Grenze International Journal of Engineering and Technology, June Issue

Policy-Induced Distributional Shift in RLHF: A Real-Time Monitoring and Penalization Framework for PPO Training

Rashmi S R[1], Burhaan-U-Deen Wani[2] and Krishnan R[3]
[1-3]Dayananda Sagar College of Engineering/Department of Computer Science & Engineering, Bengaluru, India
Email: Rashmi-cs@dayanandasagar.edu, burhaanwani823@gmail.com, krishnan-cs@dayanandasagar.edu

***Abstract*— Reinforcement learning from human feedback (RLHF) remains vulnerable to reward hacking, where policies exploit imperfections in the proxy reward model rather than im-proving genuine response quality. Existing mitigations—such as KL regularization, re-ward shaping, ensembling, and behavior-supported constraints—typically lack a direct step-wise signal of emerging overoptimization. We investigate whether policy-induced dis-tributional shift in the reward model's latent representation space can serve as such a sig-nal. We propose DSP-PPO (Distributional Shift Penalized PPO), which computes a PCA-Mahalanobis distance from the reward model's in-distribution latent support and subtracts a graded penalty from the PPO reward. The method requires no architectural changes, no reward ensembles, and only a one-time statistics pass. On Anthropic HH-RLHF using GPT-2 and an Open-Assistant DeBERTa-v3-large reward model, the shift signal corre-lates significantly with reward-model scores ($r = 0.3036$, $p = 1.25\times10^{-5}$) and temporally precedes proxy-reward inflation by 15 PPO steps ($r = 0.1825$, $p = 0.0341$), offering a pro-spective early-warning indicator absent in KL regularization alone. At $\lambda = 0.001$, DSP-PPO matches KL-only proxy reward over the final 20 steps while providing a continuous latent-drift trace at only 0.22% per-step overhead. DSP-PPO offers a practical real-time monitoring framework for standard RLHF pipelines and motivates future causal valida-tion and independent alignment evaluation.**

***Index Terms*— RLHF, reward hacking, distributional shift, PPO, reward models, Ma-halanobis distance, AI alignment.**

I. INTRODUCTION

Reinforcement learning from human feedback (RLHF) aligns language models by training a reward model on preference data and optimizing a policy against this learned proxy. While this pipeline has produced strong instruction-following systems, it remains vulnerable to reward overoptimization: poli-cies can exploit imperfections in the proxy reward model, generating high-scoring responses that are not genuinely better [1]–[6]. A key insight from prior work is that reward models are reliable primarily near the distribution of their training data. The standard PPO safeguard—KL regularization against a reference policy—measures divergence in token space but does not directly detect when responses enter poorly supported regions of the reward model's representation space.

Grenze ID: 01.GIJET.12.2.585

Recent representation-level approaches, including hidden-state regulari-zation [8], latent outlier detection via InfoRM [9], and behaviour-supported constraints [2], motivate a sharper question: can policy-induced latent drift be measured continuously during PPO and used as a graded training signal?

We answer this with DSP-PPO (Distributional Shift Penalized PPO). The method constructs a compact statistical model of the reward model's latent states from chosen preference responses before training. During PPO, each generated response is passed through the frozen reward model, projected into a PCA subspace, and scored via Mahalanobis distance from the in-distribution support. This shift score serves as both a real-time monitor and a small additive penalty in the PPO reward.

The design is lightweight: it requires no architectural changes, no reward ensembles, and only a one-time statistics pass.

This paper makes four contributions:

- It defines policy-induced distributional shift in reward-model latent space and implements a real-time PCA-Mahalanobis monitor.
- It introduces a graded DSP-PPO penalty compatible with standard PPO.
- It provides empirical evidence that the measured shift signal temporally precedes proxy-reward inflation.
- It transparently reports limitations: at λ=0.001, DSP-PPO matches (but does not outperform) KL-only proxy reward, with independent alignment-quality evaluation left as future work.

II. Related Work

Reward overoptimization. Prior work has extensively studied reward hacking in RLHF. Rafailov et al. [1] demonstrate scaling laws for overoptimization in direct alignment algorithms, even when bypassing explicit reward models.

Kwa et al. [4] show that KL regularization can fail under heavy-tailed reward misspecification, motivating complementary representation-level safeguards. Fu et al. [5] and Khalaf et al. [6] analyse reward shaping and inference-time mitigation strategies, respectively. Collectively, these studies establish that proxy reward alone is an unreliable measure of true alignment quality.

