Robust Preference Optimization through Reward Model Distillation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Language model (LM) post-training (or alignment) involves maximizing a reward function that is derived from preference annotations. Direct Preference Optimization (DPO) is a popular offline alignment method that trains a policy directly on preference data without the need to train a reward model or apply reinforcement learning. However, typical preference datasets have only a single, or at most a few, annotation per preference pair, which causes DPO to overconfidently assign rewards that trend towards infinite magnitude. This frequently leads to degenerate policies, sometimes causing even the probabilities of the *preferred* generations to go to zero. In this work, we analyze this phenomenon and propose *distillation* to get a better proxy for the true preference distribution over generation pairs: we train the LM to produce probabilities that match the distribution induced by a reward model trained on the preference data. Moreover, to account for uncertainty in the reward model we are distilling from, we optimize against a *family of reward models* that, as a whole, is likely to include at least one reasonable proxy for the preference distribution. Our results show that distilling from such a family of reward models leads to improved robustness to distribution shift in preference annotations, while preserving the simple supervised nature of DPO.

1 Introduction

Language model (LM) post-training (or alignment) aims to steer language model policies towards responses that agree with human preferences. Early state-of-the-art approaches have focused on reward learning from human feedback. In this paradigm, preference annotations are used to train reward models, which then guide the optimization of the language model policy through online reinforcement learning (an approach broadly referred to as RLHF). Recent research on offline "Direct Preference Optimization" [DPO; 23] and extensions thereof [3; 31], however, has demonstrated that it is also possible to directly optimize policies on the preference data, which bypasses the need for a separate reward model—and its offline nature also leads to faster, and simpler, training frameworks.

While this direct approach to preference optimization is attractive in terms of its simplicity and efficiency, it also raises important questions about the effectiveness and robustness of the resulting policies—as well as the broader utility of using an explicit reward model. In this paper, we argue that explicit reward modeling can, in fact, offer substantial practical advantages that are not captured by DPO's formulation. In particular, we theoretically show that relying solely on the preference data can be a precarious strategy, with few natural brakes in place to prevent policies trained under the DPO objective from careening off towards degenerate policies when the preference data exhibits certain idiosyncratic properties. On the other hand, explicit reward models can easily be regularized and understood—regardless of whether they are Bradley-Terry models [4], margin-based ranking models [40], or simply any other kind of function that correlates well with human preferences [31; 17].

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

Taking a step back from pure direct preference optimization, we propose a method that merges the best of both worlds: an efficient reward model distillation algorithm that (i) operates effectively in the offline setting, (ii) makes minimal assumptions about the true, optimal reward we aim to maximize, and (iii) demonstrates greater robustness to the specific distribution of prompt/response data used for policy alignment. Drawing inspiration from prior knowledge distillation techniques [14; 26; 35; 10], we leverage the same change of variables trick employed in DPO to express the language model policy in terms of its implicit reward model [23]. We then train the policy to match our desired, explicit reward via an L_2 loss that directly regresses the pairwise differences in target rewards for any two generation pairs (x, y_1) and (x, y_2). We theoretically establish the equivalence between optimizing this distillation loss over a sufficiently diverse offline dataset of unlabeled examples and optimizing the traditional online RLHF objective.

Our reward model distillation approach, however, is not immune to some of the same challenges facing DPO-style learning of policies. In particular, reward model distillation requires having a reliable reward model—but having a reliable reward requires having a reliable method for extracting a reward model from a potentially noisy preference dataset. To address the uncertainty surrounding the "right" reward model, we introduce a pessimistic extension to our approach. This extension aims to maximize the worst-case improvement of our model across a plausible family of reward models (e.g., those sufficiently consistent with annotated preference data). This strategy aligns with that of existing work in conservative offline reinforcement learning [5; 16]. Interestingly, we derive that this pessimistic objective can be equivalently expressed and optimized by adding a simple additional KL-divergence regularization to the original distillation objective.

Empirically, we find that reward model distillation, particularly pessimistic reward model distillation, leads to similar performance to prior direct preference optimization methods in settings where the preference datasets used are unbiased, but significantly better performance in settings where the preference datasets are biased, when compared to DPO and the Identity Preference Optimization (IPO) framework of [3], which was introduced as a more robust alternative to DPO. To further support these empirical observations, we provide an extensive theoretical analysis that both (i) sheds more light on the degenerative tendencies of DPO and issues inherent to its objective, and (ii) highlights relative advantages of our explicitly regularized approaches.

2 Preliminaries

We begin with a brief review of Direct Preference Optimization (DPO) [23] and its analysis. Proofs of all theoretical results provided here, and in the rest of the paper, are deferred to Appendix A.

2.1 The preference alignment problem

Let x be an input prompt, and let $y \sim \pi_\theta(\cdot \mid x)$ be the language model policy π_θ's response to x. Given some reward function $r^*(x, y)$ and another reference policy $\pi_{\text{ref}}(y \mid x)$, the goal of alignment is to solve for the "aligned" policy $\pi_{\theta^*}(y \mid x)$ that maximizes the following RLHF objective, i.e.,

$$\pi_{\theta^*}(y \mid x) = \operatorname*{argmax}_{\pi_\theta} \mathbb{E}_{\mu(x)} \left[\mathbb{E}_{\pi_\theta(y|x)}[r^*(x, y)] - \beta \mathbb{D}_{\text{KL}}[\pi_\theta(\cdot \mid x) \| \pi_{\text{ref}}(\cdot \mid x)] \right], \tag{1}$$

where $\mu(x)$ is a fixed distribution over prompts, and the KL-divergence term prevents the aligned policy from being dramatically different from the anchoring reference policy, $\pi_{\text{ref}}(y \mid x)$. Here, the reward function r^* is typically not known in advance, but rather inferred from collected human preference data in the form of (x, y^w, y^ℓ), where x is the prompt, y^w is the "winning", or preferred, response, and y^ℓ is the "losing", or dispreferred, response. A common approach is to assume that pairs (y_1, y_2) follow a Bradley-Terry model [4], under which the probability that y_1 is preferred to y_2 given the reward function r^* and prompt x is $p^*(y_1 \succ y_2 \mid x) = \sigma(r^*(x, y_1) - r^*(x, y_2))$, where $\sigma(\cdot)$ is the sigmoid function and $\succ$ denotes preference. Under this model, we can use the preference data $(x, y^w, y^\ell) \sim \mathcal{D}_{\text{pref}}$ to estimate r^* via maximum likelihood estimation, i.e.,

$$\hat{r} \in \operatorname*{argmin}_r \mathbb{E}_{(y^w, y^\ell, x) \sim \mathcal{D}_{\text{pref}}} \left[-\log \sigma(r(x, y^w) - r_\phi(x, y^\ell)) \right]. \tag{2}$$

With $\hat{r}$ in hand, Eq. (1) can be optimized using standard reinforcement learning algorithms [27; 29; 6].

2.2 Direct preference optimization

DPO is a simple approach for offline policy optimization that uses preferences to directly align the language model policy, without training an intermediate reward model. Specifically, DPO leverages the fact that the optimal solution to the KL-constrained objective in (1) takes the form [15]

$$\pi_{\theta^*}(y \mid x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y \mid x)\exp\left(\frac{1}{\beta}r^*(x,y)\right), \tag{3}$$

where $Z(x) = \sum_y \pi_{\text{ref}}(y \mid x)\exp(\frac{1}{\beta}r^*(x,y))$ is the partition function. DPO reparameterizes the true reward function r^* in terms of the optimal policy π_{θ^*} that it induces, i.e.,

$$r^*(x,y) = \beta\log\left(\frac{\pi_{\theta^*}(y \mid x)}{\pi_{\text{ref}}(y \mid x)}\right) + \beta\log Z(x). \tag{4}$$

Under the Bradley-Terry model, the likelihood that $y_1 \succ y_2$ can then be written as

$$p^*(y_1 \succ y_2 \mid x) = \sigma\left(\beta\log\frac{\pi_{\theta^*}(y_1)\pi_{\text{ref}}(y_2)}{\pi_{\theta^*}(y_2)\pi_{\text{ref}}(y_1)}\right), \tag{5}$$

where now π_{θ^*} can be directly estimated on $\mathcal{D}_{\text{pref}}$ following the objective in (2), in place of the intermediate reward model $\hat{r}$, i.e., $\pi_{\hat{\theta}}(y \mid x) \in \operatorname{argmin}_{\pi_\theta} \mathcal{L}_{\text{dpo}}(\pi_\theta; \mathcal{D}_{\text{pref}})$ where

$$\mathcal{L}_{\text{dpo}}(\pi_\theta; \mathcal{D}_{\text{pref}}) = \mathbb{E}_{(y^w,y^\ell,x)\sim\mathcal{D}_{\text{pref}}}\left[-\log\sigma\left(\beta\log\frac{\pi_{\theta^*}(y^w)\pi_{\text{ref}}(y^\ell)}{\pi_{\theta^*}(y^\ell)\pi_{\text{ref}}(y^w)}\right)\right]. \tag{6}$$

2.3 Pitfalls of direct preference optimization

As argued in [3], the Bradley-Terry assumption that DPO strongly relies on for maximum likelihood estimation is sensitive to the underlying preference data. Specifically, if we have any two responses y_1 and y_2 where $p^*(y_1 \succ y_2 \mid x) = 1$, then the Bradley-Terry model dictates that $r^*(y_1) - r^*(y_2) = +\infty$, and therefore $\pi_{\theta^*}(y_2 \mid x) = 0$ for *any* finite KL-regularization strength β.

We can illustrate this phenomenon on a broader level with the following example.

