On Extending Direct Preference Optimization to Accommodate Ties

Anonymous authors
Paper under double-blind review

Abstract

We derive and investigate two DPO variants that explicitly model the possibility of declaring a tie in pair-wise comparisons. We replace the Bradley-Terry model in DPO with two well-known modeling extensions, by Rao and Kupper and by Davidson, that assign probability to ties as alternatives to clear preferences. Our experiments in neural machine translation and summarization show that explicitly labeled ties can be added to the datasets for these DPO variants without the degradation in task performance that is observed when the same tied pairs are presented to DPO. We find empirically that the inclusion of ties leads to stronger regularization with respect to the reference policy as measured by KL divergence, and we see this even for DPO in its original form. These findings motivate and enable the inclusion of tied pairs in preference optimization as opposed to simply discarding them.

1 Introduction

The original formulation of DPO (Rafailov et al., 2023) does not allow for ties. DPO requires training data consisting of paired options, $y_w \succ y_l$, and each of these pairs should represent a clear preference in judgment with no ambiguity as to which is the winner and which is the loser. From this data, the DPO learning procedure encourages the underlying policy to prefer y_w over y_l. This formulation does not allow for any ambiguity or uncertainty in the comparison of the paired examples in the training data.

This certainty is not easy to achieve in practice. A common approach is simply to discard data. Dubey et al. (2024, Sec. 4.2.1) apply DPO in post-training of Llama 3 models and note that for "DPO, we use samples that are labeled as the chosen response being significantly better or better than the rejected counterpart for training and discard samples with similar responses." Similarly, Qwen2 developers (Yang et al., 2024a, Sec. 4.3) "sample multiple responses from the current policy model, and the reward model selects the most and the least preferred responses, forming preference pairs that are used for DPO." Over-generation followed by aggressive selection is effective in producing the strongly ordered judgments needed for DPO. However the process appears wasteful: many potentially useful, and expensively collected, preference judgments are discarded simply because they are ties. As Rao and Kupper (1967) note: "any model which does not allow for the possibility of ties is not making full use of the information contained in the no-preference class."

Motivated by this, we investigate DPO variants that can incorporate ties. We replace the Bradley-Terry preference model at the heart of DPO by two well-known extensions by Rao and Kupper (1967) and by Davidson (1970) that explicitly assign probability to tied judgments alongside winners and losers. Since these models are generalizations of the Bradley-Terry model, we find that they are readily incorporated into the DPO modeling framework. In experiments in neural machine translation and summarization, we find that ties can be added to the datasets for these DPO variants without the degradation in task performance that results from adding ties to the original DPO. We also observe improved regularization, in reduced KL-divergence to the reference policy, by adding ties.

2 Methodology

2.1 DPO and the Bradley-Terry Preference Distribution

The Bradley-Terry model assigns probability that an item y_i will be preferred to item y_j in terms of their 'strength' parameters λ. In the RLHF setting, strengths are expressed as rewards r, $\lambda = e^r$ (Rafailov et al., 2023, Eq. 1), so that the preference distribution for item i over item j depends on the difference in their rewards, $d_{ij} = r_i - r_j$

$$p^{BT}(y_i \succ y_j) = \frac{\lambda_i}{\lambda_i + \lambda_j} = \frac{e^{r_i}}{e^{r_i} + e^{r_j}} = \sigma(r_i - r_j) = \sigma(d_{ij}) \tag{1}$$

One of the enabling observations made by Rafailov et al. (2023) is that when a policy π_θ is sought to maximize the KL-regularized objective $\max_{\pi_\theta} \mathbb{E}\,[r(x,y)] - \beta\, D(\pi_\theta(y|x)\;||\;\pi_{ref}(y|x))$, the reward associated with the policy has the form $r_\theta(x,y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} + \beta \log Z_\theta(x)$. This allows expressing the difference in rewards between hypotheses y_w and y_l under a parameterized policy π_θ as the reward margin

$$d_\theta(x, y_w, y_l) = r_\theta(x, y_w) - r_\theta(x, y_l) = \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)} \tag{2}$$

so that the corresponding Bradley-Terry probability that item y_w beats item y_l is

$$p_\theta^{BT}(y_w \succ_x y_l) = \sigma(d_\theta(x, y_w, y_l)) = \sigma(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}). \tag{3}$$

The DPO policy objective (Rafailov et al., 2023, Eq. 7) follows by incorporating the parameterized form of the preference distribution into a maximum likelihood objective

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{x,y_w,y_l} \log p_\theta(y_w \succ_x y_l) \tag{4}$$

$$= -\mathbb{E}_{x,y_w,y_l} \log \sigma(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}) \tag{5}$$

We note that Eq. 2 follows from the regularized risk optimization (Rafailov et al., 2023, A.1). It does not rely on any assumption that limits its use to the Bradley-Terry model.

2.2 Bradley-Terry Extensions that Accommodate Ties

An observed weakness of the Bradley-Terry model is that it does not allow for ties. Unless two items have exactly the same strengths (so that $d_{ij} = 0$), the model always assigns a higher probability of winning to the stronger item. This may be reasonable if one item is much stronger than the other, but when items are relatively comparable it may be desirable to allow some probability for tied outcomes.