Representation and distributional shift. Several works highlight the importance of representation-level signals. Kirk et al. [7] show that RLHF improves OOD generalization while reducing output diversity. Yang et al. [8] regularize hidden states during reward-model training for better generalization to evolv-ing policies. Miao et al. [9] propose InfoRM, which uses a variational information bottleneck to detect latent outliers as a post-hoc overoptimization signal.

Wang et al. [10] study reward-model robustness via contrastive and meta-learning techniques. DSP-PPO differs by treating the frozen reward model's latent support as an online monitoring and penalization signal during PPO, rather than a training-time modification or post-hoc diagnostic.

Mitigation methods. BSPO [2] applies a binary behaviour-support constraint but requires continuous offline corpus access.

Liu et al. [3] show that SFT loss implicitly regularizes static dataset shift, yet cannot address dynamic drift during RL. Other approaches include ODIN's disentangled reward heads [11], causal data augmentation for robust reward models (RRM) [12], and Wasserstein DRO for DPO [13].

DSP-PPO is complementary: it adds a lightweight, graded online penalty that layers directly onto standard PPO without retraining the reward model or incurring ensemble costs. Table I positions DSP-PPO against representative methods across key dimensions.

TABLE I. COMPARISON OF REPRESENTATIVE PRIOR WORK AND GAPS RELATIVE TO THE PROPOSED FRAMEWORK

Method	Online	Graded	No Offline	No Ensemble
KL regularization [4]	✓	✓	✓	✓
BSPO [2]	✓	×	×	✓
InfoRM diagnostic [9]	×	✓	✓	✓
SFT adversarial reg. [3]	×	✓	✓	✓
Yang et al. [8]	×	~	✓	✓
ODIN [11]	×	~	✓	✓
RRM [12]	×	×	✓	✓
Wasserstein DRO [13]	×	~	✓	✓
DSP-PPO (Ours)	✓	✓	✓	✓

III. METHODOLOGY

A. Problem Formulation

Let Xdenote prompts and Ydenote responses. A reward model $r_\phi: X \times Y \to \mathbb{R}$ is trained on preference triples $\mathcal{D}_{pref} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^N$ with the Bradley-Terry loss:

$$\mathcal{L}_{\mathrm{RM}}(\phi) = -\mathbb{E}\log\sigma(\mathrm{r}_\phi(\mathrm{x},\mathrm{y}^+) - \mathrm{r}_\phi(\mathrm{x},\mathrm{y}^-)). \qquad (1)$$

PPO then optimizes the policy objective with a KL penalty against a reference policy:

$$\mathrm{J}_{\mathrm{PPO}}(\theta) = \mathbb{E}[\mathrm{r}_\phi(\mathrm{x},\mathrm{y})] - \beta\,\mathrm{KL}(\pi_\theta(\cdot|\,\mathrm{x}) \parallel \pi_{\mathrm{ref}}(\cdot|\,\mathrm{x})). \qquad (2)$$

Definition 1 (Policy-Induced Distributional Shift). Let $f_\phi(x, y) \in \mathbb{R}^d$ be the reward model's penultimate representation. Let $Z_{\mathrm{ID}} = \{f_\phi(x_i, y_i^+)\}_{i=1}^N$ be the latent support induced by chosen training responses. At PPO step t, the generated response $y_t$has latent state $z_t = f_\phi(x, y_t)$. Policy-induced distributional shift is the step-dependent deviation of z_t from the statistical structure of Z_{ID}.

B. Theoretical Motivation

We use Mahalanobis distance because it accounts for the full covariance structure of the latent space, is invariant to linear transformations, and yields a statistically interpretable scalar (approximately χ^2-distributed under Gaussian assumptions).

The raw 1024-dimensional latent space with only $N = 500$in-distribution samples makes the covariance matrix rank-deficient and ill-conditioned. We therefore apply PCA to $d' = 64$components (retaining 99.2% variance), reducing the condition number from $> 10^6$to 4,560 after Tikhonov regularization. This choice was empirically validated for stability and correlation preservation.