Assumption 1. *Suppose we are given a preference dataset of (context-free) pairs* $\mathcal{D}_{\text{pref}} = \{(y_i^w, y_i^\ell)\}_{i=1}^n$*, the pairs* (y_i^w, y_i^ℓ) *are mutually disjoint in both the elements. Further suppose that we optimize the DPO objective on* $\mathcal{D}_{\text{pref}}$ *with a single parameter* θ_y *for each* y*.*

Proposition 1. *Under Assumption 1, for any* (y, y') *such that* $y = y_i^w$ *and* $y' = y_i^\ell$ *for some* i*, we have* $\frac{\pi_{\theta^*}(y)\pi_{\text{ref}}(y')}{\pi_{\theta^*}(y')\pi_{\text{ref}}(y)} \to \infty$*, for all global minimizers* π_{θ^*} *of the DPO objective in (6), for any* $\beta > 0$*.*

Corollary 1. *Under Assumption 1, further assume that* $0 < \pi_{\text{ref}}(y) < 1$ *for all* y*. Then* π_{θ^*} *is a global minimizer of the DPO objective in (6) iff* $\pi_{\theta^*}(\mathcal{C}(y^\ell)^c) \to 1$ *with* $\pi_{\theta^*}(y_i^w) > 0\ \forall i \in [n]$*, where* $\mathcal{C}(y^\ell)^c$ *is the complement of the set of all responses* y *that appear as a dispreferred* y_i^ℓ *for any* $i \in [n]$*.*

Additional analysis of the training dynamics of DPO is also provided in §5. A significant, and non-obvious, implication of Corollary 1 is that the set of global optima of the DPO loss also includes policies that can shift nearly all probability mass to responses that never even appear in the training set—and even assign near zero probability to all of the training data responses that do in fact correspond to winning generations, y^w, a phenomenon that has been observed empirically [e.g., 20]. Stated differently, Corollary 1 implies that any θ^* merely satisfying $\pi_{\theta^*}(y_i^\ell) = 0$ with $\pi_{\theta^*}(y_i^w) > 0\ \forall i \in [n]$ is a global minimizer of the DPO objective in this setting. Though simplistic, the scenario in Assumption 1 is closer to reality than might first be appreciated: in many practical situations we can almost always expect the finite-sample preference data to contain one (or at most a few) preference annotations per example (x, y_1, y_2), while the policies π_θ can have billions of parameters ($\gg n$). Of course, this issue can also be viewed as a classic instance of overfitting—with the additional caveat that as opposed to *overpredicting* responses within the training set, we might overfit to *almost never* producing anything like the "good" responses that do appear within the training set. Furthermore, without additional regularization (beyond β), we can expect this degeneration to easily happen in typical preference datasets.

3 Uncertainty-aware reward model distillation

As discussed in the previous section, a core issue in preference optimization is that the true preference distribution $p^*(y_1 \succ y_2 \mid x)$ is not known. Attempting to infer it from finite-sample preference data (that may further be biased or out-of-distribution with respect to the target domain) can then result in a failure to learn reasonable policies. In this section, we now propose an inherently regularized approach to direct preference optimization that uses uncertainty-aware reward model distillation.

3.1 Reward model distillation

Suppose for the moment that the reward function r^* was in fact known, and did not have to be inferred from sampled preference data. Under this setting, we can then define an efficient offline optimization procedure that is similar in spirit to DPO, but no longer relies directly on a preference dataset. Concretely, given unlabeled samples $(x, y_1, y_2) \sim \rho$ (where the number of samples can be potentially unlimited), we can define a simple "distillation" loss, $\mathcal{L}_{\text{distill}}(r^*, \pi_\theta)$, as follows:

$$\mathcal{L}_{\text{distill}}(r^*, \pi_\theta; \rho) = \mathbb{E}_{\rho(x,y_1,y_2)}\left[\left(r^*(x, y_1) - r^*(x, y_2) - \beta \log \frac{\pi_\theta(y_1 \mid x)\pi_{\text{ref}}(y_2 \mid x)}{\pi_\theta(y_2 \mid x)\pi_{\text{ref}}(y_1 \mid x)}\right)^2\right]. \quad (7)$$

Intuitively, the distillation loss seeks to exactly match *differences* in reward model scores across all generation pairs (x, y_1, y_2). It is then easy to see that under the Bradley-Terry model, this is equivalent to matching the strength of the preference relationship, $y_1 \succ y_2$. Furthermore, by only matching differences, we can still conveniently ignore the log partition term, $\log Z(x)$, in the implicit reward formulation for π_θ as shown in (4), as it is constant across different y for any given x. Finally, similar to the motivation in DPO, we can show that minimizing $\mathcal{L}_{\text{distill}}(r^*, \pi_\theta; \rho)$ indeed results in an optimally aligned policy π_{θ^*}, as long as the data distribution ρ has sufficient support.

Theorem 1. *Let $\mathcal{Y}$ denote the set of all possible responses for any model π_θ. Assume that $\text{supp}(\pi_{\text{ref}}(y \mid x)) = \mathcal{Y}$, i.e., the reference policy may generate any outcome with non-zero probability. Further, let $\text{supp}(\rho(x, y_1, y_2)) = \text{supp}(\mu(x)) \times \mathcal{Y} \times \mathcal{Y}$. Let $\pi_{\theta^*}(y \mid x) \in \text{argmin}_{\pi_\theta} \mathcal{L}_{\text{distill}}(r^*, \pi_\theta; \rho)$ be a minimizer over all possible policies, of the implicit reward distillation loss in (7), for which $r^*(x, y)$ is assumed to be deterministic, and finite everywhere. Then for any $\beta > 0$, π_{θ^*} also maximizes the alignment objective in (1).*

The above result holds for a broad class of data distributions $\rho(x, y_1, y_2)$, and makes no assumptions on r^* (e.g., it is no longer necessary for it to be defined using a Bradley-Terry model). In fact, this result can also be seen as strict generalization of the IPO framework of [3] when taking $r^*(x, y) \triangleq \mathbf{1}\{y = y_w\}$, if labeled pairs (x, y_w, y_l) are provided instead of the unlabeled pairs (x, y_1, y_2).

Of course, the true reward r^* is usually not known in practice. Still, as in standard RLHF, we can go about constructing good proxies by using the preference data to identify plausible target reward models r_{tgt}—further guided by any amount of regularization and inductive bias that we desire. A natural choice is to first learn r_{tgt} on the preference data $\mathcal{D}_{\text{pref}}$ using standard methods, and then reuse $\mathcal{D}_{\text{pref}}$ to distill π_θ, which is similar to classical settings in teacher-based model distillation [14; 26]. Furthermore, as r_{tgt} is a real-valued model, at a bare minimum it is guaranteed to induce a regularized Bradley-Terry preference distribution $p_{\text{tgt}}(y_1 \succ y_2 \mid x) > 0$, $\forall x, y_1, y_2 \in \mathcal{X} \times \mathcal{Y}$, and thereby avoid some of the degeneracies identified in §2.3 for the maximum likelihood estimate under DPO.

3.2 Pessimistic reward model distillation

Choosing a single reward model r_{tgt} for anchoring the LM policy can naturally still lead to degenerate behavior if r_{tgt} is a poor approximation of the true r^* that accurately reflects human preferences. However, we can easily extend our framework to handle uncertainty in the right target reward function by defining a confidence *set* of $k \geq 1$ plausible target reward models, $\mathcal{S} = \left\{r^1_{\text{tgt}}, \ldots, r^k_{\text{tgt}}\right\}$, and training $\pi_{\theta^*}(y \mid x)$ to maximize the following "pessimistic" form of the objective in (1):

$$\max_{\pi_\theta} \min_{r^i_{\text{tgt}} \in \mathcal{S}} \mathbb{E}_{\mu(x)}\Big[\underbrace{\mathbb{E}_{\pi_\theta(y|x)}[r^i_{\text{tgt}}(x, y)] - \mathbb{E}_{\pi_{\text{ref}}(y|x)}[r^i_{\text{tgt}}(x, y)]}_{\text{advantage over baseline policy}} - \beta \mathbb{D}_{\text{KL}}(\pi_\theta(\cdot \mid x) \| \pi_{\text{ref}}(\cdot \mid x))\Big]. \quad (8)$$

In this pessimistic objective we are no longer optimizing π_θ for a single reward, but optimizing π_θ to produce generations that are scored favorably on average, even by the worst-case reward model in the set $\mathcal{S}$, relative to the generations of the baseline policy π_{ref}.When the set $\mathcal{S} = \{r^*\}$ consists of only the ground-truth reward, the objective (8) is equivalent to standard RLHF (1), up to a constant offset independent of θ. More generally, whenever $\mathcal{S}$ includes a good proxy $\widetilde{r}$ for r^*, the pessimistic advantage evaluation ensures that the the policy π^*_θ that maximizes eq. (8) still has a large advantage over π_{ref} under all $r \in \mathcal{S}$, including $\widetilde{r}$. This use of pessimism to handle uncertainty in the knowledge of the true reward is related to similar techniques in the offline RL literature [16; 5].

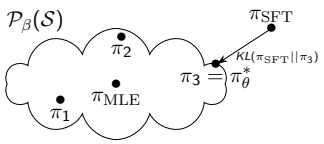

Figure 1: A toy illustration of Theorem 2, which states that the optimal π_{θ^*} for (8) is the policy in $\mathcal{P}_\beta(\mathcal{S})$ with the lowest forward-KL from π_{SFT}. The set $\mathcal{P}_\beta(\mathcal{S})$ contains a (potentially infinite) set of policies $\pi_1, \pi_2, \ldots$ corresponding to target reward models. Here, π_{SFT} assigns equal mass to y^w and y^ℓ, π_{MLE} is the MLE solution for the DPO objective, which puts all probability mass on y^w, and π_3 is the policy in $\mathcal{P}_\beta(\mathcal{S})$ with lowest forward-KL.

