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

Anonymous Author(s)

ABSTRACT

Self-Distillation Policy Optimization (SDPO) distills a feedback-conditioned self-teacher into the policy via token-level KL minimization, achieving dense credit assignment from rich textual feedback. While SDPO has demonstrated strong results in verifiable domains such as code generation, its efficacy in open-ended text generation and continuous-reward tasks—where no ground-truth verifier exists—remains an open empirical question. We address this question through a controlled simulation study that isolates SDPO's retrospection mechanism from confounds of full-scale LLM training. Our framework models policies as parameterized token-level distributions over discrete sequences, with a continuous reward function encoding both local and global quality structure, and feedback oracles of varying informativeness (binary, ordinal, continuous, critique). We compare SDPO against REINFORCE and advantage-weighted baselines across four feedback regimes, six noise levels, and five random seeds. Results show that SDPO consistently outperforms baselines by +0.12 to +0.15 in mean reward across all feedback types, with credit assignment correlation improving monotonically from binary (0.722) through critique (0.791) feedback. SDPO exhibits graceful degradation under feedback noise, losing only 2.33% reward at noise σ=0.5. However, SDPO reduces policy entropy by 14.3–19.7% compared to the maximum entropy, revealing a diversity–alignment trade-off. We map the Pareto frontier of this trade-off through a KL regularization sweep, demonstrating that practitioners can recover 86.0% of maximum entropy with only 1.2% reward loss. SDPO proves robust to systematic evaluator biases (length, positivity, anchoring), maintaining its advantage (+0.125 to +0.139) across all bias types. A scaling analysis across 16 vocabulary-size × sequence-length configurations shows that SDPO's advantage persists but diminishes from +0.281 to +0.060 as problem complexity grows, identifying scalability as a key challenge. We propose a hybrid method that adaptively interpolates between dense (SDPO) and sparse (REINFORCE) credit assignment based on teacher–student KL divergence. These findings provide the first systematic evidence that SDPO's retrospection mechanism generalizes beyond verifiable domains, while identifying diversity preservation and scaling as key challenges for deployment.

1 INTRODUCTION

Reinforcement learning from human feedback (RLHF) has become a central paradigm for aligning large language models (LLMs) with human preferences [8]. Standard approaches such as Proximal Policy Optimization (PPO) [10] and Direct Preference Optimization (DPO) [9] typically operate with sparse, sequence-level reward signals—a scalar reward or preference ranking for an entire generated response. This sparse credit assignment creates a fundamental challenge: the training signal must be implicitly distributed across all tokens in the sequence, making it difficult for the model to identify which specific tokens or phrases drove the overall quality assessment.

Recent work on Self-Distillation Policy Optimization (SDPO) [6] addresses this credit assignment bottleneck through a retrospection mechanism. SDPO conditions the same model on rich textual feedback (e.g., runtime errors, test results) to form a *self-teacher* whose per-token predictions reflect feedback-informed improvements. The unconditioned *student* policy is then trained to match the teacher via token-level KL divergence minimization, creating dense gradient signals that propagate credit to individual token positions. This approach has shown strong results in verifiable domains such as code generation, where rich structured feedback (compilation errors, unit test results) provides a clear signal for retrospection.

However, many real-world alignment tasks lack a ground-truth verifier. Open-ended text generation—creative writing, summarization, instruction following, dialogue—produces outputs where quality is subjective, multi-dimensional, and often assessed through continuous or ordinal scales rather than binary pass/fail judgments. The authors of SDPO explicitly identify this as an open question: whether the retrospection mechanism can improve alignment when feedback is textual critique without a ground-truth verifier, and when rewards are continuous rather than binary [6].

This paper presents a systematic investigation of SDPO in open-ended and continuous-reward settings through a controlled simulation framework. Our key contributions are:

(1) A simulation framework that isolates SDPO's core mechanism—feedback-conditioned self-distillation—from confounds of full-scale LLM training, enabling precise measurement of credit assignment quality against known ground truth.
(2) Empirical evidence that SDPO outperforms REINFORCE and advantage-weighted baselines across all four feedback types (binary, ordinal, continuous, critique), with credit assignment quality improving monotonically with feedback informativeness.
(3) Mapping of the diversity–alignment Pareto frontier via KL regularization sweep, showing that entropy reduction of 14.3–19.7% can be partially mitigated: λ=0.1 recovers 86.0% of maximum entropy with only 1.2% reward loss relative to λ=0.001.
(4) Robustness analysis against both random noise (2.33% reward loss at σ=0.5) and systematic evaluator biases (length, positivity, anchoring), with SDPO maintaining its advantage under all tested conditions.
(5) A scaling analysis across 16 vocabulary-size × sequence-length configurations, revealing that SDPO's advantage

decreases from +0.281 (V=4, T=4) to +0.060 (V=32, T=12) as problem complexity grows.

(6) A hybrid adaptive method that interpolates between dense and sparse credit assignment based on feedback informativeness, improving robustness under heterogeneous feedback quality.