The Rao-Kupper (Rao and Kupper, 1967) model assigns win and tie probabilities as:

$$p^{RK}(y_i \succ y_j) = \frac{\lambda_i}{\lambda_i + \nu_{RK}\lambda_j} \qquad \text{item } y_i \text{ beats item } y_j \tag{6}$$

$$p^{RK}(y_i \sim y_j) = \frac{(\nu_{RK}^2 - 1)\lambda_i\lambda_j}{(\lambda_i + \nu_{RK}\lambda_j)(\lambda_j + \nu_{RK}\lambda_i)} \qquad \text{items } y_i \text{ and } y_j \text{ tie} \tag{7}$$

while the Davidson (Davidson, 1970) model assigns win and tie probabilities as:

$$p^{D}(y_i \succ y_j) = \frac{\lambda_i}{\lambda_i + \lambda_j + 2\nu_D\sqrt{\lambda_i\lambda_j}} \qquad \text{item } y_i \text{ beats item } y_j \tag{8}$$

$$p^{D}(y_i \sim y_j) = \frac{2\nu_D\sqrt{\lambda_i\lambda_j}}{\lambda_i + \lambda_j + 2\nu_D\sqrt{\lambda_i\lambda_j}} \qquad \text{items } y_i \text{ and } y_j \text{ tie} \tag{9}$$

The probabilities of the three outcomes sum to one for both of these Bradley-Terry extensions: $p(y_i \succ y_j) + p(y_j \succ y_i) + p(y_i \sim y_j) = 1$. For both models, $p(y_i \sim y_j) = p(y_j \sim y_i)$ and

$p(y_i \sim y_j)$ tends towards 0 if $\lambda_j \gg \lambda_i$. Both variants have parameters ν that control how much probability is allocated to ties. Apart from $\nu_{RK} = 1$ or $\nu_D = 0$, when both variants agree with Bradley-Terry, some probability is reserved for tied outcomes.

The Rao-Kupper and Davidson models arise from different considerations. Rao and Kupper (1967) begin with the formulation $p^{BT}(y_i \succ y_j) = \frac{1}{4}\int_{-(r_i-r_j)}^{\infty} \text{sech}^2(y/2)dy$ (Bradley, 1953, Eq. 13) and note its sensitivity to the difference in values $r_i - r_j$. They note that some judges "may not be able to express any real preference" in paired-comparisons if their "sense of perception is not sharp enough" to detect small differences. They reason that a "threshold of sensory perception" is needed such that if the observed difference is less than the threshold, a judge declares a tie. They introduce the sensitivity threshold α_{RK} as follows, $p^{RK}(y_i \succ y_j) = \frac{1}{4}\int_{-(r_i-r_j)+\alpha_{RK}}^{\infty} \text{sech}^2(y/2)dy$, and Eqs. 6 and 7 follow for $\nu_{RK} = e^{\alpha_{RK}}$.

Davidson (1970) starts from Luce's "choice axiom" (Luce, 1959a) which states that a complete system of choice probabilities should satisfy $p(y_i \succ y_j)/p(y_j \succ y_i) = \lambda_i/\lambda_j$, which the Rao-Kupper model fails to do. Davidson (1970) observes that it is desirable for the probability of a tie to "be proportional to the geometric mean of the probabilities of preference". Adding this requirement $p(y_i \sim y_j) \propto \sqrt{p(y_i \succ y_y)p(y_j \succ y_i)}$ to the choice axioms yields Eqs. 8 and 9 as a preference model that allows for ties and also satisfies the choice axiom.

The Rao-Kupper win and tie probabilities can be written in a form more useful for DPO (Appendix B.1), with $\nu_{RK} = e^{\alpha_{RK}}$, as

$$p_\theta^{RK}(y_w \succ_x y_l) = \sigma(d_\theta(x, y_w, y_l) - \alpha_{RK}) \tag{10}$$

$$\begin{aligned} p_\theta^{RK}(y_w \sim_x y_l) &= (\nu_{RK}^2 - 1)\,\sigma(-d_\theta(x, y_w, y_l) - \alpha_{RK})\,\sigma(d_\theta(x, y_w, y_l) - \alpha_{RK}) \\ &= (\nu_{RK}^2 - 1)\,\sigma(-d_\theta(x, y_w, y_l) - \alpha_{RK})\,p_\theta^{RK}(y_w \succ_x y_l) \end{aligned} \tag{11}$$

and the Davidson win and tie probabilities can be written as

$$p_\theta^{D}(y_w \succ_x y_l) = \frac{1}{1 + e^{-d_\theta(x,y_w,y_l)} + 2\nu_D e^{-d_\theta(x,y_w,y_l)/2}} \tag{12}$$

$$p_\theta^{D}(y_w \sim_x y_l) = 2\,\nu_D\, e^{-d_\theta(x,y_w,y_l)/2}\, p_\theta^{D}(y_w \succ_x y_l) \tag{13}$$

Although their parametric forms are different, their treatments of wins and ties are similar (Appendix B.1, Fig. 5). For pairs (x, y_w, y_l) treated as wins, higher likelihood is assigned for higher values of the reward margin $d_\theta(x, y_w, y_l)$. For the Rao-Kupper this is particularly clear, in that the Bradley-Terry preference distribution is simply shifted by α_{RK}. Conversely, for pairs (x, y_w, y_l) treated as ties, the probability of declaring a tie is high for small reward margins $d_\theta(x, y_w, y_l)$.