C. PCA-Mahalanobis Shift Score

The reward-model hidden state has dimension $d = 1024$ while the in-distribution sample used for support estimation has N=500 chosen responses. We project ZID into a $d' = 64$ dimensional PCA subspace retaining 99.20% of variance. Let z' t be the projected generated-response state, and let μ' and Σ' be the mean and covariance of the projected in-distribution states. We apply Tikhonov regularization:

$$\Sigma_\varepsilon = \Sigma' + \varepsilon I, \varepsilon = 10^{-6}. \qquad (3)$$

The scalar shift score is the PCA-Mahalanobis quadratic form:

$$S_t = (z' - \mu')^\top \Sigma_\varepsilon^{-1}(z' - \mu'). \qquad (4)$$

In the final run, the PCA covariance condition number was 4560.13, confirming numerical stability after regularization. The computation requires no access to D_{pref} during training and no modifications to either model.

D. DSP-PPO Objective

DSP-PPO modifies only the scalar reward passed to PPO:

$$\tilde{r}(x, y_t) = r_\phi(x, y_t) - \lambda S_t, \qquad (5)$$

yielding the optimized objective

$$J_{\mathrm{DSP}}(\theta) = \mathbb{E}[\tilde{r}(x, y_t)] - \beta\,\mathrm{KL}(\pi_\theta(\cdot|\,x) \parallel \pi_{\mathrm{ref}}(\cdot|\,x)). \qquad (6)$$

The coefficient $\lambda = 0.001$is kept deliberately conservative, treating DSP-PPO primarily as monitoring plus light penalization.

E. Implementation and Complexity

Algorithm 1 — DSP PPO (high level)

- Build $\mathrm{Z_{ID}} = \{\mathrm{f}_\phi(\mathrm{x_i}, \mathrm{y}^+)\}$ from chosen responses (one-time statistics pass).
- Fit PCA on $\mathrm{Z_{ID}}$; compute $\mu'$and Σ'; cache PCA projection and Σ_ε^{-1}.
- Initialize $\pi_\theta \leftarrow \pi_{\mathrm{ref}}$.
- For each PPO step t: sample prompts $\mathrm{x_i}$, generate $\mathrm{y_i} \sim \pi_\theta(\cdot|\,\mathrm{x_i})$; compute $\mathrm{r_i} = \mathrm{r}_\phi(\mathrm{x_i}, \mathrm{y_i})$; compute $\mathrm{S_i}$via forward hook and Eq. (4); set $\tilde{\mathrm{r}}_\mathrm{i} = \mathrm{r_i} - \lambda \mathrm{S_i}$; update $\pi_\theta$with clipped PPO using KL coefficient β.
- Return π_θ.

Latent states are captured via a registered forward hook on the reward model's pooler layer (no extra forward pass). The only per-step overhead is the PCA projection and quadratic form (0.22% measured, 0.1636 ms vs. 74.4 ms for reward-model inference). Rewards are normalized and clipped for stability. No access to $D_{pref}$is needed during training.

IV. Experiments

A. Setup

Experiments use the Anthropic HH-RLHF dataset (4,951 training and 496 test preference pairs). The policy is GPT-2 (124M parameters) and the reward model is Open-Assistant DeBERTa-v3-large (435M parameters), with penultimate activations captured via a forward hook. The reward model achieves 76.0% preference accuracy on the held-out set.

PPO configuration: learning rate 2×10^(-6), batch size 4 (mini-batch 2), maximum new tokens 64, β=0.1, λ=0.001, T=150steps (seed 42). Rewards are normalized and clipped for stability. We log proxy reward, shift score, KL divergence, and response length at every step.

B. Baselines and Stability Guards

We compare against:

- KL-Only (λ=0),
- BSPO-Style: binary penalty of 1.0 when shift exceeds the 90th percentile of in-distribution scores (requires offline data),
- Simulated Ensemble: Gaussian perturbation (σ=0.1).

All runs share an instability guard that saves the stable partial log upon severe objective degradation. BSPO-Style terminated at step 125/150.

V. Results and Analysis

A. Main Results

Table II reports final-window metrics (last 20 steps, 95% bootstrap CIs). At λ=0.001, DSP-PPO matches KL-Only proxy reward exactly. This is a transparent negative result: the light penalty does not improve proxy reward over the conservative KL baseline but adds a continuous latent-drift trace at 0.22% overhead. BSPO-Style and the simulated ensemble reach slightly higher proxy reward, but BSPO terminates early (step 125/150) under the shared instability guard.