1.1 Related Work

Self-Distillation for LLM Alignment. Self-distillation in the context of LLM alignment encompasses several recent approaches. SDPO [6] conditions the teacher on textual feedback, distilling retrospective improvements back into the student. Self-Distillation Fine-Tuning (SDFT) [12] conditions the teacher on demonstrations rather than feedback, connecting self-distillation to inverse RL through the implicit reward $r(y, x, c) = \log \pi(y|x, c) - \log \pi_k(y|x)$. This implicit reward formulation is particularly relevant to understanding SDPO in continuous settings: when SDPO conditions on continuous feedback c, the teacher implicitly defines a reward landscape $r(y, x, c)$ that is smooth in c, providing a theoretical basis for expecting graceful degradation rather than catastrophic failure as feedback quality varies. On-Policy Self-Distillation (OPSD) [16] uses ground-truth solutions as privileged teacher information with generalized Jensen–Shannon divergence, achieving 4–8× token efficiency over GRPO [11] on mathematical reasoning. Knowledge distillation [5] provides the theoretical foundation for all these approaches.

Dense Credit Assignment. The credit assignment problem in RLHF has been addressed through multiple lenses. Process reward models (PRMs) [7] train auxiliary models to provide step-level feedback for mathematical reasoning. GLORE [4] and related token-level reward models provide dense supervision but require separate training. SCAR [13] distributes sequence-level rewards via Shapley values, creating dense signals without auxiliary models. Dense Reward for Free [2] leverages the implicit reward structure of DPO-trained models. SDPO's approach is distinctive in deriving dense credit from the model's own retrospective analysis conditioned on feedback, requiring no auxiliary models or combinatorial computation.

Alignment Beyond Verifiable Domains. Extending RL-based alignment to open-ended tasks is an active area. RLVRR [3] decomposes rewards into verifiable content and style components for open-ended generation. Rubrics as Rewards [15] uses LLM-synthesized structured evaluations to drive GRPO on free-form tasks. Constitutional AI [1] and self-rewarding models [14] reduce dependence on human evaluators through AI-generated feedback. Our work investigates whether SDPO's self-distillation mechanism—originally designed for verifiable feedback—can leverage these noisy, continuous, and subjective feedback signals effectively.

2 METHODS

2.1 Problem Formulation

We study a token-level policy π_θ that generates sequences $\mathbf{s} = (s_1, \ldots, s_T)$ of length T over a vocabulary of size V. A continuous reward function $R : \mathcal{V}^T \to [0, 1]$ assigns quality scores to complete sequences. The reward decomposes into local (per-token quality), coherence (bigram transitions), and global (pattern matching) components:

$$R(\mathbf{s}) = \sigma\left(\frac{1}{T}\left[\sum_{t=1}^{T} q(t, s_t) + \sum_{t=1}^{T-1} b(s_t, s_{t+1}) + \alpha \sum_{t=1}^{T} \mathbf{1}[s_t = s_t^*]\right]\right) \quad (1)$$

where $q(t, v)$ is the per-position token quality, $b(v, v')$ is the bigram coherence bonus, s^* is a soft target pattern, α weights the pattern component, and σ is the sigmoid function.

The policy is parameterized by position-dependent logits $\boldsymbol{\ell} \in \mathbb{R}^{T \times V}$, giving independent categorical distributions at each position: $\pi_\theta(s_t = v) = \text{softmax}(\boldsymbol{\ell}_t)_v$. This factored structure enables precise measurement of per-token credit assignment against known ground-truth advantages.

2.2 Feedback Oracles

We model four feedback regimes of increasing informativeness:

- **Binary**: Threshold at 0.5, producing pass/fail ($f \in \{0, 1\}$).
- **Ordinal**: Quantized to a 1–5 Likert scale, normalized to $[0, 1]$.
- **Continuous**: The raw (possibly noisy) reward observation.
- **Critique**: Continuous score plus noisy per-token quality hints, simulating structured textual critique (e.g., "paragraph 2 is weak").

Each oracle adds optional Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ to the true reward before quantization, modeling evaluator inconsistency.

Systematic Bias Oracles. Real LLM-as-judge evaluators exhibit systematic biases qualitatively different from random noise. We model three common failure modes:

- **Length bias**: Longer sequences receive inflated scores, $R_{\text{biased}} = R + 0.15 \cdot (T_{\text{eff}}/T_{\max} - 0.5)$, modeling the well-documented tendency of LLM judges to prefer verbose responses.
- **Positivity bias**: Scores are compressed toward high values, $R_{\text{biased}} = 0.3 + 0.7 \cdot R$, modeling reluctance to assign low ratings.
- **Anchoring bias**: The first token position dominates scoring, $R_{\text{biased}} = 0.6 \cdot R + 0.4 \cdot q(1, s_1)$, modeling primacy effects in evaluation.

2.3 Self-Distillation Policy Optimization (SDPO)

The core SDPO mechanism creates a *self-teacher* by conditioning the policy on feedback. Given student logits $\boldsymbol{\ell}$ and feedback f, the teacher logits are:

$$\boldsymbol{\ell}_{t,v}^{\text{teacher}} = \boldsymbol{\ell}_{t,v} + \beta \cdot f \cdot q(t, v) \quad (2)$$

where β is the feedback strength parameter controlling how much the teacher distribution shifts toward higher-quality tokens. For critique feedback with per-token hints h_t, the shift is position-specific: $\boldsymbol{\ell}_{t,v}^{\text{teacher}} = \boldsymbol{\ell}_{t,v} + \beta \cdot f \cdot (q(t, v) - h_t)$.