Balancing Wins and Ties. In the special case of two evenly matched players ($\lambda_i = \lambda_j$), we are interested in the probability of a tie $p(y_i \sim y_j)$ versus a clear win by either player, $p(y_i \succ y_j) + p(y_j \succ y_i)$. It follows that $P_{RK}(\text{tie}) = \frac{\nu_{RK}-1}{2} P_{RK}(\text{no tie})$ and $P_D(\text{tie}) = \nu_D P_D(\text{no tie})$. This shows that the parameters ν determine the probability that equally-matched items are judged as tied or not. ν can be tuned, but in our work, we assume that equally-matched items will tie with a probability of 1/2 and so we set $\nu_{RK} = 3$ and $\nu_D = 1$.

2.3 Incorporating Rao-Kupper and Davidson Models into DPO

We extend the DPO policy objective (Eq. 4) to include a binary flag t to indicate a tie:

$$\mathcal{L}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{x,y_w,y_l,t=0} \log p_\theta(y_w \succ_x y_l) - \mathbb{E}_{x,y_w,y_l,t=1} \log p_\theta(y_w \sim_x y_l) \tag{14}$$

where $p_\theta(y_w \succ y_l)$ and $p_\theta(y_w \sim y_l)$ are taken from either the Rao-Kupper model (Eqs. 10, 11 or the Davidson model (Eqs. 12, 13). Note that in Eq. 14 preference pairs in the dataset are unambiguously either wins ($t = 0$) or ties ($t = 1$). The policy objectives for these two DPO variants are:

$$\begin{aligned} \mathcal{L}_{RK}(\pi_\theta; \pi_{ref}) = &-\mathbb{E}_{x,y_w,y_l,t=0}\Big[\log \sigma(d_\theta(x, y_w, y_l) - \alpha_{RK})\Big] \\ &- \mathbb{E}_{x,y_w,y_l,t=1}\Big[\log \sigma(-d_\theta(x, y_w, y_l) - \alpha_{RK}) + \log \sigma(d_\theta(x, y_w, y_l) - \alpha_{RK}) - \log(\nu_{RK}^2 - 1)\Big] \end{aligned} \tag{15}$$

and

$$\mathcal{L}_D(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{x,y_w,y_l,t=0}\left[\log \frac{1}{1 + e^{-d_\theta(x,y_w,y_l)} + 2\nu_D e^{-d_\theta(x,y_w,y_l)/2}}\right]$$
$$- \mathbb{E}_{x,y_w,y_l,t=1}\left[\log \frac{2\nu_D e^{-d_\theta(x,y_w,y_l)/2}}{1 + e^{-d_\theta(x,y_w,y_l)} + 2\nu_D e^{-d_\theta(x,y_w,y_l)/2}}\right] \quad (16)$$

We refer to these DPO variants as DPO-RK and DPO-D. Like DPO, these objectives depend on the policy π_θ through the reward margin $d_\theta(x, y_w, y_l)$ (Eq. 2). Unlike DPO, the training objective Eq. 14 consists of two competing terms. For pairs (x, y_w, y_l) labeled as wins the objective is to find π_θ to increase the reward margin $d_\theta(x, y_w, y_l)$. However, for pairs labeled as ties the objective is to find π_θ to minimize $|d_\theta(x, y_w, y_l)|$. To simultaneously achieve both these objectives, the underlying policy should learn to model both wins and ties.

2.3.1 DPO-RK and DPO-D Updates

Rafailov et al. (2023) show that DPO dynamically adjusts the gradient according to how well the preference objective is optimized for each sample

$$\nabla_\theta \log p_\theta^{BT}(y_w \succ_x y_l) = \underbrace{\sigma(-d_\theta(x, y_w, y_l))}_{\substack{\text{higher weight when reward} \\ \text{estimate is wrong}}} \beta \nabla_\theta \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} \quad (17)$$

DPO-RK and DPO-D also adjust their gradients dynamically (Appendix B.2). We define the gradient scale factors Δ_{win} and Δ_{tie} to illustrate the DPO-RK and DPO-D gradient updates on wins and ties:

$$\nabla \log p_\theta^{RK}(y_w \succ_x y_l) = \underbrace{\sigma(\alpha - d_\theta(x, y_w, y_l))}_{\Delta_{win}^{RK}(d_\theta)} \beta \nabla_\theta \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} \quad (18)$$

$$\nabla_\theta \log p_\theta^{RK}(y_w \sim_x y_l) = \underbrace{\left[\sigma(\alpha - d_\theta(x, y_w, y_l)) - \sigma(\alpha + d_\theta(x, y_w, y_l))\right]}_{\Delta_{tie}^{RK}(d_\theta)} \beta \nabla_\theta \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} \quad (19)$$

$$\nabla_\theta \log p_\theta^{D}(y_w \succ_x y_l) = \underbrace{\frac{e^{-d_\theta} + \nu e^{-d_\theta/2}}{1 + e^{-d_\theta} + 2\nu e^{-d_\theta/2}}}_{\Delta_{win}^{D}(d_\theta)} \beta \nabla_\theta \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} \quad (20)$$

$$\nabla_\theta \log p_\theta^{D}(y_w \sim_x y_l) = \underbrace{\left[\Delta_{win}^{D}(d_\theta) - \frac{1}{2}\right]}_{\Delta_{tie}^{D}(d_\theta)} \beta \nabla_\theta \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} \quad (21)$$

