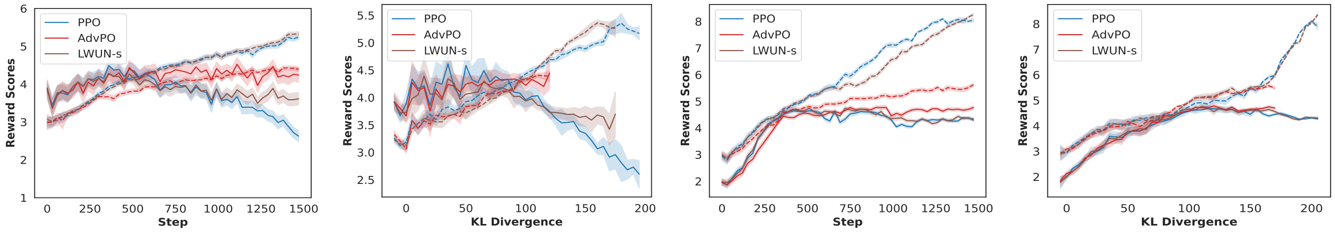
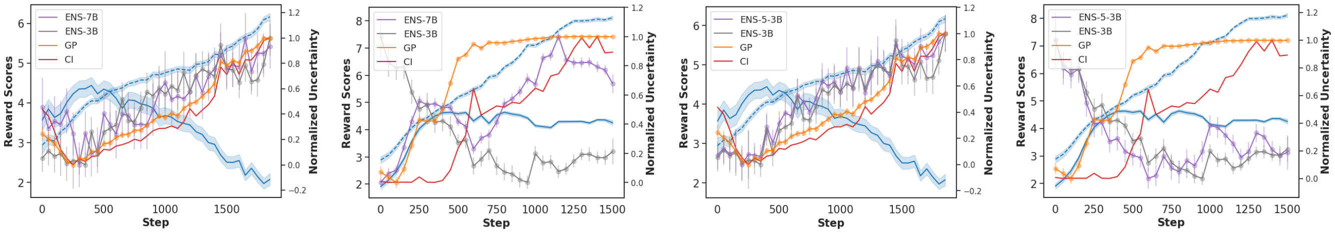


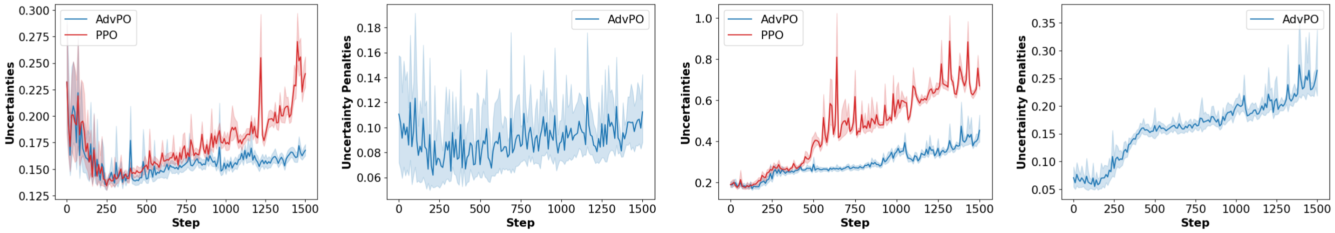
(a) Anthropic HH (Step) (b) Anthropic HH (KL) (c) TL;DR (Step) (d) TL;DR (KL)

Figure 1. Experimental results demonstrating the mitigation of overoptimization in RLHF with ADVPO and LWUN-s. The gold reward is represented by the solid line, while the dashed line corresponds to the proxy reward. The x-axis of Figure 1b and Figure 1d have a square-root scale. **LWUN-s helps mitigate overoptimization, particularly on the Anthropic dataset, as evidenced by the steady gold rewards in Figure 1a and the smaller KL divergence in Figure 1b. However, its performance is not as effective as AdvPO.**



(a) Anthropic HH with ENS-7B (b) TLDR with ENS-7B (c) Anthropic HH with ENS-3B-5 (d) TLDR with ENS-3B-5