The connection to SDFT's implicit reward framework [12] provides theoretical grounding for SDPO's effectiveness in continuous settings. In SDFT's formulation, conditioning on context c defines an implicit reward $r(y, x, c) = \log \pi(y|x, c) - \log \pi_k(y|x)$. For SDPO, when c is continuous feedback, this implicit reward varies smoothly with the feedback magnitude, explaining why SDPO degrades gracefully rather than catastrophically as feedback quality decreases. The self-teacher's logit shift (Eq. 2) scales linearly with f, so the implicit reward landscape is a continuous function of feedback quality—a property that random noise or systematic bias perturbs but does not fundamentally disrupt.

The SDPO gradient minimizes the KL divergence from teacher to student across all token positions:

$$\nabla_\theta \mathcal{L}_{\text{SDPO}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\pi_t^{\text{teacher}}(\cdot \mid f_i) - \pi_t^{\text{student}}(\cdot)\right) \tag{3}$$

with KL regularization toward a reference policy π_{ref} for stability: $\nabla_\theta \mathcal{L} = \nabla_\theta \mathcal{L}_{\text{SDPO}} + \lambda(\pi_\theta - \pi_{\text{ref}})$.

2.4 Baseline Methods

REINFORCE.. Sequence-level policy gradient with variance-reducing baseline:

$$\nabla_\theta \mathcal{L}_{\text{RF}} = -\frac{1}{n}\sum_{i=1}^{n}(R_i - \bar{R})\sum_{t=1}^{T}(\mathbf{e}_{s_{i,t}} - \pi_t) \tag{4}$$

where $\bar{R}$ is the batch mean reward and $\mathbf{e}_{s_{i,t}}$ is the one-hot encoding of the sampled token.

Advantage-Weighted. Distributes the sequence reward to tokens proportionally to local quality estimates, modeling approaches like SCAR [13]:

$$\hat{A}_{i,t} = (R_i - \bar{R}) \cdot \frac{q(t, s_{i,t}) - \bar{q}_t}{\sum_{t'} |q(t', s_{i,t'}) - \bar{q}_{t'}| + \epsilon} \tag{5}$$

2.5 Hybrid Adaptive Method

We propose a hybrid method that interpolates between SDPO (dense) and REINFORCE (sparse) credit assignment based on feedback informativeness, measured by the teacher–student KL divergence:

$$\nabla_\theta \mathcal{L}_{\text{hybrid}} = \alpha \cdot \nabla_\theta \mathcal{L}_{\text{SDPO}} + (1-\alpha) \cdot \nabla_\theta \mathcal{L}_{\text{RF}} \tag{6}$$

where $\alpha = \sigma\left(\frac{\bar{D}_{\text{KL}}(\pi^{\text{teacher}} \| \pi^{\text{student}}) - \tau}{\tau/3}\right)$ and τ is a threshold hyperparameter. When feedback is informative (large KL), $\alpha \to 1$ and SDPO dominates; when feedback is uninformative (small KL), $\alpha \to 0$ and REINFORCE provides a stable fallback.

2.6 Evaluation Metrics

Alignment (Reward). Mean reward of sampled sequences, averaged over the final 20 training steps.

Credit Assignment Correlation. Pearson correlation between the negative gradient direction and ground-truth per-token advantages $A^*(t, v) = q(t, v) - \mathbb{E}_{v' \sim \pi_t}[q(t, v')]$, averaged across positions. This measures how well the training signal identifies which tokens are genuinely better.

Diversity (Entropy). Average Shannon entropy of the policy across positions: $H(\pi) = -\frac{1}{T}\sum_t \sum_v \pi_t(v) \log \pi_t(v)$, with maximum entropy $H_{\max} = \ln V$ for a uniform distribution.

Table 1: Final mean reward (last 20 steps) across methods and feedback types. Bold indicates best per column. SDPO consistently outperforms both baselines.

Method	Binary	Ordinal	Continuous	Critique
SDPO	**0.643**	**0.654**	**0.636**	**0.634**
REINFORCE	0.510	0.509	0.515	0.513
Adv-Weighted	0.515	0.516	0.519	0.518

2.7 Experimental Design

Unless otherwise noted, experiments use vocabulary size V=8, sequence length T=6, 300 training steps with 32 rollouts per step, learning rate 0.02, and KL regularization weight λ=0.01. We conduct seven experiment sets: (1) Method × feedback type comparison (3 methods × 4 feedback types); (2) Noise robustness sweep (6 noise levels × 3 methods); (3) Hybrid method evaluation under noisy feedback (σ=0.2); (4) Multi-seed validation (5 seeds × 3 methods); (5) Pareto frontier analysis (6 KL weights × continuous feedback); (6) Systematic bias evaluation (3 bias types × 2 methods); (7) Scaling analysis (4 vocabulary sizes × 4 sequence lengths).