TABLE II. Main Results on Anthropic HH-RLHF. Metrics Are Averaged over the Last 20 Logged PPO Steps

Method	Proxy	95% CI	Len.	Offline
DSP PPO	−4.1458	[−4.3707, −3.9254]	40.3	No
KL Only	−4.1458	[−4.3707, −3.9254]	40.3	No
BSPO Style	−3.7416	[−4.0012, −3.5044]	40.7	Yes
Ensemble	−3.9584	[−4.1795, −3.7349]	40.6	No

Fig. 1. Proxy reward curves. DSP-PPO and KL-Only overlap closely throughout training.

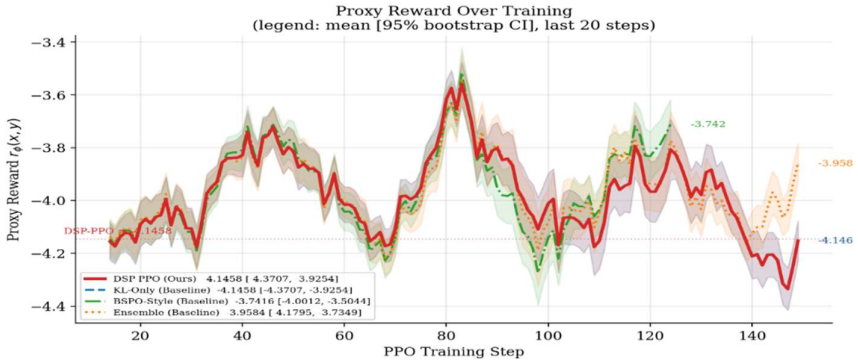

Fig 1. Proxy reward over 150 PPO steps. Legend values are final 20-step means with 95% bootstrap CIs. DSP-PPO and KL-Only overlap at λ=0.001. BSPO-Style achieves higher proxy reward but early-stops at step 125/150 under the instability guard.

B. Latent Shift, KL Divergence, and Response Length

Fig. 2 (left) shows DSP-PPO maintains lower, more stable shift scores than BSPO-Style. The right panel confirms KL divergence remains comparable across methods, indicating that latent shift provides complementary information to token-level KL. All methods remain in a narrow 38–43 token band, ruling out length-based hacking.

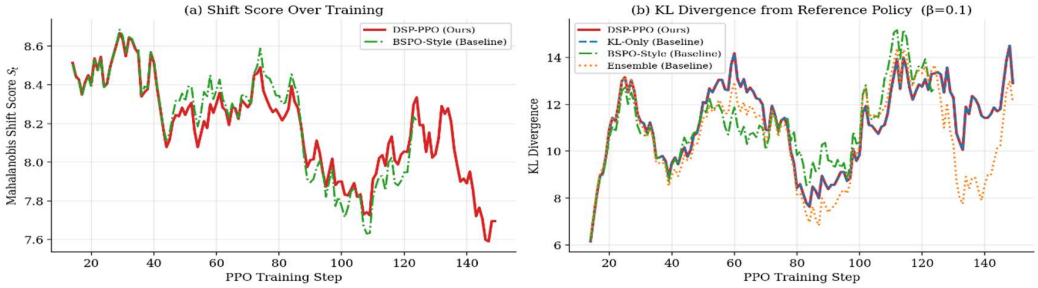

Fig 2. Left: PCA-Mahalanobis shift score over training. DSP-PPO maintains lower and more stable shift than BSPO-Style. Right: KL divergence from the reference policy (β=0.1). Latent shift provides complementary information to token-level KL.

C. Temporal Precedence

Fig. 3 shows the shift score S_ttemporally precedes proxy-reward inflation, with peak lagged correlation at k=15 steps (r=0.1825, p=0.0341). This supports its value as a prospective early-warning signal, though causality requires further intervention studies.