For the objective to be meaningful, the set $\mathcal{S}$ has to be chosen carefully. When $\mathcal{S}$ is small, it might not include any good proxy for r^*. Conversely, if $\mathcal{S}$ is too rich, it forces π_{θ^*} to be nearly identical to π_{ref}, since any deviations from π_{ref} might be penalized by some reward model in $\mathcal{S}$. Consequently, we want to design $\mathcal{S}$ to be the smallest possible set which contains a reasonable approximation to r^*.

To optimize (8), it turns out that we can formulate it as an equivalent constrained offline optimization problem, that we will show to conveniently admit a similar loss form as (7).

Theorem 2 (Pessimistic distillation)**.** *Define the constrained minimizer*

$$\pi_{\theta^*}(y \mid x) \in \operatorname*{argmin}_{\pi_\theta \in \mathcal{P}_\beta(\mathcal{S})} \beta \mathbb{E}_{\mu(x)} \mathbb{D}_{\text{KL}}(\pi_{\text{ref}}(\cdot \mid x) \| \pi_\theta(\cdot \mid x)), \tag{9}$$

where $\mathcal{P}_\beta(\mathcal{S})$ is the set of all possible policies with implicit reward models that are consistent with any target reward model $r^i_{\text{tgt}} \in \mathcal{S}$, i.e., $\mathcal{P}_\beta(\mathcal{S}) \triangleq \{\pi_{\theta_i}\}_{i=1}^{|\mathcal{S}|}$ where $\pi_{\theta_i} \propto \pi_{\text{ref}}(y \mid x) \exp \frac{1}{\beta} r^i_{\text{tgt}}(x, y)$. Then for any $\beta > 0$, π_{θ^*} also maximizes the pessimistic alignment objective in (8).

To unpack this result, Theorem 2 stipulates that the π_θ that maximizes the pessimistic objective in (8) is the policy in $\mathcal{P}_\beta(\mathcal{S})$ that is closest in *forward* KL-divergence to π_{ref} (see Figure 1).[1] In addition, this policy also maximizes the expected reward of one of the $r^i_{\text{tgt}} \in \mathcal{S}$ (minus the additional weighted reverse KL-divergence penalty term). Intuitively, the forward KL-divergence term serves the role of biasing the model towards optimizing for reward models that are similar to the implicit reward that π_{ref} already maximizes. Otherwise, there might exist a target reward model $r^i_{\text{tgt}} \in \mathcal{S}$ for which the advantage of π_θ relative to π_{ref} will be low, or even negative (a solution that we would like to avoid).

3.2.1 Optimization

The constraint in (9) can then be relaxed and approximately optimized by introducing an objective with a Lagrangian-style penalty with strength $\alpha > 0$ on a form of distillation loss as (7), i.e.,

$$\min_{\pi_\theta} \beta \mathbb{E}_{\mu(x)} \mathbb{D}_{\text{KL}}(\pi_{\text{ref}}(y \mid x) \| \pi_\theta(y \mid x)) + \alpha \min_{r^i_{\text{tgt}} \in \mathcal{S}} \mathcal{L}_{\text{distill}}(r^i_{\text{tgt}}, \pi_\theta; \rho), \tag{10}$$

where in practice we divide by α and instead optimize[2]

$$\mathcal{L}_{\text{pdistill}}(\mathcal{S}, \pi_\theta; \rho) = \min_{r^i_{\text{tgt}} \in \mathcal{S}} \mathcal{L}_{\text{distill}}(r^i_{\text{tgt}}, \pi_\theta; \rho) + \gamma \mathbb{E}_{\mu(x)} \mathbb{D}_{\text{KL}}(\pi_{\text{ref}}(\cdot \mid x) \| \pi_\theta(\cdot \mid x)), \tag{11}$$

where $\gamma = \beta\alpha^{-1}$. In reality, minimizing (11) for $\gamma > 0$ is equivalent to solving the constrained optimization problem in (9) with an implicitly larger set of possible reward models $\mathcal{S}_\gamma \supseteq \mathcal{S}$ indexed by γ. More specifically, $\mathcal{S}_\gamma$ also contains all reward models $\tilde{r}$ that are approximately consistent with the anchoring reward models r^i_{tgt} contained in $\mathcal{S}$, as the following result states.

[1] Note that the objective in (9) minimizes the *forward* KL-divergence $\mathbb{D}_{\text{KL}}(\pi_{\text{ref}}(\cdot \mid x) \| \pi_\theta(\cdot \mid x))$ even though the pessimistic objective in (8) is regularized with *reverse* KL-divergence $\mathbb{D}_{\text{KL}}(\pi_\theta(\cdot \mid x) \| \pi_{\text{ref}}(\cdot \mid x))$.

[2] In practice, we compute and optimize the $\min$ over reward models per each mini-batch of examples.

Proposition 2 (Soft pessimistic distillation)**.** *Assume the same conditions as Theorem 1. Then for any $0 < \gamma < \infty$, there exists a $\lambda \geq 0$ such that $\pi_{\theta^*}(y \mid x) \in \operatorname{argmin}_{\pi_\theta} \mathcal{L}_{\text{pdistill}}(\mathcal{S}, \pi_\theta; \rho)$, where π_{θ^*} is a minimizer over all possible policies, is a solution to (9) for the effective reward model set*

$$\mathcal{S}_\gamma = \bigcup_{r^i_{\text{tgt}} \in \mathcal{S}} \Big\{ \tilde{r} \colon \mathbb{E}_{\rho(x, y_1, y_2)} \left[(r^i_{\text{tgt}}(x, y_1) - r^i_{\text{tgt}}(x, y_2) - \tilde{r}(x, y_1) + \tilde{r}(x, y_2))^2 \right] \leq \lambda \Big\}. \tag{12}$$

As a result, optimizing (11) even when using the singleton $\mathcal{S} = \{r_{\text{tgt}}\}$ yields an implicitly pessimistic objective, in which the pessimism is over all reward models $\tilde{r}$ that are consistent up to λ with r_{tgt}.

3.3 Pessimistic DPO

We can also observe that Proposition 2 can be leveraged to obtain an alternative, implicitly pessimistic, objective that uses DPO directly instead of distillation. Consider the following regularized DPO loss:

$$\mathcal{L}_{\text{pdpo}}(\pi_\theta; \mathcal{D}_{\text{pref}}) = \mathcal{L}_{\text{dpo}}(\pi_\theta; \mathcal{D}_{\text{pref}}) + \gamma \mathbb{E}_{\mu(x)} \mathbb{D}_{\text{KL}}(\pi_{\text{ref}}(y \mid x) \| \pi_\theta(y \mid x)). \tag{13}$$

Following a similar analysis as in Proposition 2, we can derive that this implicitly corresponds to maximizing the pessimistic objective in (8) for the reward model set

$$\mathcal{S}_\gamma = \Big\{ r_{\pi_\theta} \colon \mathcal{L}_{\text{dpo}}(\pi_\theta; \mathcal{D}_{\text{pref}}) \leq \min_{\pi'_\theta} \mathcal{L}_{\text{dpo}}(\pi'_\theta; \mathcal{D}_{\text{pref}}) + \lambda \Big\}, \tag{14}$$

where $r_{\pi_\theta}(x, y) \triangleq \beta \log \pi_\theta(y \mid x) / \pi_{\text{ref}}(y \mid x) + \beta \log Z(x)$ is the implicit reward model defined by π_θ. $\mathcal{S}_\gamma$ then corresponds to the set of reward models r_{π_θ} that are all approximate minimizers of the DPO loss. This not only includes the MLE, but also all other estimators that obtain nearly the same loss. In principle, this can be expected to help ameliorate some of the issues of §2.3: since driving the reward to $\pm\infty$ only marginally decreases the $\mathcal{L}_{\text{dpo}}$ loss past a certain point, the set $\mathcal{S}$ will also include finite reward functions $|r_{\pi_\theta}(x, y)| < \infty$ for any $\gamma > 0$. These rewards would then be preferred if they induce a policy with a smaller (forward) KL-divergence to π_{ref} than the degenerate, infinite rewards.

4 Experimental results

The main motivation for reward distillation and pessimism is to increase alignment robustness in challenging settings where it is difficult to learn good policies directly from the preference data. To demonstrate the effectiveness of our approach, we run experiments on the popular TL;DR summarization task [29; 32], in which we simulate a scenario where the preference data has a spurious correlation between the *length* of a summary and whether or not it is preferred.[3]

4.1 Experimental setup

We first train an "oracle" reward model on the TL;DR preference data training set [29] and relabel all preference pairs with this oracle. This enables us to use the oracle reward model for evaluation, without worrying about the gap to true human preferences. After relabeling, longer responses (where longer is defined as y_1 having at least 10% more tokens than y_2) are preferred in 61% of the examples.