$\nabla \log p_\theta(y_w \succ_x y_l)$: For data labeled as wins, the DPO-RK gradient scale factor has the same form as DPO, but shifted by α_{RK} (Fig. 6). DPO-D has a symmetric scale factor that is not as steep as DPO-RK. All three methods work to increase the reward margin $d_\theta(x, y_w, y_l)$.

$\nabla \log p_\theta(y_w \sim_x y_l)$: For data labeled as ties, the DPO-D and DPO-RK gradient scale factors are odd and work to drive $d_\theta(x, y_w, y_l)$ towards zero, although the DPO-RK scale factor is more aggressive. This is a mechanism not present in DPO.

2.3.2 Rao-Kupper and Davidson Classifiers

The above DPO variants yield probability distributions $p_\theta(y_w \succ_x y_l)$ and $p_\theta(y_w \sim_x y_l)$ in terms of the policy π_θ and the reference model π_{ref}. We can use these distributions as classifiers to label a pair (x, y_1, y_2) as either a win ($y_1 \succ_x y_2$ or $y_2 \succ_x y_1$) or a tie ($y_1 \sim_x y_2$), whichever has the highest probability under either the Rao-Kupper or the Davidson model (Eqs. 10, 11, or 12, 13). We will evaluate classification performance on held-out data not used in training to see if policies produced by our DPO variants learn to distinguish wins from ties.

3 Experiments

3.1 Adding Ties to DPO

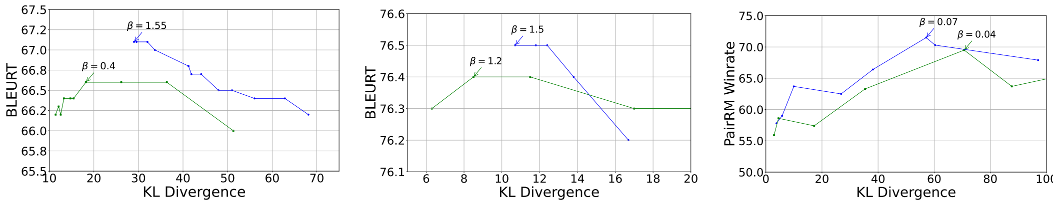
DPO in its original formulation relies on a static dataset of comparisons $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$ where $y_w^{(i)}$ and $y_l^{(i)}$ are preferred and dispreferred responses to a prompt $x^{(i)}$ (Rafailov et al., 2023). These preferences are assumed to be sampled from some latent reward model and we refer to this dataset as **Clear Preference Pairs** (**CPs**, for short) because they are typically selected to reflect a clear preference between winner and loser as assessed either by human judges or by some trusted automatic metric. We distinguish these Clear Preference Pairs from **Tied Pairs (TPs)**. Tied Pairs also consist of a winner and a loser, but are very similar in quality. Human judges might be less consistent, or have less confidence, in selecting the winner in a tied pair, and automatic metrics will assign more similar quality scores to Tied Pairs than to Clear Preference Pairs. As noted, DPO datasets typically are constructed to include only Clear Preference Pairs. We will extend the data selection procedures to generate Tied Pairs along with Clear Preference Pairs so that we can investigate how DPO changes when Tied Pairs are included in the training data. We report experiments on Neural Machine Translation (NMT) and Summarization. Appendix C gives experiment details.

Clear Preference Pairs and Tied Pairs in NMT. We use DPO to improve translation quality similar to that done in Yang et al. (2024b). We apply DPO with BLOOMZ-mt-7b (Muennighoff et al., 2023) as the baseline model. Translation quality is measured with BLEURT (Sellam et al., 2020) on the WMT21 ZH-EN and IWSLT17 FR-EN translation test sets (Appendix C.1). To construct a DPO preference dataset for the WMT21 ZH-EN test set, we use BLOOMZ-mt-7b to generate 32 translations (via sampling) for each source sentence in the WMT20 ZH-EN test set. For each source sentence, the translations are ranked by their BLEURT scores computed with respect to the reference translations. The highest and lowest scoring translations form the Clear Preference Pairs; for each source sentence, these are the two translations with the greatest difference in BLEURT score. By contrast, we take the Tied Pairs as the two non-identical translations with the minimum absolute BLEURT difference; the translation with higher BLEURT is labeled as the winner of each Tied Pair. This yields ca. 16K CPs and TPs for use in DPO. The same procedure is applied to the IWSLT17 validation set, yielding ca. 800 CPs and TPs for use as DPO preference datasets.