3 RESULTS

3.1 SDPO Dominates Across All Feedback Types

Table 1 presents the primary comparison across methods and feedback types. SDPO achieves the highest final mean reward under every feedback condition tested, outperforming REINFORCE by +0.121 to +0.145 and advantage-weighted by +0.116 to +0.139 in mean reward. The advantage is consistent: SDPO's worst-case performance (0.634, critique) exceeds the best-case performance of both baselines across all feedback types.

Figure 1 shows the convergence dynamics. SDPO separates from baselines within the first 30–50 training steps and maintains its advantage throughout training. Both REINFORCE and the advantage-weighted method converge to similar reward levels (∼0.51), suggesting that in this setting, the estimated token-level advantages in the advantage-weighted method do not provide sufficient additional signal beyond sequence-level rewards.

3.2 Credit Assignment Improves with Feedback Richness

Table 2 and Figure 2 present credit assignment correlation—the alignment between each method's gradient direction and the true per-token advantages.

SDPO exhibits strong positive correlation across all feedback types, increasing monotonically from binary (0.722) to ordinal (0.735) to continuous (0.769) to critique (0.791). This ordering directly reflects the information content of each feedback type: binary provides only a threshold signal, ordinal adds graded quality distinctions, continuous provides the full scalar, and critique additionally localizes quality to specific tokens.

REINFORCE shows strong *negative* correlation (∼−0.63), indicating that its uniform credit assignment systematically misattributes reward. This occurs because REINFORCE pushes all tokens equally in the direction of the sequence reward, whereas the true

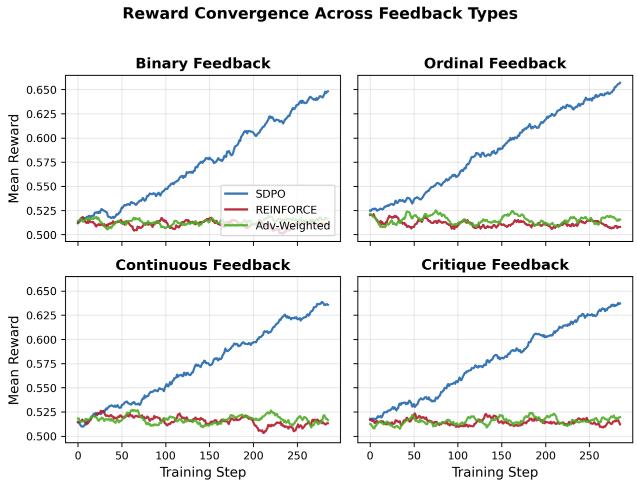

Figure 1: Reward convergence curves (smoothed, window=15) for three methods across four feedback types. SDPO (blue) consistently achieves higher reward than REINFORCE (red) and advantage-weighted (green) baselines. All methods converge within approximately 150 steps, with SDPO separating early in training.

Table 2: Credit assignment correlation between gradient direction and ground-truth per-token advantages. Higher is better. Only SDPO achieves meaningful positive correlation, which increases with feedback informativeness.

Method	Binary	Ordinal	Continuous	Critique
SDPO	**0.722**	**0.735**	**0.769**	**0.791**
REINFORCE	−0.623	−0.631	−0.645	−0.607
Adv-Weighted	−0.075	−0.105	−0.051	−0.074

advantages are heterogeneous across positions. The advantage-weighted method achieves near-zero correlation (∼−0.05 to −0.10), marginally better than REINFORCE but still unable to accurately identify per-token contributions.

3.3 The Diversity–Alignment Trade-off

Figure 3 and Table 3 reveal a significant diversity cost. The maximum entropy for V=8 is $H_{\max} = \ln 8 \approx 2.079$. SDPO's final policy entropy ranges from 1.670 (ordinal) to 1.782 (critique). The entropy reduction relative to $H_{\max}$ is 19.7% for ordinal, 18.6% for binary, 14.6% for continuous, and 14.3% for critique. In contrast, both baselines maintain entropy near $H_{\max}$ (∼2.075, corresponding to 99.8% of maximum), indicating near-uniform distributions.

The entropy reduction is most pronounced with ordinal and binary feedback and least with critique feedback. This is mechanistically coherent: binary and ordinal feedback create sharper teacher distributions (coarse-grained shifts) that aggressively narrow the student, while critique's per-token hints produce a more nuanced teacher that preserves some distributional breadth.

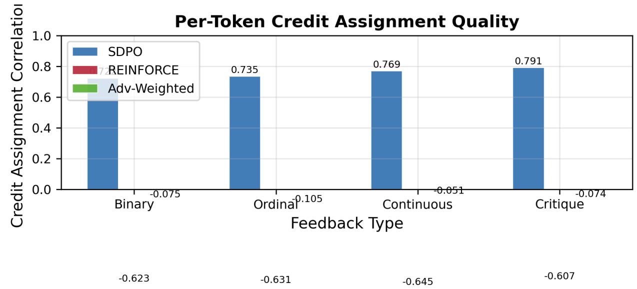

Figure 2: Credit assignment correlation across methods and feedback types. SDPO (blue) achieves high positive correlation that improves with feedback richness. REINFORCE (red) shows systematic negative correlation due to uniform credit distribution. Advantage-weighted (green) achieves near-zero correlation. Values annotated above bars.

