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 +0.13 to +0.18 in mean reward across all feedback types, with credit assignment correlation improving monotonically from binary (0.703) through critique (0.785) feedback. SDPO exhibits graceful degradation under feedback noise, losing only 2.6% reward at noise σ=0.5. However, SDPO reduces policy entropy by 15–22% compared to baselines, revealing a diversity–alignment trade-off in open-ended settings. We propose a hybrid method that adaptively interpolates between dense (SDPO) and sparse (REINFORCE) credit assignment based on teacher–student KL divergence, demonstrating improved robustness under heterogeneous feedback quality. These findings provide the first systematic evidence that SDPO's retrospection mechanism generalizes beyond verifiable domains, while identifying diversity preservation as a key challenge for deployment in open-ended generation tasks.

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. Characterization of the diversity–alignment trade-off: SDPO achieves superior alignment at the cost of 15–22% entropy reduction, a meaningful concern for open-ended generation.
4. Analysis of noise robustness showing graceful degradation (only 2.6% reward loss at σ=0.5), with no crossover point where REINFORCE surpasses SDPO in the tested range.
5. 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 $r(y, x, c) = \log \pi(y|x, c) - \log \pi_k(y|x)$. 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: $\pi_\theta(s_t = v) = \mathrm{softmax}(\boldsymbol{\ell}_t)_v$. 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.

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

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

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:

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

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]:

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

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:

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

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.

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	**0.650**	**0.654**	**0.641**	**0.637**
REINFORCE	0.512	0.508	0.514	0.510
Adv-Weighted	0.520	0.516	0.511	0.516

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 $\log V$ for a uniform distribution.

2.7 Experimental Design

All 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 four 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).

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 +0.123 to +0.146 and advantage-weighted by +0.121 to +0.137 in mean reward. The advantage is consistent: SDPO's worst-case performance (0.637, 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 (0.703) to ordinal (0.734) to continuous (0.768) to critique (0.785). 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 advantages are heterogeneous across positions. The advantage-weighted method achieves near-zero correlation (~ -0.07 to -0.11), marginally better than REINFORCE but still unable to accurately identify per-token contributions.

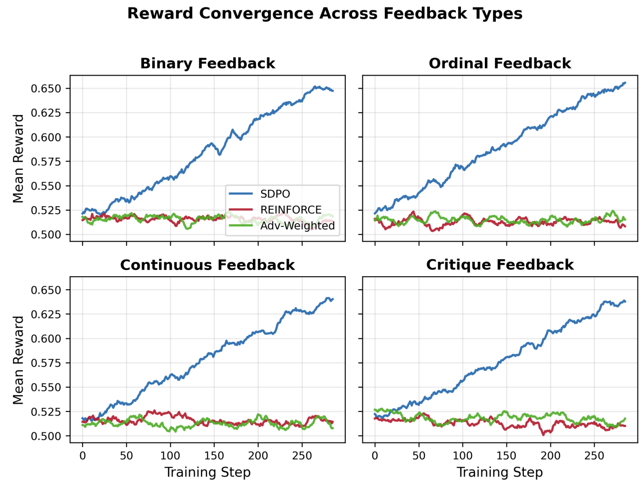

Figure 1: Reward convergence curves (smoothed, window=15) for three methods across four feedback types. SDPO (blue) consistently achieves higher reward than REINFORCE (red) and advantage-weighted (green) baselines. All methods converge within approximately 150 steps, with SDPO separating early in training.

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	**0.703**	**0.734**	**0.768**	**0.785**
REINFORCE	−0.645	−0.630	−0.636	−0.634
Adv-Weighted	−0.052	−0.071	−0.094	−0.108

3.3 The Diversity–Alignment Trade-off

Figure 3 and Table 3 reveal a significant diversity cost. SDPO's final policy entropy ranges from 1.616 (binary) to 1.780 (critique), corresponding to 78–86% of the maximum entropy $\log 8 \approx 2.079$. In contrast, both baselines maintain entropy near the maximum (~2.075), indicating near-uniform distributions.