Figure 2. Comparison among lightweight uncertainty estimations with ENS-7B and ENS-3B-5. The blue lines with shaded areas depict the reward dynamics concerning optimization steps in PPO, where the solid and dashed lines represent gold and proxy rewards, respectively. The lines with dots denote the results from different uncertainty estimation methods. The reward values are indexed on the left y-axis, while the uncertainty is indexed on the right y-axis. **In Figure 2a and 2b, we include ENS-7B, which quantifies uncertainty through three LLama7B reward ensembles. Particularly on the TLDR datasets, while ENS-3B shows no increasing trend correlated with the divergence between gold and proxy rewards, ENS-7B generally exhibits an increasing trend with some perturbations, indicating its effectiveness in capturing reward uncertainty. We also include ENS-3B-5 in Figure 2c and 2d. ENS-3B-5 employs five LLama3B reward ensembles to quantify uncertainty. However, ENS-3B-5 fails to demonstrate a consistently increasing trend in the reward difference between gold and proxy rewards similar to ENS-3B in TLDR dataset (Figure 2d). This suggests that perhaps the size of the reward model is more crucial than the number of ensembles.**



(a) Anthropic HH (UN) (b) Anthropic HH (UN Penalty) (c) TL;DR (UN) (d) TL;DR (UN Penalty)

Figure 3. Analysis of ADVPO. Figure 3a and 3c illustrate the evolution of the average uncertainty of generated responses by PPO and ADVPO across optimization steps. Figure 3b and Figure 3d depict the average uncertainty penalization of ADVPO over optimization steps. **We can observe from Figure 3a and 3c that the average uncertainties of generated responses remain stable under ADVPO, in contrast to the significant increase in uncertainty observed with PPO.**

1. Revised Part of Section 3

Uncertainty Quantification in Neural Bandits. Following previous work (Xu et al., 2020; Riquelme et al., 2018),we demonstrate in Theorem A.1 that with probability $1-\delta$, the following inequality concerning the uncertainty, or the width of the confidence interval of the estimated reward, holds:

$$|r^*(x,y) - r_\varphi(x,y)| \leq b\sqrt{e(x,y)^\top M_D^{-1} e(x,y)}, \qquad (1)$$

where b is a function of δ (typically the smaller δ is, the larger b is). And M_D summarizes all last layer embeddings observed in the training data for the reward model, i.e.,$M_D = \lambda I + \sum_{i=1}^N \sum_{y \in \{y_c^i, y_r^i\}} e(x_i, y) e(x_i, y)^\top$. For ease of illustration, we then denote $U_{x,y}^{CI} = \sqrt{e(x,y)^\top M_D^{-1} e_(x,y)}$. Intuitively, if the new prompt-response pair (x,y) is similar to samples in the training data, applying the inverse of M_D would result in a small uncertainty $U_{x,y}^{CI}$; otherwise, the uncertainty will be high.

2. Revised Version of Section 4

Given that the lightweight uncertainty estimation methods in Section 3 capture the reliability of estimated rewards and demonstrate potential in mitigating the overoptimization issue, this section devises an effective way to leverage them to improve policy optimization. For subsequent sections, we primarily adopt CI in Eq.(4) to quantify uncertainties in reward modelling, as it demonstrates strong effectiveness, as shown in Figure 1.

The reward uncertainties enable the construction of a confidence interval containing the golden reward with high probability, as defined in Eq.(4). Consequently, to mitigate the risk of overoptimization caused by high rewards with high uncertainty, we should not solely optimize towards the potentially incorrect point estimate $r_\varphi(x,y)$. Instead, it is imperative to optimize the policy towards the reward model's prediction confidence interval, considering that the gold reward could assume any value within the interval with equal probability.

As derived in Theorem A.1, the reward confidence interval indeed arises from the confidence region of the projection weight $\hat{\phi}$, expressed as $\|\phi^* - \hat{\phi}\|_{M_D}^2 \leq B$, where $\phi^* \in \mathbb{R}^d$ and $\hat{\phi} \in \mathbb{R}^d$ denote the optimal and estimated projection weights for the last layer to predict rewards, respectively.

Thus, instead of optimizing toward the reward confidence interval for each sample individually, AdvPO employs distributionally robust optimization, where the inner optimization directly seeks a pessimistic projection weight ϕ within the confidence ball for reward calculation,

$$\max_{\pi_\theta} \min_{\|\phi-\hat{\phi}\|_{M_D}^2 \leq B} \mathbb{E}_{x,y\sim\pi_\theta(\cdot|x)}\left[r_\phi(x,y)\right] - \mathbb{E}_{x,y_{\text{ref}}}\left[r_\phi(x,y_{\text{ref}})\right]$$
$$- \beta\mathbb{E}_{x,y\sim\pi_\theta(\cdot|x)}\left[\mathbb{D}_{\text{KL}}\left[\pi_\theta(y|x)\|\pi_{\text{SFT}}(y|x)\right]\right], \qquad (2)$$

Here, with a bit of abuse of notation, we use $r_\phi(\cdot)$ to denote the reward obtained when using the projection weight ϕ, while keeping the representation encoder unchanged.