Clear Preference Pairs and Tied Pairs in Summarization. We follow Amini et al. (2024a) in DPO fine-tuning of Pythia-2.8B (Biderman et al., 2023) on the TL;DR dataset (Stiennon et al., 2020) with evaluation via win-rate against human-written summaries. Previous works use GPT-4 to compute the win-rate (Rafailov et al., 2023; Amini et al., 2024b). We find that the judgments of PairRM (Jiang et al., 2023) agree well with those of GPT-4 (Appendix C.3) and opt to use PairRM win-rate as a cost-effective automatic metric. In the TL;DR task, each prompt is associated with a collection of paired summaries, with a winner and a loser identified for each pair. There is no immediately obvious way to distinguish tied pairs from clear preference pairs in the collection and so we use DPO itself to select tied pairs. We first apply DPO with $\beta = 0.1$ on the full TL;DR training dataset. Using the reward model formed by this model and the reference model, we compute the reward margins of all pairs of summaries in the training split. For each prompt, the pair with minimal reward margin is treated as a tied pair, with all other pairs kept as clear preference pairs, yielding ca. 14k (15.3%) TPs. See Appendix C.4 for a study of this selection strategy.

3.1.1 Task Performance vs. KL to the Reference Policy

Following prior work (Rafailov et al., 2023; Amini et al., 2024b; Park et al., 2024), we evaluate DPO in terms of task performance versus KL divergence to the reference policy. For each of the three tasks we form two training sets: CP, which contains the Clear Preference Pairs; and CP+TP, which contains both the Clear Preference Pairs and the Tied Pairs. We refer to DPO training on these sets as DPO(CP) and DPO(CP+TP) (Figure 1).



(a) WMT21 ZH-EN. Baseline = 66.0

(b) IWSLT17 FR-EN. Baseline = 75.7

(c) TL;DR PairRM Win-Rate; Baseline = 37.5%

Figure 1: Task Performance vs. KL to the reference policy for DPO systems trained on Clear Preference Pairs (DPO(CP), blue) and on Clear Preference Pairs and Tied Pairs (DPO(CP+TP), green). KL is estimated over 256 test set policy samples; β is noted for best performing systems. Full details are in Appendix C.5.1.

The obvious conclusion from these experiments is that including tied pairs in DPO is not good for task performance. All best performing systems are obtained by DPO(CP), with DPO(CP+TP) underperforming for nearly all values of KL relative to the reference policy. This performance degradation from including ties is consistent with common practice in the DPO literature which only keeps pairs with clear preference, filtering others to obtain the best-performing system (Yang et al., 2024a; Dubey et al., 2024). Consistent with this, the TL;DR results show that removing tied pairs from the DPO dataset leads to improved summarization performance, even when ties are identified by a DPO model in an unsupervised manner. These results also suggest that tied pairs in the DPO datasets can enhance regularization. By this we mean that including tied pairs causes DPO to find models that are closer to the reference policy as measured by KL divergence. The overall effect of the reduced task performance and more regularization is to shift the frontier 'down and to the left'.

Theorem 3.1 of Chen et al. (2024) suggests how these regularization effects might arise. The ideal DPO policy π^* should follow (Appendix D):

$$\frac{\pi^*(y_w|x)}{\pi^*(y_l|x)} = \frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\text{ref}}(y_l|x)} \Big(\frac{\gamma(x, y_w, y_l)}{1 - \gamma(x, y_w, y_l)} \Big)^{1/\beta} \tag{22}$$

where $\gamma(x, y_w, y_l)$ is the true preference probability of $y_w \succ y_l$ under prompt x. If we assume that tied pairs have a true preference probability $\gamma(x, y_w, y_l)$ of 0.5, from Equation 22 we have $\frac{\pi^*(y_w|x)}{\pi^*(y_l|x)} = \frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\text{ref}}(y_l|x)}$, where π^* is the ideal DPO policy[1]. By this analysis, the ideal DPO model should maintain the same chosen/rejected likelihood ratio as the reference model on tied pairs, and this constraint serves as a form of regularization. In our NMT experiments (Figures 8a, 8b), where half of the pairs are constructed to be ties, the regularization effect is especially pronounced as the DPO model should keep to the reference model likelihood ratio on 50% of the training data. Regularization is less pronounced on TL;DR (Figure 1c) where only 1/8 of the pairs are ties. Furthermore, Eq 22 can be rearranged as follows:

$$d^*_\theta(x, y_w, y_l) = \beta \Big(\log \frac{\pi^*(y_w|x)}{\pi_{ref}(y_w|x)} - \log \frac{\pi^*(y_l|x)}{\pi_{ref}(y_l|x)} \Big) = \beta \log \frac{\gamma(x, y_w, y_l)}{1 - \gamma(x, y_w, y_l)} \tag{23}$$

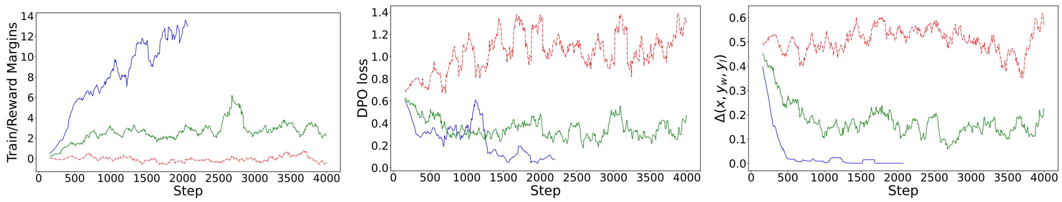
From this it follows that the reward margin on tied pairs should ideally be close to zero, which we verify experimentally in the next section.