To test the effect of a spurious correlation on preference-based policy optimization, we select as a training set 30K examples from the relabeled data such that the longer output is preferred in ρ fraction of examples, with $\rho \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. Each such training set is denoted $\mathcal{D}_\rho$. At each $\mathcal{D}_\rho$, we compare our approach to DPO [23] and IPO [3], which are currently the most commonly used offline alignment methods. We test the following variants of distillation and pessimism:

- **Distilled DPO** (d-DPO): Trains a reward model r_ρ on $\mathcal{D}_\rho$, and then optimizes $\mathcal{L}_{\text{distill}}(r_\rho, \pi_\theta; \rho)$.
- **Pessimistic DPO** (p-DPO): A pessimistic version of DPO as described in §3.3, trained on $\mathcal{D}_\rho$.
- **Pessimistic Distilled DPO** (pd-DPO): Combines the above two by training a reward model r_ρ on $\mathcal{D}_\rho$ and optimizing the pessimistic distillation objective (Eq. (11)) with confidence set $\mathcal{S} = \{r_{\text{tgt}}\}$.
- **Pessimistic Ensemble DPO** (e-DPO): To create ensembles of reward models, we subsample from each $\mathcal{D}_\rho$ five preference datasets, $\mathcal{D}_{\rho,b}$, at $b \in \mathcal{B} = \{0.2, 0.4, 0.5, 0.6, 0.8\}$, such that the fraction

[3] Length has been repeatedly shown in the past to correlate with reward [28; 21].

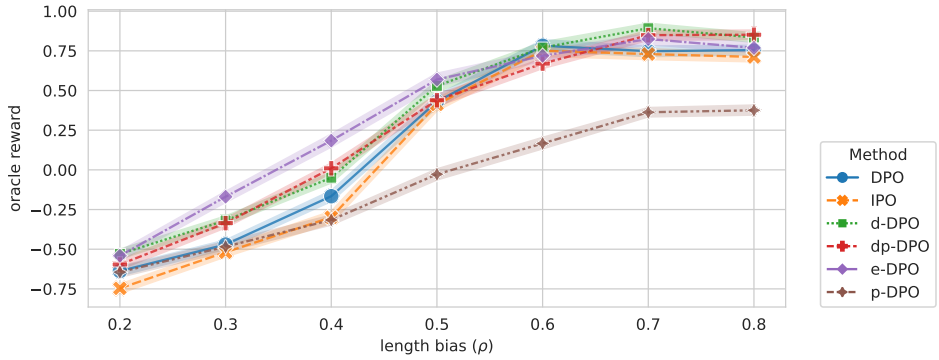

Figure 2: **Main results**, showing the advantage in oracle reward compared to the initial finetuned policy. Errorbars correspond to bootstrap 95% confidence intervals for finite sample variance. Ensemble DPO (e-DPO) is significantly better than DPO and IPO in the challenging setup where shorter responses are preferred ($\rho \leq 0.5$), and is generally the best-performing method overall in this regime. Distilled DPO (d-DPO) performs best when longer responses are preferred ($\rho > 0.6$).

of pairs where the longer response is preferred is b, and train reward models $r_{\rho,b}$ on those subsets. Consequently, sensitivity to length should vary across ensemble members. We then apply the same procedure as pd-DPO above, with a confidence set $\mathcal{S}_\rho = \{r_{\rho,b}\}_{b=1}^{\mathcal{B}}$.

All reward models and policies are initialized from Palm-2-XS [2]. Policies also go through a supervised finetuning step on human-written summaries from the original TL;DR training set [32] prior to alignment, and we term this policy π_{SFT}. We evaluate performance by sampling summaries for test set prompts, evaluating the average reward according to the oracle reward model, and computing the advantage in average reward compared to π_{SFT} (before alignment). We train policies for 10^4 steps with batch size 16 and learning rate 10^{-6}, and reward models for $3k$ steps with batch size 64 and learning rate 4×10^{-6}. We use the validation set for model selection during policy training and to choose the following hyperparameters. For all DPO variants, we sweep over $\beta \in \{.01, .1, 1, 3, 10, 30, 100\}$. For IPO, we sweep over $\tau \in \{0.01, 0.1, 1, 3, 5, 10, 25\}$. For all pessimistic methods we anneal $\gamma = \alpha/\beta$ from 10^{-4} to 10^{-2} linearly during the $10k$ training steps.

4.2 Results

We present the results of our experiment in Figure 2. As can be seen in the plot, the more challenging setting is when $\rho < 0.5$, which corresponds to a sample of preference annotations in which shorter outputs are generally preferred. This distribution shift is more difficult because as mentioned the oracle reward model (trained on human annotations) has a bias in favor of longer outputs [28]. Nevertheless we get sizable improvements compared to the reference policy π_{SFT} for all length bias values.

All approaches that invoke distillation (d-DPO, e-DPO, dp-DPO) outperform IPO and DPO ($p < .01$ by a Wald test) for $\rho \leq 0.5$, where shorter responses are preferred. Pessimistic ensemble DPO (e-DPO) performs particularly well in these settings, generally outperforming all methods that use a single reward model. When longer responses are preferred ($\rho > 0.6$), single reward distillation (d-DPO) leads to the highest performance, significantly outperforming both DPO and IPO ($p < .01$ by a Wald test). Interestingly, p-DPO does not provide empirical benefits relative to the distillation based methods, indicating that the distillation loss itself is quite important. For the effect of hyper-parameter selection, see Figure D.1. In DPO-based methods, the optimal value of β is inversely correlated with the bias; in IPO the same holds for the τ hyperparameter.

To better understand the utility of reward ensembles in e-DPO, in particular when $\rho < 0.5$, we examine the role of each reward model in the ensemble across different biases. Specifically, given the final e-DPO policy per length bias, for each example we identify the reward model $r_{\rho,b}$ that best matches the implicit reward of this policy, i.e., for which reward model is $\mathcal{L}_{\text{distill}}$ minimized on that example (see Eq. (7) and (11)). We find that when the policy is trained on data where shorter preference are preferred ($\rho < .5$), the reward model that best matches the policy often has the opposite bias (b is high), and vice versa. Thus, the success of e-DPO may be explained by its ability to distill from reward models that do not suffer from the bias in the policy training data, which is particularly

helpful when $\rho \leq .5$ as this bias is also not shared by the oracle RM. We provide the full distribution over reward models for all ρ and β in App. C. Overall, these results demonstrate the efficacy of training a policy by distilling from a reward model in the presence of distribution shifts, and that a careful design of an ensemble to mitigate spurious correlations can lead to further performance gains.[4]

5 Theoretical analysis

This section characterizes problems with the DPO objective and solutions offered by pessimistic DPO and distillation, focusing on the simplified scenario in which we optimize with respect to a single preference pairs (y^w, y^ℓ). Once again, all proofs are deferred to Appendix A.

In its Lagrangian formulation, pessimistic DPO adds a forward KL term to the DPO objective (§3.3). For the sake of analysis, we assume that the preference annotations are sampled from the reference distribution, $\mu(x) \times \pi_{\text{ref}}(y \mid x) \times \pi_{\text{ref}}(y \mid x)$. Then a finite-sample approximation of the forward KL term is $\hat{\Omega}(\Theta) := \sum_{(y^w, y^\ell) \in \mathcal{D}_{\text{Pref}}} -(\log \pi_\theta(y^\ell) + \log \pi_\theta(y^w))$. By applying this finite-sample approximation, *p-DPO has a finite optimum, unlike DPO*, as shown in Proposition 1. Note that this analysis is limited in two ways: (1) as mentioned, we compute the KL term over the completions in the preference data; (2) we directly optimize the probability ratios $\psi_w = \pi_\theta(y^w)/\pi_{\text{ref}}(y^w)$ and $\psi_\ell = \pi_\theta(y^\ell)/\pi_{\text{ref}}(y^\ell)$, rather than optimizing them jointly through the parameters. For sufficiently expressive π_θ, however, this approximation captures the behavior of the two algorithms reasonably well.

Proposition 3. *Let $\hat{\mathcal{L}}_{\text{pdpo}}$ represent a finite-sample approximation to $\mathcal{L}_{\text{pdpo}}$ with the empirical forward KL term $\hat{\Omega}(\Theta)$. For a fixed $\hat{\pi}_\theta(y_i^w)$ and $\alpha > 1$, the $\operatorname{argmin}_{\pi_\theta(y^\ell)} \hat{\mathcal{L}}_{\text{pdpo}}$ is $\min\left(1 - \hat{\pi}_\theta(y_i^w), \hat{\pi}_\theta(y_i^\ell)\right)$, with $\log \hat{\pi}_\theta(y_i^\ell) = -\frac{1}{\beta}\log(\alpha - 1) + \log \hat{\pi}_\theta(y_i^w) + \log \frac{\pi_{\text{ref}}(y_i^\ell)}{\pi_{\text{ref}}(y_i^w)}$.*

The optimum in Proposition 3 corresponds to $\log \psi_w/\psi_\ell = \beta^{-1}\log(\alpha - 1)$. Recall that IPO seeks to assign a constant value to this ratio by minimizing $(\log \frac{\psi_w}{\psi_\ell} - \tau^{-1})^2$; the (unconstrained) optima are identical for $\tau^{-1} := \beta^{-1}\log(\alpha - 1)$, but the loss surfaces are different (see Appendix B). DPO sets $\pi_\theta(y_i^\ell) \to 0$, as shown in Corollary 1; this is due not only to competition from $\pi_\theta(y_i^w)$ but from DPO penalizing positive probability on y_i^ℓ. Analysis of the distilled loss gives a similar result:

Proposition 4. *For any fixed $\hat{\pi}_\theta(y_i^w)$ and $\beta > 0$, the argmin of the distilled DPO objective (eq. (7)) is $\min(1 - \hat{\pi}_\theta(y_i^w), \hat{\pi}_\theta(y_i^\ell))$, with $\log \hat{\pi}_\theta(y_i^\ell) = \frac{1}{\beta}(r_t(x, y_i^\ell) - r_t(x, y_i^w)) + \log \hat{\pi}_\theta(y_i^w) + \log \frac{\pi_{\text{ref}}(y_i^\ell)}{\pi_{\text{ref}}(y_i^w)}$.*

While the setting is simplistic, the results are comforting: here the additional regularization effects of both distillation and pessimism (in the case of p-DPO) clearly help to avoid degenerate optima.