Furthermore, to better align our adversarial search with how the reward model is obtained in Eq. (1), we also incorporate a reference response through pairwise preference comparison in our objective function. The reference response can be any acceptable answer, such as an annotated good response from users or a response generated by the SFT model. As demonstrated in Lemma A.2, the inclusion of reference responses prevents AdvPO from becoming overly pessimistic or optimizing incorrectly. This is achieved by directing policy optimization toward the reference responses while also optimizing against pessimistic rewards.

Theorem D.1 in Appendix D demonstrates the inner minimization of Eq.(2) has a closed-form solution, and thus the optimization problem in Eq.(2) has an equivalent but easier to operate form:

$$\max_{\pi_\theta} \quad \mathbb{E}_{x,y\sim\pi_\theta(\cdot|x)}[r_{\hat{\phi}}(x,y) - \frac{1}{\lambda^*}e(x,y)^\top M_D^{-1}g] \qquad (3)$$
$$- \mathbb{E}_{x,y_{\text{ref}}}[r_{\hat{\phi}}(x,y_{\text{ref}}) - \frac{1}{\lambda^*}e(x,y_{\text{ref}})^\top M_D^{-1}g],$$
$$- \beta\mathbb{E}_{x,y\sim\pi_\theta(\cdot|x)}\left[\mathbb{D}_{\text{KL}}\left[\pi_\theta(y|x)\|\pi_{\text{SFT}}(y|x)\right]\right],$$

where $g = \mathbb{E}_{x,y\sim\pi_\theta(\cdot|x)}\left[e(x,y)\right] - \mathbb{E}_{x,y_{\text{ref}}}\left[e(x,y_{\text{ref}})\right]$ and $\lambda^* = \sqrt{\frac{g^\top M^{-1} g}{B}}$.

Comparison to previous approaches in utilizing uncertainty against overoptimization. Previous work (Coste et al., 2023; Eisenstein et al., 2023) utilizes reward uncertainty on a per-sample basis, i.e., penalizing each sample's reward based on its individual uncertainty, as illustrated in Eq. (3). While both per-sample uncertainty penalization and AdvPO adopt a pessimistic approach to leveraging reward uncertainty, the degree of pessimism is crucial (Jin et al., 2021; Rashidinejad et al., 2021). Excessive pessimism, i.e., penalizing rewards too heavily based on uncertainties, is known to impede the discovery of the correct direction for optimization, thus failing to find a good policy.

In Lemma D.2 of Appendix D, we theoretically demonstrate that directly penalizing uncertainty in a sample-wise manner, as in previous work (Coste et al., 2023; Eisenstein et al., 2023), results in a more pessimistic objective compared to the max-min objective in Eq.(3) of AdvPO. This suggests that AdvPO is more likely to contribute to improving policy performance while mitigating overoptimization.

References

Coste, T., Anwar, U., Kirk, R., and Krueger, D. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

Eisenstein, J., Nagpal, C., Agarwal, A., Beirami, A., D'Amour, A., Dvijotham, D., Fisch, A., Heller, K., Pfohl, S., Ramachandran, D., et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Riquelme, C., Tucker, G., and Snoek, J. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling, 2018.

Xu, P., Wen, Z., Zhao, H., and Gu, Q. Neural contextual bandits with deep representation and shallow exploration, 2020.

A. Theoretical Proofs

Theorem A.1. *With probability 1-δ, the following inequality concerning the uncertainty, or the width of the confidence interval of the estimated reward $r_\varphi(x,y)$, holds:*

$$|r^*(x,y) - r_\varphi(x,y)| \le b_{D,\delta}\sqrt{e(x,y)^\top M_D^{-1} e(x,y)} + const, \tag{4}$$

where $b_{D,\delta}$ is a term related to preference dataset D and δ (typically the smaller δ is, the larger $b_{D,\delta}$ is). And M_D summarizes all last layer embeddings observed in the perference dataset D for the reward model, i.e., $M_D = \lambda I + \sum_{i=1}^N \sum_{y\in\{y_c^i, y_r^i\}} e(x_i,y)e(x_i,y)^\top$.