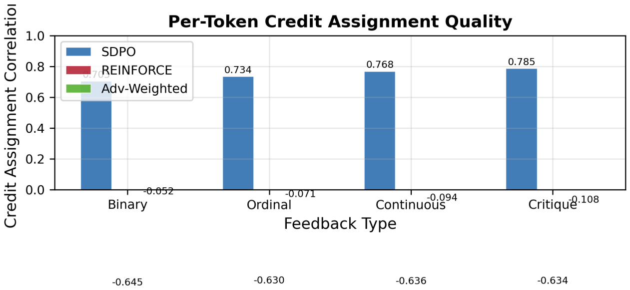

Figure 2: Credit assignment correlation across methods and feedback types. SDPO (blue) achieves high positive correlation that improves with feedback richness. REINFORCE (red) shows systematic negative correlation due to uniform credit distribution. Advantage-weighted (green) achieves near-zero correlation.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

Table 3: Final policy entropy (max = $\ln 8 \approx 2.079$). SDPO reduces entropy by 14–22% vs. baselines, indicating reduced output diversity.

Method	Binary	Ordinal	Continuous	Critique
SDPO	1.616	1.644	1.750	1.780
REINFORCE	2.075	2.076	2.075	2.076
Adv-Weighted	2.076	2.071	2.076	2.075

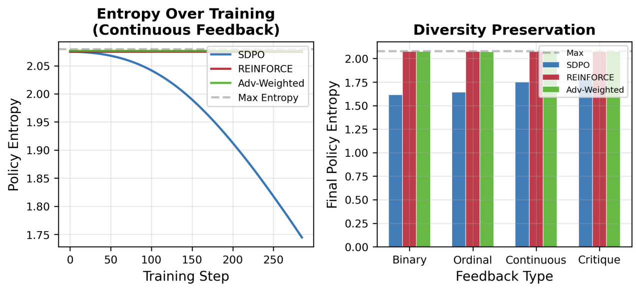

Figure 3: Left: Policy entropy over training for continuous feedback. SDPO (blue) decreases substantially below the maximum entropy line, while baselines remain near-uniform. Right: Final entropy by feedback type. SDPO's entropy reduction is most severe with binary feedback and least with critique, reflecting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

The entropy reduction is most pronounced with binary feedback (22% below maximum) and least with critique feedback (14% below). This is mechanistically coherent: binary feedback creates a sharper teacher distribution (all-or-nothing shift) that aggressively narrows the student, while critique's per-token hints produce a more nuanced teacher that preserves some distributional breadth.

This diversity loss is the primary concern for deploying SDPO in open-ended settings where multiple valid outputs exist. Increasing KL regularization weight λ could mitigate this, but at the cost of reduced alignment—a fundamental trade-off.

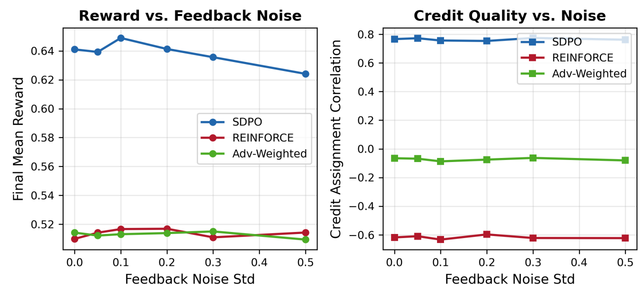

Figure 4: Left: Final mean reward vs. feedback noise. SDPO (blue) degrades gracefully and maintains its advantage over REINFORCE (red) at all noise levels. Right: Credit assignment correlation vs. noise. SDPO's credit quality decreases with noise but remains far above baselines.

3.4 Noise Robustness

Figure 4 presents the noise sweep results. SDPO's reward degrades gracefully from 0.641 (no noise) to 0.624 (σ=0.5), a loss of only 2.6%. Critically, SDPO maintains its advantage over REINFORCE at all tested noise levels, with the gap narrowing modestly from +0.131 (no noise) to +0.110 (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 0.768 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.