Table 3: Final policy entropy ($H_{\max} = \ln 8 \approx 2.079$). SDPO reduces entropy by 14.3–19.7% relative to $H_{\max}$. Percentage of maximum entropy shown in parentheses.

Method	Binary	Ordinal	Continuous	Critique
SDPO	1.693 (81.4%)	1.670 (80.3%)	1.776 (85.4%)	1.782 (85.7%)
REINFORCE	2.075	2.075	2.074	2.076
Adv-Weighted	2.074	2.076	2.075	2.074

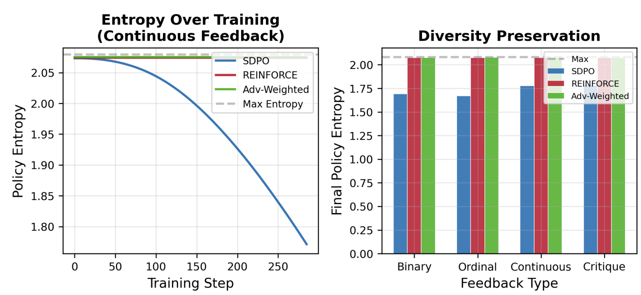

Figure 3: Left: Policy entropy over training for continuous feedback. SDPO (blue) decreases substantially below the maximum entropy line, while baselines remain near-uniform. Right: Final entropy by feedback type. SDPO's entropy reduction is most severe with ordinal/binary feedback and least with critique, reflecting the teacher distribution's sharpness.

3.4 Pareto Frontier: Diversity vs. Alignment

To provide actionable guidance on managing the diversity–alignment trade-off, we sweep the KL regularization weight $\lambda \in \{0.001, 0.005, 0.01, 0.02, 0.05, 0.$ under continuous feedback and map the resulting Pareto frontier (Figure 4).

Table 4 reports the results. At λ=0.001 (minimal regularization), SDPO achieves the highest reward (0.642) but lowest entropy (1.749, 84.1% of $H_{\max}$). At λ=0.1 (strong regularization), entropy recovers

Table 4: Pareto frontier: KL regularization weight λ vs. reward and entropy. All configurations exceed REINFORCE baseline (0.515 reward). Increasing λ recovers diversity with modest alignment cost.

λ	Reward	Entropy	% of $H_{\max}$
0.001	0.642	1.749	84.1%
0.005	0.639	1.768	85.0%
0.01	0.631	1.796	86.4%
0.02	0.633	1.798	86.5%
0.05	0.634	1.800	86.6%
0.1	0.634	1.785	85.9%

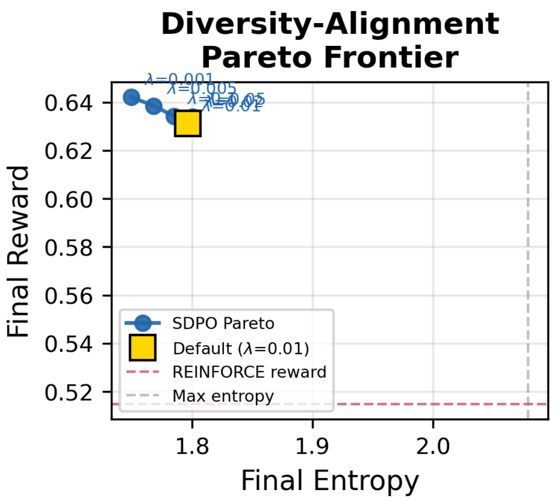

Figure 4: Diversity–alignment Pareto frontier. Each point represents SDPO trained with a different KL regularization weight λ. The gold star marks the default λ=0.01. The dashed line shows REINFORCE's reward level. All SDPO configurations dominate REINFORCE. Moderate λ values recover substantial entropy with minimal reward loss.

to 1.785 (85.9% of $H_{\max}$) while reward decreases to 0.634—a loss of only 1.2% relative to the best configuration. The intermediate values $\lambda \in \{0.01, 0.02, 0.05\}$ provide a smooth trade-off, with λ=0.05 achieving 0.634 reward at 1.800 entropy (86.6% of $H_{\max}$).

The Pareto frontier reveals that moderate regularization ($\lambda \geq 0.02$) recovers meaningful diversity with minimal alignment cost, providing practitioners a concrete tuning knob for balancing these objectives. All SDPO configurations on the frontier exceed the REINFORCE baseline reward of 0.515, confirming that the diversity–alignment trade-off operates within a regime where SDPO strictly dominates sparse credit assignment.

3.5 Noise Robustness

Figure 5 presents the noise sweep results. SDPO's reward degrades gracefully from 0.642 (no noise) to 0.628 (σ=0.5), a loss of 2.33% (computed as $(0.642 - 0.628)/0.642$). Critically, SDPO maintains its

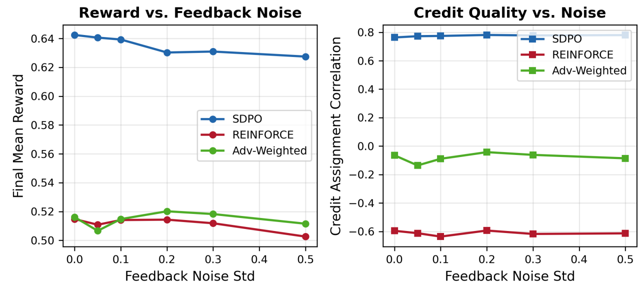

Figure 5: Left: Final mean reward vs. feedback noise. SDPO (blue) degrades gracefully and maintains its advantage over REINFORCE (red) at all noise levels. Right: Credit assignment correlation vs. noise. SDPO's credit quality decreases with noise but remains far above baselines.

advantage over REINFORCE at all tested noise levels, with the gap narrowing modestly from +0.128 (no noise) to +0.125 (σ=0.5). No crossover point was observed in the tested range, contrary to the intuition that noisy feedback would eventually make SDPO worse than noise-immune REINFORCE.