Proof. For ease of illustration, we denote the parameters in $e(x,y)$ as $\hat{w}$, thus $\varphi = (\hat{\phi}, \hat{w})$. Additionally, We also use ϕ^* and w^* to denote the unknown ground-truth of $\hat{\phi}$ and $\hat{w}$ respectively. At a high level, the derivation of the reward uncertainty $|r^*(x,y) - r_\varphi(x,y)|$ consists the following steps:

- **Step 1**: Obtain the confidence region of the learned projection weight $\hat{\phi}$ with preference dataset D, such that with probability $1-\delta$, $\|\phi^* - \hat{\phi}\|_{M_D} \le b_{D,\delta}$ holds, where $b_{D,\delta}$ is a term related to D and δ. Typically, the smaller δ is, the larger $b_{D,\delta}$ is.
- **Step 2**: Derive the reward uncertainty $|r^*(x,y) - r_\varphi(x,y)|$ based on the confidence region of $\hat{\phi}$.

Proof of Step 2. We first elaborate how to derive the reward uncertainty $|r^*(x,y) - r_\varphi(x,y)|$ given $\|\phi^* - \hat{\phi}\|_{M_D} \le b_{D,\delta}$.

Under some assumptions regarding properties of the neural network architecture (Assumption 4.3 in (Xu et al., 2020)), Lemma C.1 in (Xu et al., 2020) shows that $r^*(x,y)$ can be approximated by a linear function around some known init parameters points (ϕ_0, w_0), i.e. :

$$r^*(x,y) = e(x,y)^T\phi^* + \phi_0^T \cdot \nabla_w e(x,y) \cdot (w^* - w_0)$$

And the absolute value of the second term can be bounded as demonstrated by Lemma C.2 in (Xu et al., 2020). Consequently, with probability $1-\delta$, we have

$$|r^*(x,y) - r_\varphi(x,y)| = |e(x,y)^T\phi^* - e(x,y)^T\hat{\phi} + \phi_0^T \cdot \nabla_w e(x,y) \cdot (w^* - w_0)| \le |e(x,y)^T\phi^* - e(x,y)^T\hat{\phi}| + \text{const} \tag{5}$$

Utilizing the inequality $|u^Tv| \le \|u\|_A\|v\|_{A^{-1}}$ for any postive definite matrix A, derived from Cauchy-Schwarz inequality, we further obtain:

$$|e(x,y)^T\phi^* - e(x,y)^T\hat{\phi}| \le \|e(x,y)\|_{M_D^{-1}}\|\phi^* - \hat{\phi}\|_{M_D} \tag{6}$$

Proof of Step 1. Next, we present a proof for step 1, deriving the confidence region of $\hat{\phi}$. From the reward modeling, we have:

$$\hat{\phi} = argmin_\phi L_D = \sum_{(x,y_c,y_r)\in D}\sum_{k\in\{c,r\}} i_{x,k}L_{x,k} = \sum_{(x,y_c,y_r)\in D}\sum_{k\in\{c,r\}} i_{x,k}\log\left(\frac{\exp(e(x,y_k)^T\phi)}{\exp(e(x,y_c)^T\phi) + \exp(e(x,y_r)^T\phi)}\right) \tag{7}$$

where $i_{x,c} = 1$ and $i_{x,r} = 0$. Let $p_x^k := \frac{\exp(e(x,y)^T\hat{\phi})}{\exp(e(x,y_c)^T\hat{\phi}) + \exp(e(x,y_r)^T\hat{\phi})}$, $k \in \{c,r\}$. Since $\hat{\phi}$ is the minimizer regarding L_D, setting the derivative of L_D with respect to $\hat{\phi}$ as zero, we have :

$$\sum_{(x,y_c,y_r)\in D}\sum_{k\in\{c,r\}} (p_{x,k} - i_{x,k})e(x,y_k) = 0$$

Denote

$$\mu(\phi, e(x,y_k)) = \frac{\exp(e(x,y_k)^T\phi)}{\sum_{k\in\{c,r\}}\exp(e(x,y_k)^T\phi)}$$

and

$$G(\phi) = \sum_{D} \sum_{k \in \{c,r\}} [\mu(\phi, e(x, y_k)) - \mu(\phi^*, e(x, y_k))] e(x, y_k).$$

Following the Step 1 of Theorem 1 in (Li et al., 2017), we can derive :

$$\|\hat{\phi} - \phi^*\|^2_{M_D} \leq \frac{1}{\kappa^2} \|G(\hat{\phi})\|^2_{M_D^{-1}},$$

where $\kappa := \inf_{\|\phi^* - \phi\| \leq 1} \dot{\mu}(\phi, e(x, y)) \geq 0, \forall e(x, y)$ (Assumption 1 in (Li et al., 2017)). Then Lemma 3 in (Li et al., 2017) further bounds $\|G(\hat{\phi})\|^2_{M_D^{-1}}$ by a term related to D and δ, resulting $\|\phi^* - \phi\|_{M_D} \leq b_{D,\delta}$ holding with probability $1 - \delta$.