Why DPO can drive $\pi(y^w)$ to zero. In §2.3 we pointed out a peculiarity of the DPO global optima: in certain cases, it can include policies where $\pi(y^w)$ may be nearly 0 for all y^w in the training set. This undesirable behavior has also been observed in practice [20; 22; 30]. For intuition on why this may happen, consider the simplified case where the policy is a bag-of-words model, $\pi_\theta(y) \propto \exp(c(y) \cdot \theta)$ for $c(y)$ representing a vector of counts in y and θ_i representing the unnormalized log-probability of token i. Then we can formally show that DPO optimization monotonically decreases an upper bound on the probability of the *preferred* completion, $\tilde{\pi}_{\theta^{(t-1)}}(y^w) \geq \tilde{\pi}_{\theta^{(t)}}(y^w) \geq \pi_{\theta^{(t)}}(y^w)$.

Proposition 5. *Let $y^w, y^\ell \in \mathcal{V}^n$ be preferred vs. dispreferred outputs of length n, with $\pi_{\text{ref}}(y^w), \pi_{\text{ref}}(y^\ell) > 0$ and corresponding count vectors $c(y^w), c(y^\ell)$. Let $\log \pi_\theta(y) = c(y) \cdot \theta - nZ(\theta)$ for $Z(\theta) = \log \sum_i^{\mathcal{V}} e^{\theta_i}$, with upper bound $\log \tilde{\pi}_\theta(y) = c(y) \cdot \theta - n \max_j \theta_j$. Let $\theta^{(t)}$ represent the parameters of π after t steps of gradient descent on $\mathcal{L}_{\text{dpo}}(\{y^\ell, y^w, x\})$, with $\theta^{(0)} = 0$. Then $\pi_{\theta^{(t)}}(y^w) \leq \tilde{\pi}_{\theta^{(t)}}(y^w) \leq \tilde{\pi}_{\theta^{(t-1)}}(y^w)$ for all t.*

Where does the probability mass go? If $\pi_{\theta^{(t)}}(y^w)$ decreases in t, what other strings become more probable? In the following proposition, we show that under the bag-of-words model, DPO optimization moves probability mass away from y^w to sequences that contain only the tokens that maximize the difference between y^w and y^ℓ. This is a concrete example of the type of undesirable optima described in §2.3, now shown here to be realizable.

[4] We also experimented with an ensemble where members are different checkpoints across training of a reward model on the preference data and did not observe any empirical gains from this form of ensemble.

Proposition 6. *Let y^w and y^ℓ be preferred / dispreferred outputs of length n. Let $\Delta = c(y^w) - c(y^\ell)$ be the difference in unigram counts. Let $\hat{y} = [i, i, \ldots, i]$, for $i \in arg\max \Delta$, with $||c(\hat{y})||_1 = n$. Then $\pi_{\theta^{(t)}}(y^w) - \pi_{\theta^{(t)}}(\hat{y}) = \tau(t)k$ for some $k \leq 0$ and some non-decreasing $\tau : \mathbb{Z}_+ \to \mathbb{R}_+$.*

We have $k = 0$ when $c(y^w) = c(\hat{y})$, and $k \ll 0$ when $||c(y^w)||_2 \ll ||c(\hat{y})||_2 = n$ (dense $c(y^w)$) and $||\Delta||_2 = ||\Delta||_\infty$ (sparse Δ). This implies that when y^w and y^ℓ are similar, $\pi_\theta(y^w)$ will degrade more rapidly. Early stopping will therefore tradeoff between reaching the degenerate solution on such cases, and underfitting other cases in which y^w and y^ℓ are more distinct.

6 Related work

Recent work in offline alignment has focused on DPO [23] as a simpler alternative for aligning language models from preference data. Subsequent work has identified issues with DPO, including weak regularization [3] and a tendency to decrease the probability of winning generations during training [20]. Other methods have explored various avenues for improvement. These include analyzing the impact of noise on DPO alignment [11], proposing to update the reference policy during training [12], and suggesting a variant of IPO with a per-context margin [1]. Additional research has focused on token-level alignment methods [38; 22] and on developing a unified view of various offline alignment methods [31]. This work builds upon several these findings, and provides further analysis, as well as a solution based on pessimism and reward distillation.

While offline alignment methods are popular, recent evidence suggests that online alignment methods such as RLHF [6; 29], may lead to more favorable outcomes [13; 30; 8; 34]. Notably, Zhu et al. [41] proposed iterative data smoothing, which uses a trained model to softly label data during RLHF. Whether online or offline, however, policies are still succeptible to overfitting to certain degenerate phenomena. To this end, reward ensembles have been widely investigated recently as a mechanism for tackling reward hacking in RLHF [9; 7; 39; 25], and in the context of multi-objective optimization [19; 24]. We use an ensemble of rewards to represent the uncertainty with respect to reward models that are suitable given preference data. Moskovitz et al. [19] focus on "composite" rewards, with the goal of achieving high task reward while ensuring that every individual component is above some threshold—also by applying a Lagrangian relaxation. In this work, we also consider multiple reward models, but we only focus on cases where there is no known, obvious reward decomposition.

Finally, the question of using a small amount of offline data to learn high-quality policies, instead of online access to reward feedback, has been widely studied in the offline reinforcement learning (RL) literature. The predominant approach here is to use pessimism, that is, to learn a policy with the highest reward under all plausible environment models consistent with the data, with an extensive theoretical [18; 37; 33] and empirical [16; 5; 36] body of supporting work. The key insight in this literature is that without pessimism, the RL algorithm learns undesirable behaviors which are not explicitly ruled out in the training data, and pessimism provides a robust way of preventing such undesirable extrapolations, while still preserving generalization within the support of the data.

7 Conclusion

LM alignment is crucial for deploying safe and helpful assistants, but is difficult due to lack of access to perfect preference oracles. We presented a thorough theoretical analysis of some of the degeneracies that DPO is susceptible to when learning from sampled human preference data. Furthermore, our findings suggest that explicit reward modeling remains a powerful vehicle for introducing regularization into post-training. By distilling the reward assigned by a single, explicit reward model—or a family of explicit reward models—directly into the implicit reward maximized by our policies using offline data, we demonstrated that we can achieve improved robustness to variations in preference dataset quality, while maintaining the simplicity of the DPO framework.

Limitations. The empirical results in the paper are based on one dataset and form of distribution shift. For deeper understanding of pessimism and ensembling, additional settings should be explored. The theoretical aspects of the paper are sometimes based on restrictive assumptions and simplifications. Nonetheless, they provide potential explanations for phenomena observed in real-world settings.

Broader impact. We introduce new ideas to the active field of research on preference-based post-training, which we hope will help facilitate the alignment of large models, and improve understanding of current approaches—ultimately supporting the development of capable and reliable AI systems.

References

[1] Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

[2] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023.

[3] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[4] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444. URL http://www.jstor.org/stable/2334029.

[5] Ching-An Cheng, Tengyang Xie, Nan Jiang, and Alekh Agarwal. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pages 3852–3878. PMLR, 2022.

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[7] Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

[8] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. 2024.

[9] Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? Reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

[10] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1607–1616. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/furlanello18a.html.

[11] Yang Gao, Dana Alon, and Donald Metzler. Impact of preference noise on the alignment performance of generative language models. *arXiv preprint arXiv:2404.09824*, 2024.

[12] Alexey Gorbatovski, Boris Shaposhnikov, Alexey Malakhov, Nikita Surnachev, Yaroslav Aksenov, Ian Maksimov, Nikita Balagansky, and Daniil Gavrilov. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*, 2024.

[13] Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531.

[15] Tomasz Korbak, Ethan Perez, and Christopher Buckley. RL with KL penalties is better viewed as bayesian inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1083–1091, 2022.

[16] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

[17] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

[18] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274, 2020.

[19] Ted Moskovitz, Aaditya K Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca D Dragan, and Stephen McAleer. Confronting reward model overoptimization with constrained rlhf. *arXiv preprint arXiv:2310.04373*, 2023.

[20] Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

[21] Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

[22] Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function. 2024.

[23] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards, 2023.

[25] Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. Warm: On the benefits of weight averaged reward models. 2024.

[26] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *In Proceedings of ICLR*, 2015.

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. 2017.

[28] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. 2023.

[29] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[30] Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of llms should leverage suboptimal, on-policy data, 2024.

[31] Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. 2024.

[32] Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In Lu Wang, Jackie Chi Kit Cheung, Giuseppe Carenini, and Fei Liu, editors, *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4508. URL https://aclanthology.org/W17-4508.

[33] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

[34] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. 2024.

[35] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan L. Yuille. Training deep neural networks in generations: a more tolerant teacher educates better students. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33015628. URL https://doi.org/10.1609/aaai.v33i01.33015628.

[36] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems*, 34:28954–28967, 2021.

[37] Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

[38] Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. 2024.

[39] Yuanzhao Zhai, Han Zhang, Yu Lei, Yue Yu, Kele Xu, Dawei Feng, Bo Ding, and Huaimin Wang. Uncertainty-penalized reinforcement learning from human feedback with diverse reward lora ensembles. 2023.

[40] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. Slic-hf: Sequence likelihood calibration with human feedback. 2023.

[41] Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Iterative data smoothing: Mitigating reward overfitting and overoptimization in rlhf. 2024.