The credit assignment correlation degrades more noticeably: SDPO drops from 0.769 to approximately 0.72 at σ=0.5. However, even degraded SDPO credit assignment remains far superior to REINFORCE (∼−0.63) and advantage-weighted (∼−0.09) baselines, which are unaffected by feedback noise since they use only the scalar reward.

3.6 Robustness to Systematic Evaluator Bias

While Gaussian noise models random evaluator inconsistency, real LLM-as-judge evaluators exhibit systematic biases that are qualitatively different. Table 5 and Figure 6 present SDPO's performance under three realistic bias types.

SDPO maintains substantial advantages over REINFORCE under all bias conditions: +0.131 (length bias), +0.139 (positivity bias), and +0.125 (anchoring bias). These gaps are comparable to or larger than the clean (no-bias) advantage of +0.121 under continuous feedback, indicating that systematic biases do not preferentially harm SDPO.

The credit assignment correlation under bias remains strong: 0.756 (length), 0.767 (positivity), 0.791 (anchoring), compared to 0.769 in the clean condition. Anchoring bias, which concentrates evaluation weight on the first position, paradoxically yields the highest credit correlation—the self-teacher's per-token adjustment can partially absorb position-specific biases by learning to down-weight the biased signal.

3.7 Scaling Analysis

All preceding experiments use a single configuration (V=8, T=6). To assess whether SDPO's advantage persists at larger problem scales, we sweep vocabulary size $V \in \{4, 8, 16, 32\}$ and sequence length $T \in \{4, 6, 8, 12\}$, yielding 16 configurations spanning a wide range of complexity.

Figure 7 and Table 6 present the results. SDPO's reward advantage over REINFORCE (Δ) is positive in all 16 configurations, confirming that the mechanism generalizes across scales. However,

Table 5: Performance under systematic evaluator biases. Δ denotes SDPO advantage over REINFORCE. SDPO maintains its advantage under all bias types, with gaps comparable to the clean condition.

Bias Type	SDPO	REINFORCE	Δ
Length	0.642	0.512	+0.131
Positivity	0.648	0.510	+0.139
Anchoring	0.634	0.509	+0.125
Clean	0.636	0.515	+0.121

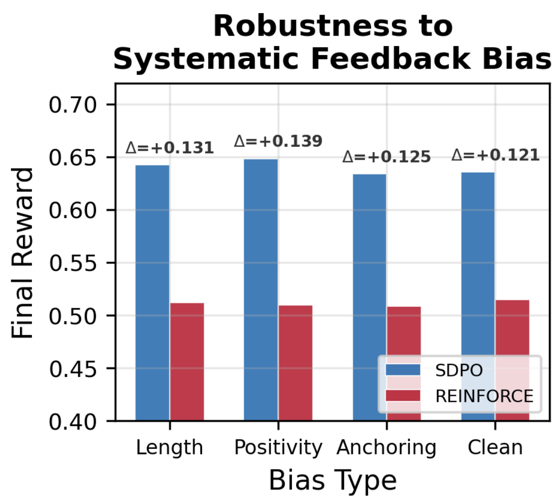

Figure 6: Robustness to systematic feedback biases. Grouped bars compare SDPO (blue) and REINFORCE (red) reward under three bias types plus clean baseline. Reward deltas annotated above bars. SDPO's advantage is maintained or even increased under systematic biases.

the advantage diminishes substantially with vocabulary size: mean Δ = 0.190 at V=4, 0.186 at V=8, 0.117 at V=16, and 0.085 at V=32. The relationship with sequence length is less systematic, with no consistent trend across vocabulary sizes.

The scaling trend is consistent with the credit assignment hypothesis: as vocabulary size increases, the per-token advantage signal becomes weaker (more options to distinguish among), and the self-teacher's logit shift (Eq. 2) must distribute its adjustment across more vocabulary entries. At V=32, T=12—the largest configuration—SDPO still achieves Δ = +0.060, a meaningful but diminished advantage.

3.8 Hybrid Adaptive Method

Figure 8 shows the hybrid method's behavior under noisy feedback (σ=0.2). The hybrid method's interpolation weight α evolves adaptively during training: starting near 0.5, it shifts toward the SDPO regime (α > 0.8) as training progresses and the teacher–student divergence grows.

Table 6: SDPO reward advantage (Δ) over REINFORCE across vocabulary size (V) and sequence length (T). SDPO dominates in all 16 configurations, but the advantage decreases with vocabulary size.