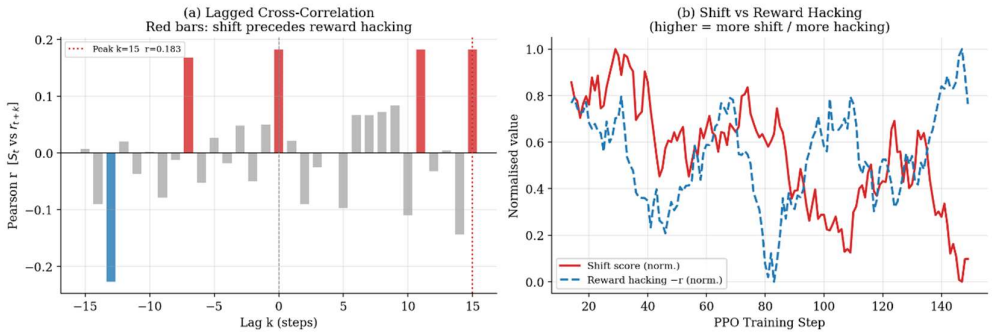

Fig 3. Lagged relationship between shift and later proxy-reward inflation. Peak positive correlation at k=15 steps (r=0.1825, p=0.0341) supports temporal precedence but not causal identification.

D. Preference-Structure Correlation

Shift correlates significantly with reward-model score (r=0.3036, p=1.25×10^{-5}), but chosen and rejected response distributions overlap substantially (Fig. 4). Thus, shift serves as a continuous diagnostic rather than a binary separator.

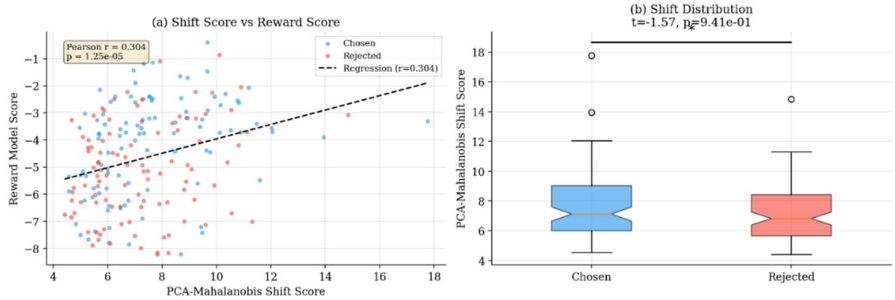

Fig 4. Shift score vs. reward-model score and chosen/rejected shift distributions. Shift correlates with reward-model score (r=0.3036, p=1.25×10^{-5}), but chosen and rejected distributions overlap substantially (p=0.941), con-firming that shift is a correlational signal rather than a binary separator.

E. Ablations, Sensitivity and Overhead

Table III confirms KL and DSP penalties operate on distinct axes: removing KL degrades performance, while the weak λ=0.001 DSP penalty alone does not alter KL-Only results. Proxy reward is stable for λ≤0.002. Per-step overhead is only 0.22% (Table III), dominated by the reward-model forward pass.

Table III. Ablation and Overhead Results.

TABLE III: BETA ABLATION (AT Λ=0.001) AND MEASURED PER-STEP OVERHEAD (200-TRIAL MEAN, SINGLE GPU). REMOVING KL DEGRADES PERFORMANCE WHILE THE WEAK DSP PENALTY HAS NEGLIGIBLE EFFECT ALONE.

Method	β	λ	Proxy Reward	Component	Time (ms)	Relative Cost
DSP-PPO	0.1	0.001	-4.1458	Reward-model forward pass	74.412	100%
DSP-PPO, β=0	0.0	0.001	-3.8765	PCA-Mahalanobis step	0.1636	0.22%
KL-Only	0.1	0.000	-4.1458	—	—	—

IV. DISCUSSIONS, LIMITATIONS, FUTURE WORK

The core contribution of this work is not solving reward hacking outright, but demonstrating that policy-induced latent support drift is a measurable and low-cost online signal during PPO training. At λ=0.001, DSP-PPO matches KL-only proxy reward while providing a continuous latent-drift trace unavailable from KL regularization alone. This second monitoring axis is valuable because proxy reward alone cannot reliably indicate true alignment quality.

In our experiments, the shift signal correlates with reward-model scores (r=0.3036, p=1.25×10^(-5)) and temporally precedes proxy-reward inflation by 15 steps (r=0.1825, p=0.0341). Ablations confirm that latent shift and token-level KL operate on distinct axes. The graded DSP penalty also proved more stable than the binary BSPO-Style approach under the same instability guard.

Limitations. All results are from single-run experiments at GPT-2/DeBERTa scale (124M/435M). Multi-seed evaluations, larger models (7B+), and independent alignment metrics (e.g., human preferences) are essential. The temporal precedence is correlational, not causal. The method was validated only on penultimate representations of one reward model architecture.