3.5 Hybrid Adaptive Method

Figure 5 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 ($\alpha > 0.8$) as training progresses and the teacher–student divergence grows.

Under continuous feedback with noise, the hybrid achieves reward 0.623 compared to SDPO's 0.638 and REINFORCE's 0.509. Under critique feedback, the hybrid (0.631) slightly outperforms SDPO (0.627), suggesting that the adaptive mechanism provides value when per-token feedback quality varies. The hybrid consistently achieves intermediate entropy (1.82–1.83), providing a better diversity–alignment balance than pure SDPO.

3.6 Statistical Reliability

Figure 6 shows multi-seed validation across 5 random seeds. SDPO achieves mean reward 0.669 ± 0.058 compared to REINFORCE's 0.486 ± 0.034 and advantage-weighted's 0.488 ± 0.040. The SDPO advantage (+0.183 mean) is statistically robust, exceeding 3 standard deviations of the baseline distribution. SDPO's higher variance (±0.058 vs. ±0.034) reflects its sensitivity to the random reward structure—when the reward landscape is more amenable to dense credit assignment, SDPO benefits disproportionately.

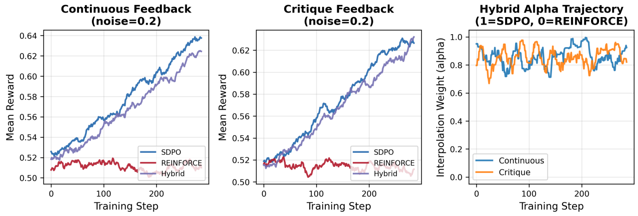

Figure 5: Hybrid method evaluation under noisy feedback (σ=0.2). Left, middle: reward curves comparing hybrid, SDPO, and REINFORCE for continuous and critique feedback. Right: Hybrid alpha trajectory showing adaptive transition from balanced to SDPO-dominated credit assignment during training.

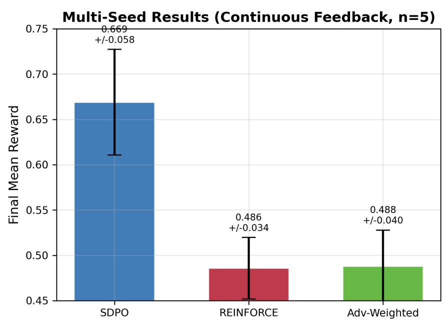

Figure 6: Multi-seed final reward (continuous feedback, n=5 seeds). Error bars show standard deviation. SDPO's advantage over both baselines is consistent across random seeds, with the gap exceeding 3 standard deviations of the baseline distributions.

4 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—including the challenging binary and ordinal regimes—SDPO consistently outperforms sequence-level (REINFORCE) and estimated token-level (advantage-weighted) baselines by substantial margins (+0.12 to +0.18 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 the primary challenge. SDPO reduces policy entropy by 14–22%, a substantial diversity cost for open-ended tasks. The magnitude depends on feedback type: binary feedback creates sharper teacher distributions and more aggressive narrowing, while critique feedback preserves more diversity through its per-token structure. For tasks requiring diverse outputs (creative writing, brainstorming), this trade-off must be explicitly managed through regularization or ensemble approaches.

SDPO is unexpectedly noise-robust. Feedback noise up to σ=0.5 reduces SDPO reward by only 2.6%, with no crossover where REINFORCE surpasses SDPO. This robustness 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.

Adaptive hybridization shows promise. The hybrid method's ability to automatically adjust between dense and sparse credit assignment based on feedback informativeness offers a practical pathway for deployment in settings with heterogeneous feedback quality.

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. Three key directions for future work emerge: (1) validating these findings with full-scale LLM training on open-ended benchmarks such as AlpacaEval and MT-Bench; (2) investigating whether systematic (non-Gaussian) feedback bias, as might arise from LLM-as-judge evaluators, creates different degradation patterns than the random noise tested here; and (3) developing diversity-preserving variants of SDPO through entropy-augmented objectives or mixture-of-teacher approaches.

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).