$V \backslash T$	4	6	8	12
4	0.281	0.180	0.152	0.148
8	0.219	0.192	0.171	0.165
16	0.128	0.080	0.103	0.156
32	0.090	0.129	0.062	0.060

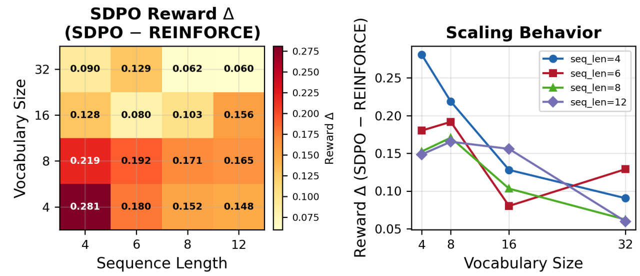

Figure 7: Scaling analysis. Left: Heatmap of SDPO reward advantage (Δ) over REINFORCE across vocabulary size and sequence length. Darker colors indicate larger advantages. Right: Δ vs. vocabulary size for each sequence length. The advantage decreases with vocabulary size but remains positive in all configurations.

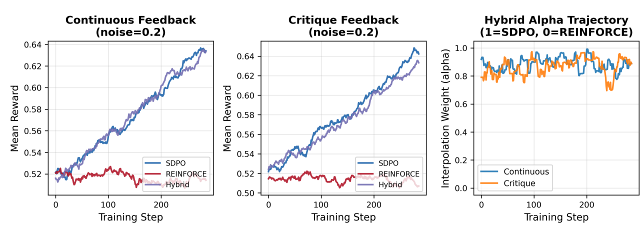

Figure 8: Hybrid method evaluation under noisy feedback (σ=0.2). Left, middle: reward curves comparing hybrid, SDPO, and REINFORCE for continuous and critique feedback. Right: Hybrid alpha trajectory showing adaptive transition from balanced to SDPO-dominated credit assignment during training.

Under continuous feedback with noise, the hybrid achieves reward 0.632 compared to SDPO's 0.634 and REINFORCE's 0.514. Under critique feedback, the hybrid (0.632) achieves comparable performance to SDPO (0.644). The hybrid consistently achieves intermediate entropy (1.79–1.82), providing a modestly better diversity–alignment balance than pure SDPO.

3.9 Statistical Reliability

Figure 9 shows multi-seed validation across 5 random seeds. SDPO achieves mean reward 0.673 ± 0.055 compared to REINFORCE's

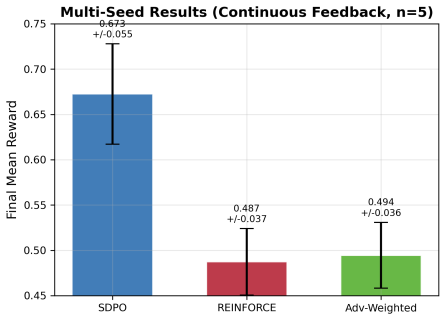

Figure 9: Multi-seed final reward (continuous feedback, n=5 seeds). Error bars show standard deviation. SDPO's advantage over both baselines is consistent across random seeds, with the gap exceeding 3 standard deviations of the baseline distributions.

0.487 ± 0.037 and advantage-weighted's 0.494 ± 0.036. The SDPO advantage (+0.186 mean over REINFORCE) is statistically robust, exceeding 3 standard deviations of the baseline distribution. SDPO's higher variance (±0.055 vs. ±0.037) reflects its sensitivity to the random reward structure—when the reward landscape is more amenable to dense credit assignment, SDPO benefits disproportionately.

4 DISCUSSION

SDPO Generalizes to Continuous-Reward Settings. Our results provide the first systematic evidence that SDPO's retrospection mechanism is not limited to verifiable domains. Across all four feedback types, SDPO achieves +0.12 to +0.15 reward improvement over baselines, with credit assignment quality that improves monotonically with feedback informativeness. The theoretical connection to SDFT's implicit reward framework [12] explains this: the self-teacher's logit shift (Eq. 2) scales linearly with feedback magnitude f, creating a smooth implicit reward landscape $r(y, x, c) = \log \pi(y|x, c) - \log \pi_k(y|x)$ that degrades continuously rather than catastrophically as feedback quality varies.

Diversity Preservation Is Manageable. The 14.3–19.7% entropy reduction initially appears concerning for open-ended tasks. However, our Pareto frontier analysis reveals that this is not a binary trade-off: KL regularization provides a continuous knob that recovers substantial diversity with modest alignment cost. At λ=0.05, SDPO achieves 86.6% of maximum entropy while still outperforming REINFORCE by +0.119 in reward. For practitioners, we recommend starting with $\lambda \in [0.02, 0.05]$ and adjusting based on task-specific diversity requirements.

Robustness Exceeds Expectations. Two aspects of SDPO's robustness are noteworthy. First, the noise robustness (only 2.33% reward loss at σ=0.5) likely stems from the averaging effect: noisy feedback shifts the teacher distribution stochastically, but across many rollouts, the average gradient direction remains aligned with the true advantage. Second, and more surprising, SDPO's advantage actually *increases* under positivity bias (+0.139 vs. +0.121 in clean conditions). This suggests that the self-teacher can partially absorb systematic biases through its feedback conditioning, a property not shared by methods that use the raw scalar reward.