A Proofs

A.1 Proof of Proposition 1

Proof. Since all the preference pairs (y, y') are mutually disjoint, and θ_y is specific to each y, the DPO objective over $\mathcal{D}_{\text{pref}}$ is convex in $\Delta = \{\Delta_1, \ldots, \Delta_n\}$, where

$$\Delta_i = \beta \log \frac{\pi_\theta(y_i^w)\pi_{\text{ref}}(y_i^\ell)}{\pi_\theta(y_i^\ell)\pi_{\text{ref}}(y_i^w)}. \tag{15}$$

Furthermore, the different Δ_i are completely independent from each other due to the preference pairs being disjoint, so they can be optimized over separately.

In particular, for every i we have that

$$\lim_{\Delta_i \to \infty} -\log\left(\sigma\left(\Delta_i\right)\right) = 0, \tag{16}$$

which implies that $\Delta^* = \{\infty\}^n$ is the unique global minimizer of the DPO loss over $\mathcal{D}_{\text{pref}}$ in the space of Δ's, and any θ^* that is a global minimizer must therefore satisfy

$$\log \frac{\pi_\theta(y_i^w)\pi_{\text{ref}}(y_i^\ell)}{\pi_\theta(y_i^\ell)\pi_{\text{ref}}(y_i^w)} = \infty. \tag{17}$$

□

A.2 Proof of Corollary 1

Proof. Following the same argument of the proof of Proposition 1, we have that all global minimizers θ^* of the DPO satisfy $\Delta_i^* = \infty$, which in turn implies that

$$\frac{\pi_{\theta^*}(y_i^w)\pi_{\text{ref}}(y_i^\ell)}{\pi_{\theta^*}(y_i^\ell)\pi_{\text{ref}}(y_i^w)} = \infty. \tag{18}$$

Since $\pi_{\text{ref}}(y)$ is assumed to satisfy $0 < \pi_{\text{ref}}(y) < 1$ for all y, this implies that all θ^* satisfy

$$\frac{\pi_{\theta^*}(y_i^w)}{\pi_{\theta^*}(y_i^\ell)} = \infty, \tag{19}$$

which further implies that $\pi_{\theta^*}(y_i^\ell) = 0$ and $\pi_{\theta^*}(y_i^w) > 0$ for all $i \in [n]$, as $\pi_{\theta^*}(y_i^w) \leq 1$ for any y_i^w. Aggregating

$$\mathcal{C}(y_\ell) = \{y \colon \exists i \in [n] \text{ s.t } y_i^\ell = y\} \tag{20}$$

then gives that

$$\pi_{\theta^*}(\mathcal{C}(y_\ell)) = \sum_{y \in \mathcal{C}(y_\ell)} \pi_{\theta^*}(y) = 0 \Longrightarrow \pi_{\theta^*}(\mathcal{C}(y_\ell)^c) = 1. \tag{21}$$

□

To prove the converse, let $\pi_{\theta'}$ be a policy that satisfies $\pi_{\theta'}(\mathcal{C}(y^\ell)^c) = 1$, with $\pi_{\theta'}(y_i^w) > 0, \forall i \in [n]$,. As $\pi_{\theta'}(y) \geq 0$ for all y, this implies that $\pi_{\theta'(y_i^\ell)} = 0 \; \forall i \in [n]$. Then, we have

$$\frac{\pi_{\theta'}(y_i^w)}{\pi_{\theta'}(y_i^\ell)} = \infty, \tag{22}$$

which by Proposition 1 implies that $\pi_{\theta'}$ is a global optimum.

A.3 Proof of Theorem 1

Proof. We know that the optimal policy for the RLHF objective (1) is given by $\pi_{\theta^*}(y|x) \propto \pi_{\text{ref}}(y|x)\exp(r^*(x,y)/\beta)$. Plugging this policy into the distillation objective (7), we see that $\mathcal{L}_{\text{distill}}(r^*, \pi_{\theta^*}, \rho) = 0$ for all ρ. In fact, the loss is equal to 0 pointwise, meaning that π_{θ^*} is

a global minimizer of the distillation objective (7). Further, let π be some other minimizer of $\mathcal{L}_{\text{distill}}(r^*, \cdot, \rho)$. Then π also has to attain a loss of 0 at all (x, y, y') in the support of ρ, meaning that $\log \pi(y|x) - \log \pi(y'|x) = \log \pi_{\theta^*}(y|x) - \log \pi_{\theta^*}(y|x)$ for all (x, y, y') in the support of ρ. Consequently, the two policies coincide in the support of ρ (due to the normalization constraint, there is no additional offset term allowed as the support of ρ covers all of $\mathcal{Y}$). Finally, noting that the support of the chosen ρ is such that π_{θ^*} puts no mass outside its support due to the KL constraint in (1), we complete the proof. □

A.4 Proof of Theorem 2

Proof. Consider the pessimistic objective:

$$\max_{\pi_\theta} \min_{r_{\text{tgt}} \in \mathcal{S}} \mathbb{E}_{\mu(x)} \Big[\mathbb{E}_{\pi_\theta(y|x)}[r_{\text{tgt}}(x, y)] - \mathbb{E}_{\pi_{\text{ref}}(y|x)}[r_{\text{tgt}}(x, y)] \Big] - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}). \tag{23}$$

As it is linear in r_{tgt} and convex in π, we can switch the order of $\min$ and $\max$:

$$\min_{r_{\text{tgt}} \in \mathcal{S}} \left[\max_{\pi \in \Pi} \mathbb{E}_{\mu(x)} \Big[\mathbb{E}_{\pi(y|x)}[r_{\text{tgt}}(x, y)] - \mathbb{E}_{\pi_{\text{ref}}(y|x)}[r_{\text{tgt}}(x, y)] \Big] - \beta \mathbb{D}_{\text{KL}}(\pi \| \pi_{\text{ref}}) \right]. \tag{24}$$

Note that every $r_{\text{tgt}} \in \mathcal{S}$ can be written in terms of the KL-constrained policy $\pi^*_{r_{\text{tgt}}}$ it induces, i.e.,

$$r_{\text{tgt}}(x, y) = \beta \log \frac{\pi^*_{r_{\text{tgt}}}(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x, r_{\text{tgt}}), \tag{25}$$

where

$$\pi^*_{r_{\text{tgt}}} = \operatorname*{argmax}_{\pi_\theta} \mathbb{E}_{\mu(x)} \mathbb{E}_{\pi_\theta(y|x)}[r_{\text{tgt}}(x, y)] - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \tag{26}$$

which has the form

$$\pi^*_{r_{\text{tgt}}}(y \mid x) = \frac{1}{Z(x, r_{\text{tgt}})} \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r_{\text{tgt}}(x, y) \right) \tag{27}$$

where $Z(x, r_{\text{tgt}})$ is the partition function:

$$Z(x, r_{\text{tgt}}) = \sum_{y \in \mathcal{Y}} \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r_{\text{tgt}}(x, y) \right). \tag{28}$$

Substituting $\pi^*_{r_{\text{tgt}}}$ in for $\max_\pi$ and writing r_{tgt} in terms of $\pi^*_{r_{\text{tgt}}}$, we get the simplified objective

$$\begin{aligned}
&\min_{r_{\text{tgt}} \in \mathcal{S}} \left[\max_{\pi \in \Pi} \mathbb{E}_{\mu(x)} \Big[\mathbb{E}_{\pi(y|x)}[r_{\text{tgt}}(x, y)] - \mathbb{E}_{\pi_{\text{ref}}(y|x)}[r_{\text{tgt}}(x, y)] \Big] - \beta \mathbb{D}_{\text{KL}}(\pi \| \pi_{\text{ref}}) \right] \\
&\quad = \min_{r_{\text{tgt}} \in \mathcal{S}} \Bigg[\mathbb{E}_{\mu(x)} \Bigg[\mathbb{E}_{\pi^*_{r_{\text{tgt}}}(y|x)} \left[\beta \log \frac{\pi^*_{r_{\text{tgt}}}(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x, r_{\text{tgt}}) \right] \\
&\qquad\qquad - \mathbb{E}_{\pi_{\text{ref}}(y|x)} \left[\beta \log \frac{\pi^*_{r_{\text{tgt}}}(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x, r_{\text{tgt}}) \right] \\
&\qquad\qquad - \beta \mathbb{D}_{\text{KL}}(\pi^*_{r_{\text{tgt}}} \| \pi_{\text{ref}} \mid x) \Bigg] \Bigg] \\
&\quad = \min_{r_{\text{tgt}} \in \mathcal{S}} \beta \left[\mathbb{E}_{\mu(x)} \left[\mathbb{D}_{\text{KL}}(\pi^*_{r_{\text{tgt}}} \| \pi_{\text{ref}} \mid x) + \mathbb{D}_{\text{KL}}(\pi_{\text{ref}} \| \pi^*_{r_{\text{tgt}}} \mid x) - \mathbb{D}_{\text{KL}}(\pi^*_{r_{\text{tgt}}} \| \pi_{\text{ref}} \mid x) \right] \right] \\
&\quad = \min_{r_{\text{tgt}} \in \mathcal{S}} \beta \mathbb{E}_{\mu(x)} \left[\mathbb{D}_{\text{KL}}(\pi_{\text{ref}} \| \pi^*_{r_{\text{tgt}}} \mid x) \right].
\end{aligned} \tag{29}$$

□

A.5 Proof of Proposition 2

Proof. The proof is a standard Lagrangian duality argument, which we reproduce here for completeness. For two functions $f(z)$ and $g(z)$, let us define

$$z^* = \operatorname*{argmin}_z f(z) + \alpha g(z). \tag{30}$$

Let us also consider the constrained problem

$$z' = \operatorname*{argmin}_z f(z) \quad \text{s.t. } g(z) \leq g(z^*). \tag{31}$$

Suppose by contradiction that z^* is not a minimizer of (31). Since z^* is feasible for the constraint by construction, we get that $f(z') < f(z^*)$. Consequently, we further have

$$f(z') + \alpha g(z') < f(z^*) + \alpha g(z^*),$$

where the inequality follows from the feasibility of z' in (31). This contradicts the optimality of z^* in (30), meaning that z^* must be a minimizer of (31). Applying this general result with $f = \beta \mathbb{E}_{\mu(x)} \mathbb{D}_{\mathrm{KL}}(\pi_{\mathrm{ref}}(y \mid x) \| \pi_\theta(y \mid x))$, $g = \min_{r^i_{\mathrm{tgt}} \in \mathcal{S}} \mathcal{L}_{\mathrm{distill}}(r^i_{\mathrm{tgt}}, \pi_\theta; \rho)$, and $z = \pi_\theta$ completes the proof, since we recognize the set $\mathcal{S}_\gamma$ in (12) to be equivalent to $\bigcup_{r^i_{\mathrm{tgt}} \in \mathcal{S}} \mathcal{L}_{\mathrm{distill}}(r^i_{\mathrm{tgt}}, \pi_\theta; \rho) \leq \lambda$.