3.1.2 Convergence Behaviour

We analyse how the inclusion of tied pairs affects the detailed behaviour of DPO. We study DPO on the BLOOMZ-mt-7b datasets with $\beta = 0.7$ for WMT21 ZH-EN as these systems show both strong regularization effects and task performance degradation when tied pairs

[1]In Appendix D, we show that the ideal policy can also be derived for DPO-D which includes the ideal DPO policy as a special case.

are added. Figure 2 shows the evolution of reward margins, DPO loss, and gradient scale factors (Equations 2, 5, 24) during training.



(a) Reward margins (b) DPO loss (c) Gradient scale factor

Figure 2: DPO(CP) (blue) and DPO(CP+TP) training statistics on WMT21 ZH-EN. For DPO(CP+TP), margins, loss, and gradient scale factor are shown separately on CPs (green) and on TPs (red).

DPO(CP) is well behaved: the reward margins on the CP data increase over the epoch (Fig. 2a (blue)); the DPO losses on the CP dataset decrease over the epoch (Fig. 2b (blue)); and the DPO gradient scale factor shows that learning slows and stabilizes after the 500^{th} batch (Fig. 2c (blue)).

Adding tied pairs to the DPO dataset alters this behaviour for both tied pairs and clear preference pairs. DPO(CP+TP) does yield some gains in reward margins for clear preference pairs, but these are well below that of DPO(CP) (Fig. 2a (blue vs green)). By contrast, DPO(CP+TP) fails almost entirely to find any improvement in the reward margins for its tied pair data (Fig. 2a (red)). While this is less than ideal from a modeling perspective, we note that it provides empirical support for the observation in the previous section that the reward margins on tied pairs should ideally remain close to zero. Similar behaviour is observed in the DPO loss (Fig. 2b). Decreases in loss over clear preference pairs are offset by loss increases on the tied pairs. This is reflected in the gradient scale factors. The gradient scale factors remain high as DPO(CP+TP) searches for a better policy.

3.2 Adding Ties to DPO-RK and DPO-D

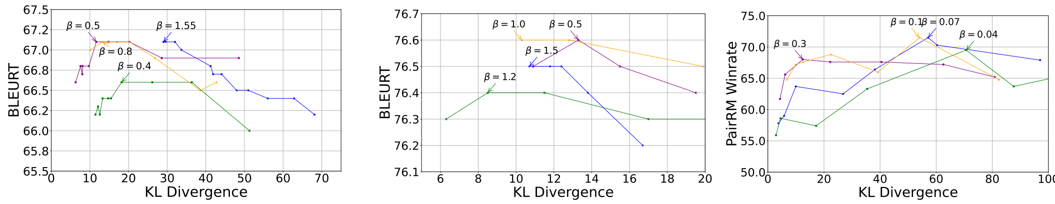
In the previous section we investigated the effects of including tied preference pairs in DPO datasets. Using the same data we now evaluate DPO-RK and DPO-D as DPO variants that explicitly model both ties and clear preferences. We use the DPO datasets CP+TP (Sec. 2.2) with the DPO-D and DPO-RK algorithms to produce models DPO-D(CP+TP) and DPO-RK(CP+TP). We follow the protocols of Sec. 3.1 so that results are directly comparable to earlier DPO(CP) and DPO(CP+TP) results. For all experiments we set $\nu^{RK} = 3$ and $\nu^{D} = 1$ for DPO-RK and DPO-D (as described in Sec. 2.2).

3.2.1 Task Performance vs. KL to the Reference Policy

When tied pairs are added to the dataset, DPO-D and DPO-RK do not suffer the same drops in task performance that DPO exhibits (Fig. 3, orange and purple vs. green). DPO-RK(CP+TP) and DPO-D(CP+TP) reach similar levels of task performance to each other, and to DPO(CP), but do so at smaller KL values than DPO (Fig. 3, orange and purple vs. blue). These are the regularization effects of including tie pairs in the DPO datasets reported in Section 3.1, but without decrease in task performance. For a given level of KL to reference policy, DPO-D(CP+TP) and DPO-RK(DP+TP) yield higher task performance than DPO(CP). Compared to DPO as it is usually done, DPO-RK and DPO-D frontiers are shifted leftwards, showing similar task performance but stronger regularization.

3.2.2 Preference Pair Classification Accuracy

We assess the performance of the Rao Kupper and Davidson classifiers introduced in Sec.2.3.2 in terms of their ability to label preference pairs as either clear preferences or ties. Ideally, classification performance will improve: (1) as tied pairs are added to the clear preference



(a) WMT21 ZH-EN frontiers. Baseline BLEURT = 66.0

(b) IWSLT FR-EN frontiers. Baseline BLEURT = 75.7

(c) TL;DR frontiers. Baseline win-rate = 37.5%

Figure 3: KL-Performance frontiers with DPO(CP) in blue, DPO(CP+TP) in green, DPO-RK(CP+TP) in purple, and DPO-D(CP+TP) in orange. Full details in Appendix C.5.

data sets (CP vs CP+TP); and (2) with margins generated from models produced by DPO variants that emphasize the distinction between tied pairs and clear preference pairs (DPO-D(CP+TP), DPO-RK(CP+TP)).