Future Work. Near-term priorities include multi-seed lambda sweeps with stronger quality metrics to find settings where DSP-PPO yields stable improvements. Longer-term directions include online updating of in-distribution statistics, preference-aware representation learning, and extensions to direct alignment algorithms such as DPO and RAFT.

IV. CONCLUSION

We introduced DSP-PPO, a lightweight method for real time monitoring and graded penalization of policy-induced distributional shift in reward-model latent space. The method requires a one-time PCA-Mahalanobis statistics pass and adds 0.22% measured per-step overhead. In the conservative final run, DSP-PPO at λ=0.001 matches KL-only proxy reward while producing a continuous shift trace that is not available from KL alone. That trace correlates with reward-model score (r=0.3036, p=1.25×10−5) and temporally precedes later proxy-reward inflation by 15 PPO steps (r=0.1825, p=0.0341). These results support DSP-PPO as a practical monitoring tool and low-cost penalty mechanism, while leaving causal validation and independent alignment-quality evaluation as necessary future work.

REFERENCES

[1] R. Rafailov, Y. Chittepu, R. Park, H. Sikchi, J. Hejna, W. B. Knox, C. Finn, and S. Niekum, "Scaling laws for reward model overoptimization in direct alignment algorithms," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 37, 2024.

[2] J. Dai, T. Chen, Y. Yang, Q. Zheng, G. Pan, "Mitigating reward over optimization in RLHF via behavior-supported regularization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2025.

[3] Z. Liu, M. Lu, S. Zhang, B. Liu, H. Guo, Y. Yang, J. Blanchet, and Z. Wang, "Provably mitigating overoptimization in RLHF: Your SFT loss is implicitly an adversarial regularizer," in Adv. NeurIPS, vol. 37, 2024.

[4] T. Kwa, D. Thomas, and A.G. Alonso, "Catastrophic Goodhart: Regular izing RLHF with KL divergence does not mitigate heavy-tailed reward misspecification," in Adv. NeurIPS, vol. 37, 2024.

[5] J. Fu, X. Zhao, C. Yao, H. Wang, Q. Han, and Y. Xiao, "Reward shaping to mitigate reward hacking in RLHF," arXiv:2502.18770, 2025.

[6] H. Khalaf, C. M. Verdun, A. Oesterling, H. Lakkaraju, and F. P. Calmon, "Inference-time reward hacking in large language models," in Adv. NeurIPS (Spotlight), vol. 38, 2025.

[7] R. Kirk, I. Mediratta, C. Nalmpantis, J. Luketina, E. Hambro, E. Grefen stette, R. Raileanu, “Understanding the effects of RLHF on LLM generalisation and diversity,” in Proc. ICLR, 2024.
[8] R. Yang, R. Ding, Y. Lin, H. Zhang, and T. Zhang, “Regularizing hidden states enables learning generalizable reward model for LLMs,” in Adv. NeurIPS, vol. 37, 2024.
[9] Y. Miao, Sen. Zhang, L. Ding, R. Bao, L. Zhang, D. Tao, “InfoRM: Mitigating reward hacking in RLHF via information-theoretic reward modeling,” in Adv. NeurIPS, vol. 37, 2024.
[10] B. Wang, R. Zheng, L. Chen, Y. Liu, S. Dou, C. Huang, W. Shen, and S. Jin et al., “Secrets of RLHF in large language models Part II: Reward modeling,” arXiv:2401.06080, 2024.
[11] L. Chen, C. Zhu, D. Soselia, J. Chen, T. Zhou, T. Goldstein, H. Huang, M. Shoeybi, and B. Catanzaro, “ODIN: Disentangled reward mitigates hacking in RLHF,” in Proc. Int. Conf. Mach. Learn. (ICML), 2024.
[12] T. Liu, W. Xiong, J. Ren, L. Chen, J. Wu, R. Joshi et al., “RRM: Robust reward model training mitigates reward hacking,” in Proc. ICLR, 2025.
[13] Z. Xu, S. Vemuri, K. Panaganti, D. Kalathil, R. Jain, and D. Ramachan dran, “Robust LLM alignment via distributionally robust direct preference optimization,” arXiv:2502.01930, 2025.