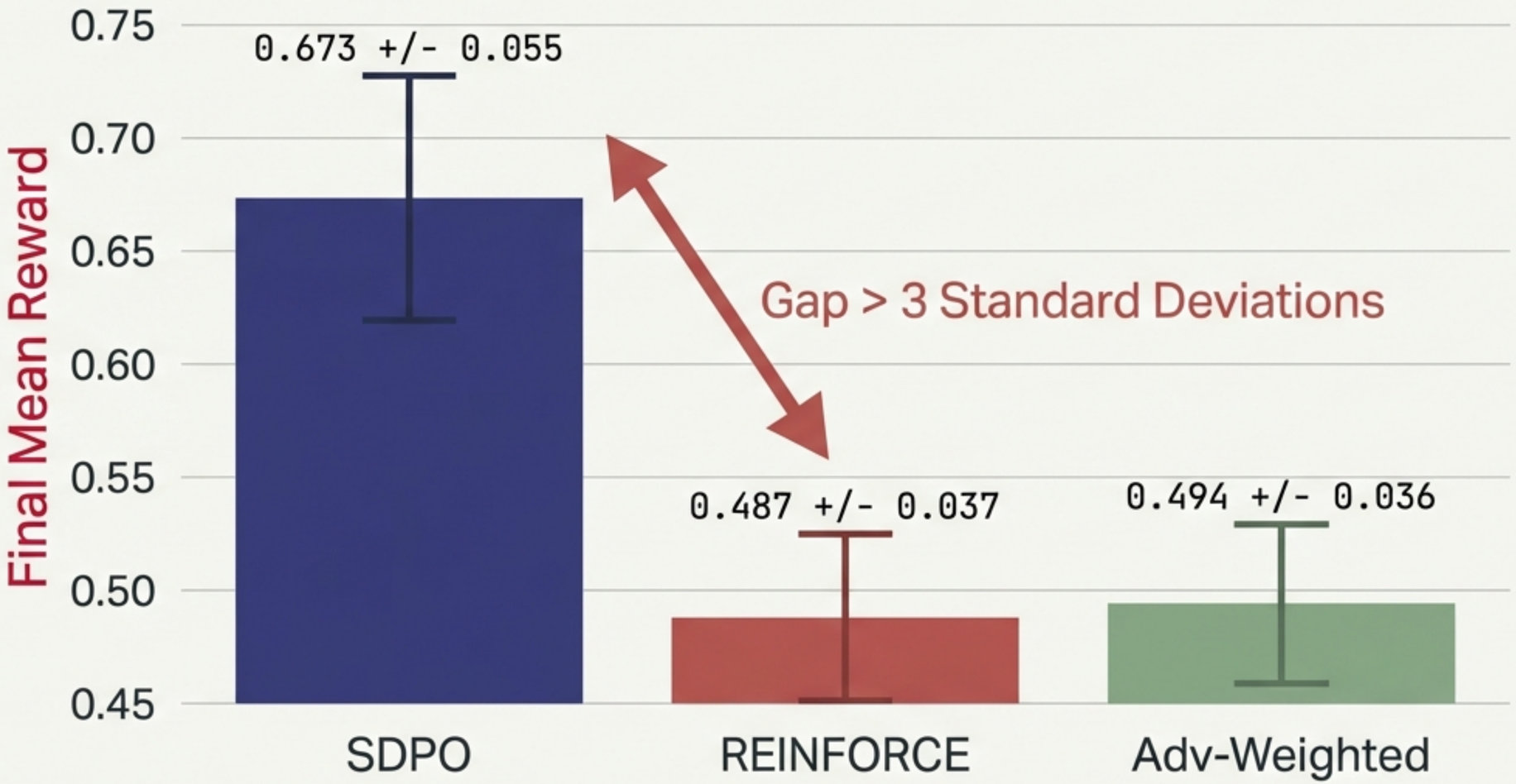
The Scaling Challenge: Advantage Diminishes with Complexity



A Hybrid Method for Heterogeneous Feedback Quality



Statistical Reliability Across Random Seeds



Results aggregated over n=5 random seeds. in Crimson Pro

Summary of Findings

Generalization

Works for open-ended, continuous-reward tasks, not just code.

Dense Credit

Correctly identifies high-value tokens where baselines fail.

Robustness

Handles 50% noise and systematic biases without failure.

Trade-offs

Diversity loss is real but tunable via KL regularization.

Verdict: A Viable Candidate for Open-Ended Alignment

SDPO offers a path away from the “Sparse Reward” trap.

Benchmark Validation
(AlpacaEval, MT-Bench)

Architectural Mods
(Subspace conditioning for large vocabularies)

Hybrid Deployment
(Combining with REINFORCE)

With scaling solutions, dense credit assignment is within reach for open-ended generation.