We assess classifier performance on the held-out set created by collecting CPs and TPs from the WMT18 ZH-EN test set as was done for WMT20 ZH-EN (Sec.3.1); this yields pairs with gold labels as either clear preference pairs or tied pairs. Classification and assessment proceeds as follows: we generate reward margins for the WMT18 ZH-EN pairs using DPO(CP), DPO(CP+TP), DPO-RK(CP+TP), DPO-D(CP+TP) models; we use these reward margins to label the unseen pairs using the Davidson and Rao-Kupper classification rules (Sec. 2.3.2); and finally compute the classification accuracy relative to the gold labels.

Results are shown in Table 1. We find that smaller beta in training consistently leads to better overall RK-classification accuracy (+10% overall Acc. from $\beta = 1.0$ to $\beta = 0.1$), suggesting heavy regularization with respect to the reference model impedes preference ranking. Classifiers based on reward margins generated from DPO(CP) models perform well in identifying clear preference pairs (Acc. > 85%) but poorly in identifying tied pairs (Acc. < 35%). This imbalance is likely explained by the DPO(CP) model never having seen tied pairs in training. Adding TPs to the DPO datasets (DPO(CP+TP)) significantly improves the classification accuracy of tied pairs (+30%) with more balanced classification accuracies for CPs and TPs. The best overall classification accuracies ($\approx 73\%$) are obtained with reward margins generated by models trained to match its classifier. Across all beta values, DPO-RK(CP+TP) and DPO-D(CP+TP) achieve better overall accuracy and more-balanced CP accuracy and TP accuracy under their respective decision rules.

Model	$\beta = 0.1$	$\beta = 0.5$	$\beta = 1.0$
		Rao-Kupper Classifier	
DPO(CP)	60.1% (**87.1%**, 33.1%)	52.8% (87.3%, 18.3%)	50.1% (86.9%, 13.3%)
DPO(CP+TP)	67.0% (72.0%, 62.1%)	57.5% (69.3%, 45.7%)	51.5% (71.2%, 31.9%)
DPO-RK(CP+TP)	**73.1**% (74.5%, **71.7%**)	64.2% (73.2%, 55.3%)	58.5% (73.4%, 43.5%)
		Davidson Classifer	
DPO(CP)	65.3% (**84.4%**, 46.3%)	57.4% (83.7%, 31.0%)	53.6% (84.6%, 22.6%)
DPO(CP+TP)	71.0% (59.1%, **82.8%**)	62.1% (58.3%, 65.8%)	57.2% (62.3%, 52.2%)
DPO-D(CP+TP)	**73.8%** (79.6%, 67.9%)	66.8% (75.9%, 57.8%)	62.7% (75.2%, 50.3%)

Table 1: Preference pair classification accuracies (Overall Acc. (CP Acc., TP Acc.)) for Rao-Kupper and Davidson classification rules based on reward margins computed using DPO(CP), DPO(CP+TP), DPO-RK(CP+TP), and DPO-D(CP+TP) models as evaluated on the WMT18 ZH-EN test set.

3.2.3 Empirical Reward Margin Distributions

We now look at the reward margins on held-out pairs to determine how the DPO objective generalizes to unseen data. Ideally, a post-DPO model should assign reward margins that are large for clear preference pairs but close to zero for tied pairs. We assess this on the same held-out data as in the previous section (Sec. 3.1).

Model	$\beta = 0.1$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.1$	$\beta = 0.5$	$\beta = 1.0$
	Clear Preference Pairs			Tied Pairs		
DPO(CP)	8.2 ±12.0	9.5 ±13.2	10.0 ±11.1	0.7 ±13.2	0.6 ±9.4	0.4 ±7.9
DPO(CP+TP)	2.4 ±3.3	2.3 ±3.2	2.5 ±3.3	0.4 ±4.8	0.3 ±3.2	0.2 ±2.7
DPO-RK(CP+TP)	2.9 ±4.3	2.8 ±3.3	3.0 ±3.3	0.0 ±1.3	0.0 ±1.4	0.0 ±1.7
DPO-D(CP+TP)	4.6 ±5.8	4.8 ±6.1	4.9 ±6.3	0.0 ±2.0	0.1 ±2.3	0.0 ±2.4

Table 2: Reward margin statistics (mean ± std) for Clear Preference Pairs and Tied Pairs from WMT18 ZH-EN.

In Table 2, reward margins of DPO(CP+TP), DPO-RK(CP+TP), and DPO-D(CP+TP) are similar and well-behaved, showing means close-to-zero on TPs (< 0.4) and farther from zero for CPs (> 2.3). Reward margin standard deviations are also similar and reasonably small. However the standard deviation for both tied pairs and clear preference pairs are much higher for DPO(CP) models (≥ 11.1 on CPs and ≥ 7.9 on TPs).