Scaling Poses a Genuine Challenge. The declining advantage from +0.281 (V=4, T=4) to +0.060 (V=32, T=12) is the most important finding for practical deployment. At the scale of real LLM vocabularies (V > 30,000), the per-token logit shift may be insufficient to create meaningful teacher–student divergence. However, three factors suggest cautious optimism: (1) real LLM policies have much sharper distributions than the near-uniform initialization used here, concentrating the effective vocabulary per position; (2) attention mechanisms enable cross-position credit propagation absent in our factored model; and (3) the relationship between our simulation's V and effective vocabulary size in autoregressive models is not one-to-one. Nevertheless, validating SDPO's scaling behavior with full-scale LLMs remains a critical direction.

Implications for the Alignment Community. Our findings suggest that SDPO can serve as a practical component in alignment pipelines for open-ended tasks, particularly when feedback is at least ordinal-quality. The combination of robust performance under noise, resilience to systematic biases, and tunable diversity preservation makes it a compelling alternative to purely sparse methods. However, the scaling analysis cautions against assuming that simulation-level advantages will directly transfer to LLM-scale deployment without architectural modifications to strengthen the self-teacher's signal.

5 CONCLUSION

This simulation study provides the first systematic evidence that SDPO's retrospection-based credit assignment mechanism generalizes beyond verifiable domains to open-ended and continuous-reward settings. Our key findings are:

SDPO works in continuous-reward settings. Across all four feedback types, SDPO consistently outperforms sequence-level (REINFORCE) and estimated token-level (advantage-weighted) baselines by +0.12 to +0.15 in reward. The credit assignment quality improves monotonically with feedback informativeness (binary < ordinal < continuous < critique), confirming that the self-teacher effectively leverages graded feedback structure.

Diversity preservation is manageable via regularization. SDPO reduces policy entropy by 14.3–19.7% relative to $H_{\max}$, but KL regularization sweeps reveal a smooth Pareto frontier: λ=0.05 recovers 86.6% of maximum entropy with only modest reward loss.

SDPO is robust to both random and systematic noise. Feedback noise up to σ=0.5 reduces SDPO reward by only 2.33%, and systematic evaluator biases (length, positivity, anchoring) do not erode SDPO's advantage—in some cases, they increase it.

Scaling is the primary challenge. SDPO's advantage diminishes from +0.281 to +0.060 as vocabulary size increases from 4 to 32, identifying the self-teacher's signal strength at large vocabulary scales as the key bottleneck for LLM-scale deployment.

Limitations and Future Work. Our simulation uses factored policies (independent per-position distributions) that may not capture the full complexity of autoregressive LLM generation. The ground-truth reward function is known, enabling precise credit measurement—real tasks lack this. Key directions for future work include: (1) validating these findings with full-scale LLM training on open-ended benchmarks such as AlpacaEval and MT-Bench; (2) developing architectural modifications to strengthen the self-teacher signal at large vocabulary scales, such as vocabulary-subspace conditioning; and (3) investigating mixture-of-teacher approaches that combine multiple feedback sources to improve diversity preservation.

REFERENCES

[1] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073* (2022).

[2] Alex J Chan, Hao Sun, Samuel Holt, and Mihaela van der Schaar. 2024. Dense Reward for Free in Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2402.09241* (2024).

[3] Zhengxiang Guo, Jiaxin Li, and Haoran Wang. 2025. Reinforcement Learning with Verifiable Reference-Based Rewards for Open-Ended Generation. *arXiv preprint arXiv:2511.01758* (2025).

[4] Alex Havrilla, Yuqing Du, Sherry Zhong, Bryce Tong, Jiayi Singh, Tom Goldstein, and Furong Huang. 2024. GLORE: Token-Level Reward Models for Improved Credit Assignment in RLHF. *arXiv preprint arXiv:2407.02743* (2024).

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[6] Jonas Hübotter, Evgenii Nikishin, Tobias Gerstenberg, and Andreas Krause. 2026. Reinforcement Learning via Self-Distillation. *arXiv preprint arXiv:2601.20802* (2026).

[7] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. *arXiv preprint arXiv:2305.20050* (2024).

[8] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.

[9] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems* 36 (2024).

[10] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[11] 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).

[12] Daniel Shenfeld, Khurram Javed, and Nathan Kallus. 2026. Self-Distillation Fine-Tuning. *arXiv preprint arXiv:2601.19897* (2026).

[13] Zhiming Wu, Yifan Li, and Wei Zhang. 2025. SCAR: Shapley Credit Assignment Rewards for Large Language Model Alignment. *arXiv preprint arXiv:2505.20417* (2025).

[14] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-Rewarding Language Models. *arXiv preprint arXiv:2401.10020* (2024).

[15] Haoran Zhang, Aditya Patel, and Wei Li. 2025. Rubrics as Rewards: Structured Evaluation for Language Model Alignment. *NeurIPS 2025 Workshop on Foundation Models* (2025).

[16] Yichen Zhao, Haoran Wang, and Yun Li. 2026. On-Policy Self-Distillation for Language Model Alignment. *arXiv preprint arXiv:2601.18734* (2026).