□

A.6 Proof of Proposition 3

Proof. We differentiate $\mathcal{L}_{\mathrm{pdpo}}$ with respect to $\psi_\ell = \pi_\theta(y^\ell)/\pi_{\mathrm{ref}}(y^\ell)$ with i implicit, obtaining,

$$\frac{\partial \mathcal{L}_{\mathrm{pdpo}}}{\partial \psi_\ell} = \beta \frac{\psi_\ell^\beta}{\psi_w^\beta + \psi_\ell^\beta} \psi_\ell^{-1} - \frac{\beta}{\alpha} \psi_\ell^{-1} = \beta \psi_\ell^{-1} \left(\frac{\psi_\ell^\beta}{\psi_w^\beta + \psi_\ell^\beta} - \alpha^{-1} \right) \tag{32}$$

which is zero when,

$$\alpha \psi_\ell^\beta = \psi_w^\beta + \psi_\ell^\beta \tag{33}$$

$$\psi_\ell = \left(\frac{1}{\alpha - 1} \right)^{1/\beta} \psi_w \tag{34}$$

$$\log \psi_\ell = -\frac{1}{\beta} \log(\alpha - 1) + \log \psi_w \tag{35}$$

$$\log \pi_{\hat{\theta}}(y^\ell) = \log \pi_{\mathrm{ref}}(y^\ell) - \frac{1}{\beta} \log(\alpha - 1) + \log \pi_\theta(y^w) - \log \pi_{\mathrm{ref}}(y^w). \tag{36}$$

By the second-order condition, the critical point is a minimum. The objective $\mathcal{L}_{\mathrm{pdpo}}$ is the sum of two components: the negative log sigmoid term for $\mathcal{L}_i$ and the negative log probability for $\hat{\Omega}$. Because each component is a convex function of ψ_i, so is $\mathcal{L}_{\mathrm{pdpo}}$. As a result, the local minimum $\log \hat{\pi}_\theta(y^\ell)$ is also a global minimum. □

A.7 Proof of Proposition 4

Proof. This follows directly from differentiating eq. (7) with respect to $\pi_\theta(y_2)$. □

A.8 Proof of Proposition 5

Proof. Let $\Delta = [c(y^w) - c(y^\ell)]$ and $\rho = \pi_{\mathrm{ref}}(y^w)/\pi_{\mathrm{ref}}(y^\ell)$. The theorem assumes $|y^w| = |y^\ell|$. Then $\mathcal{L}_{\mathrm{dpo}} = -\log \sigma\left(\beta(\Delta \cdot \theta) + \beta \log \rho\right)$. The derivative with respect to θ is,

$$\frac{\partial \mathcal{L}_\beta(\theta)}{\partial \theta} = -\left(1 - \sigma(\beta(\Delta \cdot \theta) + \beta \log \rho)\right) \beta \Delta = -\Pr(y^\ell \succ y^w; \theta) \beta \Delta \prec 0. \tag{37}$$

Let $\delta_t = \beta \Pr(y^\ell \succ y^w; \theta^{(t)})$. Then,

$$\tilde{\pi}_{\theta^{(t)}} = \theta^{(t)} \cdot c(y^w) - n \max_j \theta_j^{(t)} \tag{38}$$

$$= (\theta^{(t-1)} + \delta_t \Delta) \cdot c(y^w) - n \max_j (\theta_j^{(t-1)} + \delta_t \Delta_j) \tag{39}$$

$$= \theta^{(t-1)} \cdot c(y^w) - n \max_j \theta_j^{(t-1)} + \delta_t \Delta \cdot c(y^w) - n\delta_t \max_j \Delta_j \tag{40}$$

$$= \tilde{\pi}_{\theta^{(t-1)}} + \delta_t \left(\Delta \cdot c(y^w) - n \max_j \Delta_j \right) \tag{41}$$

$$= \tilde{\pi}_{\theta^{(t-1)}} + \delta_t \sum_j^{\mathcal{V}} c_j(y^w)(\Delta_j - \max_{j'} \Delta_{j'}) \leq \tilde{\pi}_{\theta^{(t-1)}}. \tag{42}$$

We obtain $\max_j \left(\theta_j^{(t-1)} + \delta_t \Delta_j \right) = \max_j \theta_j^{(t-1)} + \max_j \delta_t \Delta_j$ from the fact that $\theta^{(0)} = 0$ and therefore $j \in \arg\max \Delta$ implies $j \in \arg\max \theta^{(t')}$ for all $t' > 0$. The second-to-last step uses $n = \sum_j^{\mathcal{V}} c_j(y^w)$ and the final step uses $\Delta_j \leq \max'_j \Delta_{j'}$. Finally, we have $\pi_{\theta^{(t)}}(y) \leq \tilde{\pi}_{\theta^{(t)}}(y^w)$ because $Z(\theta) = \log \sum_j \exp \theta_j \geq \log \max_j \exp \theta_j = \max_j \theta_j$. □

A.9 Proof of Proposition 6

Proof. Applying gradient descent with learning rate η to the gradient from Equation (37), at each step t the parameters are,

$$\theta^{(t)} = \theta^{(t-1)} + \eta\beta \Pr(y^\ell \succ y^w; \theta^{(t-1)})\Delta = \left(\sum_{t'=1}^{t} \eta\beta \Pr(y^\ell \succ y^w; \theta^{(t')}) \right) \Delta = \tau(t)\Delta. \tag{43}$$

Plugging these parameters into the likelihoods,

$$\ell_{\theta^{(t)}}(c(y^w)) - \ell_{\theta^{(t)}}(\hat{y}) = c(y^w) \cdot \theta^{(t)} - nZ(\theta^{(t)}) - c(\hat{y}) \cdot \theta^{(t)} + nZ(\theta^{(t)}) \tag{44}$$

$$= (c(y^w) - c(\hat{y})) \cdot \theta^{(t)} = (c(y^w) - c(\hat{y})) \cdot (\tau(t)\Delta) \tag{45}$$

$$= \tau(t)(c(y^w) \cdot \Delta - n \max \Delta) = \tau(t)k, \tag{46}$$

with $k \leq 0$ by $c(y^w) \cdot \Delta \leq ||c(y^w)||_1 \times ||\Delta||_\infty = n \max \Delta$. □

B Transitive closure

Both p-DPO and IPO target a constant ratio for $\log \psi_w / \psi_l$. However, the loss surfaces are different. To see this, we consider a simplified setting with three possible outputs, y_1, y_2, y_3. We observe either $\mathcal{D} = \{(y_1 \prec y_2), (y_2 \prec y_3)\}$ or $\overline{\mathcal{D}} = \mathcal{D} \cup \{(y_1 \prec y_3)\}$. If we treat this problem as a multi-arm bandit, the goal is to assign a weight to each arm, which we denote $\psi_i = \log \pi_\theta(y_i|x) + Z_x$, with Z_x an underdetermined log-partition function.

Proposition 7. *Let $\mathcal{D} = \{(i, i+1) : i \in 1, 2, \ldots, n\}$ for $n > 2$. Let $\overline{\mathcal{D}}$ be the dataset arising from the transitive closure of $\mathcal{D}$. Assume π_{ref} is indifferent to all (y_i, y_j). Let $\psi_\infty^{(\mathcal{D})} = \max_i \psi_i^{(\mathcal{D})} - \min_i \psi_i^{(\mathcal{D})}$. Then $\psi_\infty^{(\mathcal{D})} = (n-1)\tau^{-1} > \psi_\infty^{(\overline{\mathcal{D}})} = 2\frac{n-1}{n}\tau^{-1}$.*

Proof. For $\mathcal{D}$, the IPO objective can be minimized at zero, so that $\psi_\infty^{(\mathcal{D})} = (n-1)\tau^{-1}$. For $\overline{\mathcal{D}}$, each adjacent pair of completions is separated by γ, and the objective is $\sum_{i=1}^{n-1}(n-i)(i\gamma - \tau^{-1})^2$. The minimum is $\gamma = \frac{n(n+1)(n-1)/6}{n^2(n+1)(n-1)/12}\tau^{-1} = \frac{2}{n}\tau^{-1}$, so that $\psi_\infty^{(\overline{\mathcal{D}})} = (n-1)\gamma = 2\frac{n-1}{n}\tau^{-1} < (n-1)\tau^{-1} = \psi_\infty^{(\mathcal{D})}$ for $n > 2$. □

Intuitively, the observation of $(y_1 \prec y_3)$ should increase confidence that y_3 is superior to y_1, but in IPO it has the opposite effect, drawing their scores closer together. While pessimistic DPO also has a target ratio between each preference pair, its loss surface is different: in particular, it does not