Thus we conclude the proof. □

Lemma A.2. *The inclusion of reference responses prevents* AdvPO *from being overly or wrongly pessimistic by enforcing policy optimization towards the direction of the reference responses while optimizing against pessimistic rewards.*

Proof. Let $\hat{\phi}^*_{\text{ref}}$ and $\hat{\phi}^*_{\text{noref}}$ represent the derived projection weights of the inner optimization of the max-min objective in Eq.(7) with or without reference responses, respectively. Denote $g_{\pi_\theta} = \mathbb{E}_{x, y \sim \pi_\theta(\cdot|x)}\left[e(x, y)\right]$ and $z_{\text{ref}} = \mathbb{E}_{x, y_{\text{ref}}}\left[e(x, y_{\text{ref}})\right]$ Thus the policy optimization objective for max-min objective ¡em¿with¡/em¿ reference responses (i.e., Eq.(7)) is

$$J_{\text{ref}} = \max_{\pi_\theta} \quad g^T_{\pi_\theta} \hat{\phi}^*_{\text{ref}} - z^T_{\text{ref}} \hat{\phi}^*_{\text{ref}} = \max_{\pi_\theta} \quad g^T_{\pi_\theta} \hat{\phi}^*_{\text{ref}}.$$

The last equality holds because the second term is a constant given $\hat{\phi}^*_{\text{ref}}$, thus subtracting it will not affect the resulted optimal policy. Similarly, we can derive the policy optimization objective for max-min objective ¡em¿without¡/em¿ reference responses:

$$J_{\text{noref}} = \max_{\pi_\theta} \quad g^T_{\pi_\theta} \hat{\phi}^*_{\text{noref}}.$$

Following a similar procedure as Theorem D.1 in Appendix D and replacing g in Eq.(14) by $g_{\pi_\theta} - z_{\text{ref}}$ and g_{π_θ}, we can derive the closed-form solution of $\hat{\phi}^*_{\text{ref}}$ and $\hat{\phi}^*_{\text{noref}}$. By plugging them into J_{ref} and J_{noref}, we can get:

$$J_{\text{ref}} = \max_{\pi_\theta} \quad g^T_{\pi_\theta} \hat{\phi}^*_{\text{ref}} = g^T_{\pi_\theta} \hat{\phi} - \frac{1}{\lambda^*_{\text{ref}}} g^T_{\pi_\theta} M_D^{-1} g_{\pi_\theta} + \underbrace{\frac{1}{\lambda^*_{\text{ref}}} g^T_{\pi_\theta} M_D^{-1} z_{\text{ref}},}_{(A_{\text{ref}})}$$

and

$$J_{\text{noref}} = \max_{\pi_\theta} \quad g^T_{\pi_\theta} \hat{\phi}^*_{\text{noref}} = g^T_{\pi_\theta} \hat{\phi} - \frac{1}{\lambda^*_{\text{noref}}} g^T_{\pi_\theta} M_D^{-1} g_{\pi_\theta}$$

where λ^*_{ref} and λ^*_{noref} are Lagrangian multipliers derived from the optimization process.

We can observe that both J_{ref} and J_{noref} aim to prevent the policy from moving in the direction of high uncertainty by minimizing $g^T_{\pi_\theta} M_D^{-1} g_{\pi_\theta}$. However, J_{ref} includes an additional term A_{ref} compared to J_{noref}. This term encourages the policy π_θ to move towards the reference responses $z_{\text{ref}} = \mathbb{E}_{x, y_{\text{ref}}}[e(x, y_{\text{ref}})]$. With reasonably good reference responses, i.e., $z^T_{\text{ref}} \phi^* > 0$, the additional term A_{ref} guides the policy in a more accurate optimization direction, preventing AdvPO from being overly or wrongly pessimistic. □