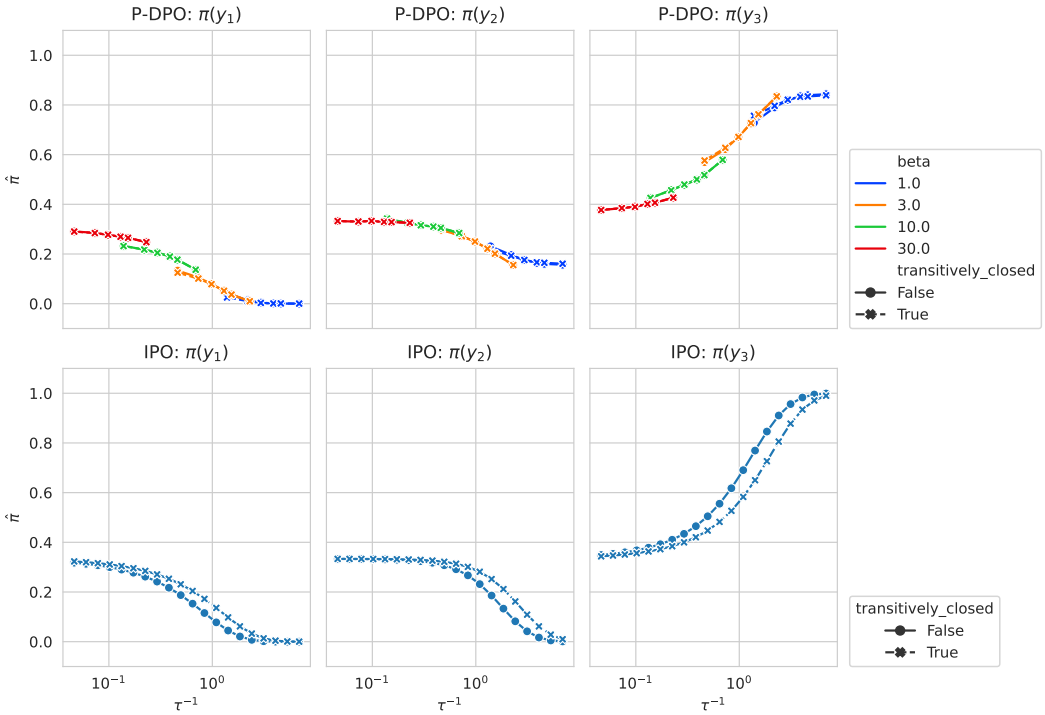

Figure B.1: **Effect of transitive closure on p-DPO and IPO solutions to preference learning in a multi-arm bandit**. Each column shows the learned policy probability for a given arm, based on the preferences $y_1 \prec y_2 \prec y_3$. The top row shows that in p-DPO, the probabilities are not materially affected by the transitive closure $y_1 \prec y_3$. The bottom row shows that in IPO, transitive closure causes the probabilities to be compressed. In each subfigure, we sweep a range of effective values of τ^{-1}, shown on the x-axis.

increase quadratically as we move away from the target. We find empirically that pessimistic DPO is robust to the transitive closure of preference annotations in the multi-arm bandit setting, as shown in Figure B.1. As discussed above, DPO will set $\psi_1 \to -\infty$ because y_1 is never preferred.

In our empirical experiments we solve the p-DPO and IPO objectives for both $\mathcal{D} = \{(y_1, y_2), (y_2, y_3)\}$ and $\overline{\mathcal{D}} = \mathcal{D} \cup \{(y_1, y_3)\}$, solving with respect to $\{\pi_\theta(y_i)\}$. IPO is solved analytically as a quadratic program; for pessimistic DPO we used projected gradient descent. We consider $\beta \in (1, 3, 10, 30)$ and $\alpha \in (5, 10, 20, 50, 100, 1000)$. As shown in Figure B.1, there are significant differences in the IPO solutions with and without transitive closure, while for p-DPO these differences are imperceptible.

C Distribution over reward models for e-DPO

Figure C.1 investigates the reason for the success of e-DPO, especially when $\rho < .5$. For every length bias, we show across all training examples the fraction of cases where a certain reward model, $r_{\rho,b}$, best matched the implicit reward of the final e-DPO policy. The policy matches different reward models in different examples. Moreover, there is inverse correlation between the data bias for policy training (ρ) and the data bias for training the reward models (b). This suggests that the ensemble in e-DPO helps as the policy is distilling from reward models that do not share the data bias of the policy training set.

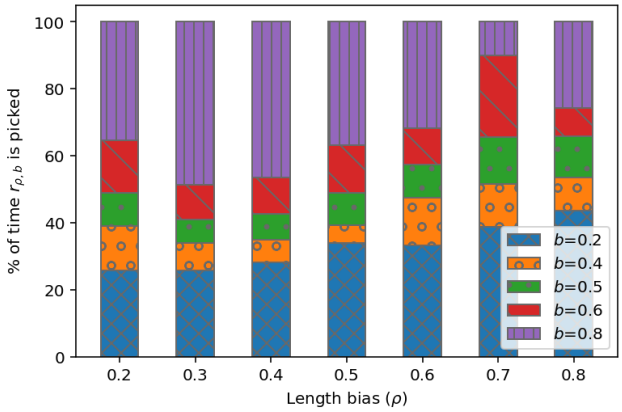

Figure C.1: We show for every length bias, ρ, the distribution over reward models that best match the final policy trained by e-DPO across all training examples. We observe that the e-DPO policy matches different reward models across examples. Moreover, when the policy is trained with data biased towards preferring short responses, the reward model that was trained on longer responses is often preferred and vice versa.

D Hyperparameters

Validation set performance across the range of hyperparameter settings is shown in Figure D.1. In pilot studies we found that these results were relatively robust to variation in the random seed, but did not conduct extensive investigation of this effect across all methods and hyperparameters due to cost.

E Compute resources

We train policies on 32 TPU v3 chips and reward models on 16 TPU v3 chips. We obtain roughly 0.1 steps per second when training, for both the policy and reward models.

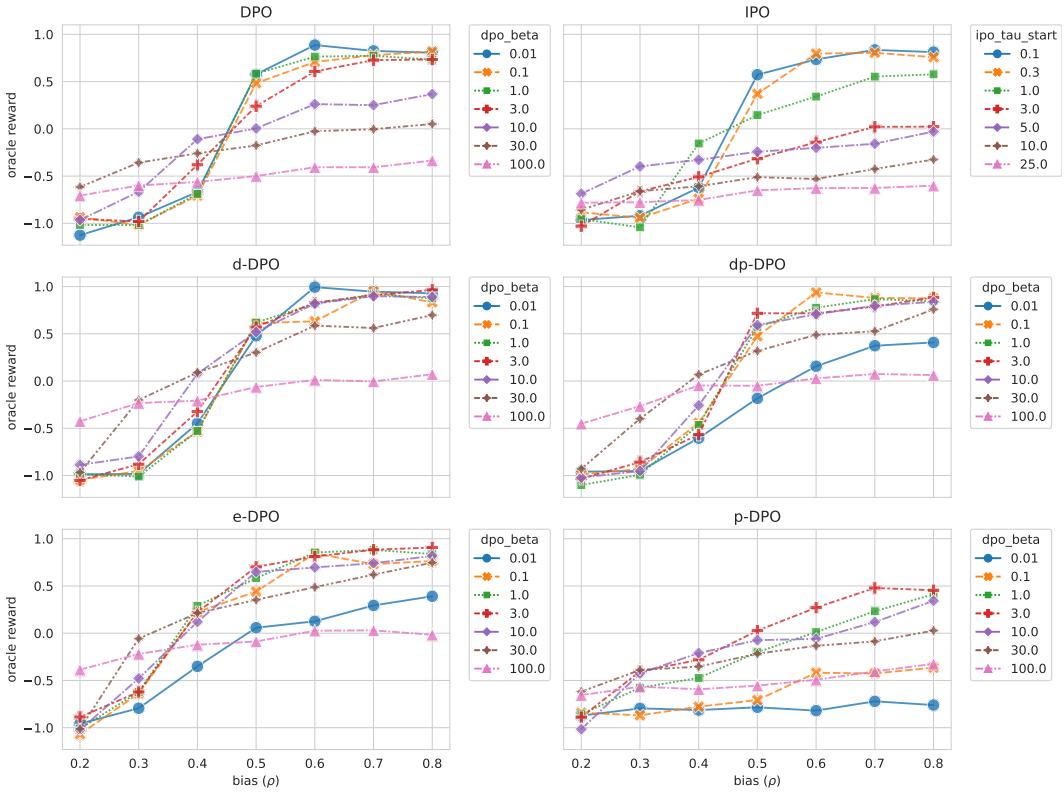

Figure D.1: **Validation set results** across hyperparameters for each method. For all methods, different values of ρ induce different optimal hyperparameters β and τ^{-1}.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In our view, the abstract and introduction accurately summarize the contributions of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: See Section 7

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See Appendix A

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details are provided in Section 4.1 and Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Experiments are on publicly-available data, but it is not possible for us to share code. We believe that the implementation should be relatively straightforward, given the mathematical descriptions presented here.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: These details are provided in Section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 4.2 includes bootstrap 95% confidence intervals on the main figure and hypothesis tests for specific comparisons between methods.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research does not involve human subjects and does not introduce new data. Its main impact should be to improve effectiveness and understanding of preference-based post-training.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 7

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No data or models are released.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The main external resource is the TLDR dataset, which we cite. Its license is CC BY 4.0.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets are introduced.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.