This can be explained by Figure 4 which shows that DPO(CP) models overwhelmingly assign preference probability values of either ~ 1.0 or ~ 0.0 to tied pairs, corresponding to very positive and very negative reward margins, respectively. This contributes to the high standard deviation and shows that for a tied pair (y_1, y_2), DPO(CP) model exhibits a strong preference for either $y_1 \succ y_2$ or $y_2 \succ y_1$, even though these are tied pairs by construction ($y_1 \sim y_2$). In contrast, DPO(CP+TP) yields well-behaved estimated preference probability distribution more centered around 0.5 for tied pairs.

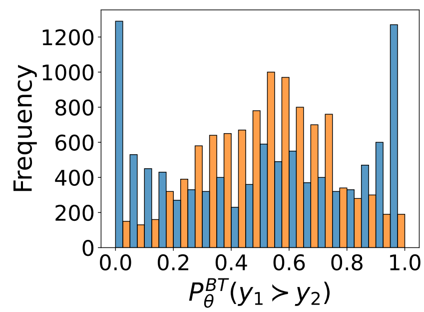


Figure 4: Empirical distribution of tied probabilities on tied pairs. DPO(CP) in blue, and DPO(CP+TP) in orange. See Appendix C.6 for an analysis of CPs.

4 Related Work

Variants of Direct Preference Optimization A range of variants of Direct Preference Optimization have been proposed based on problem-specific or theoretical motivations. Park et al. (2024) tackle excessive response length by introducing explicit length normalization in the DPO objective. SimPO (Meng et al., 2024) modifies the DPO objective to remove the need for a reference model and to include length normalization. KTO (Ethayarajh et al., 2024) is motivated by Kahneman and Tversky's prospect theory and learns from non-paired preference data. ODPO (Amini et al., 2024a) incorporates preference strength in the objective by introducing an offset parameter. In deriving ODPO, the offset parameter of Amini et al. (2024b, Theorem 3)) plays a role similar to the sensitivity threshold of Rao and Kupper (1967). To our knowledge, our work is the first to consider accommodating tied pairs in DPO. We note that the ODPO objective with a fixed offset agrees with our proposed DPO-RK objective restricted to clear preference data, but does not extend to ties.

Frameworks for Pair-wise Preference Optimization Several works propose theoretical frameworks for understanding general Preference Optimization from which DPO can be obtained as a special case. Azar et al. (2024) introduces the ΨPO formalism which allows alternative expression of the reward in terms of the model's predicted probability. IPO is derived when the identity mapping is used, and DPO arises under a log-ratio mapping. Dumoulin et al. (2024) formulate learning from pair-wise preference as learning the implicit preference generating distribution of the annotators. In this formalism, DPO is a well-specified model for the implicit preference distribution assuming that the human preference generative process follows the Bradley-Terry model. Our work can be viewed as assuming an annotator preference generating distribution that allows for the outcome of a tie (i.e. the Rao-Kupper or the Davidson model). Tang et al. (2024) propose a generalized approach to deriving offline preference optimization losses through binary classification. In this work, we consider the ternary classification with the possibility of declaring a tie. In Appendix D, we show that the 'perfect' DPO-D policy can be simulated starting from the ternary classification loss.

Pair-wise Comparison Models Hamilton et al. (2023) review the history and the range of motivations for the Bradley-Terry model, including its relation to the logistic distribution (Bradley and Gart, 1962), and the Luce choice axiom Luce (1959b). The Rao-Kupper (Rao and Kupper, 1967) and the Davidson model (David, 1988) are two notable extensions to Bradley-Terry (Sec. 2.2). We point interested readers to a review by David (1988) and a bibliography by Davidson and Farquhar (1976). Modeling ties remains an active research topic in fields such as sport team ranking (Zhou et al., 2022) and medical treatments (Gaohong Dong and Vandemeulebroecke, 2020).

5 Conclusion

We have derived and investigated two tie-compatible DPO variants, DPO-RK and DPO-D, by replacing the Bradley-Terry preference model with the Rao-Kupper model and the Davidson model, respectively. Our experiments on translation and summarization show that by explicitly modeling the probability of declaring a tie, DPO-RK and DPO-D can accommodate tied pairs in preference data without the degradation in task performance that is observed when the same tied pairs are added to the original DPO. We find empirically that the inclusion of ties in preference learning leads to stronger regularization with respect to the reference model as measured by KL divergence, gives better-behaved reward margin distribution on held-out sets and improves the trained policy's overall accuracy in classifying clear preference and tied pairs. These findings alongside with the proposed DPO variants motivate and enable the use of tied pairs in available preference data as opposed to wastefully discarding them. We discuss limitations in Appendix A.

References

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. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-

Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 Technical Report, 2024a. URL https://arxiv.org/abs/2407.10671.

PV Rao and Lawrence L Kupper. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317):

194–204, 1967.

Roger R. Davidson. On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments. *Journal of the American Statistical Association*, 65(329): 317–328, 1970. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2283595.

Ralph Allan Bradley. Some Statistical Methods in Taste Testing and Quality Evaluation. *Biometrics*, 9(1):22–38, 1953. ISSN 0006-341X. doi: 10.2307/3001630. URL https://www.jstor.org/stable/3001630. Publisher: [Wiley, International Biometric Society].

R Duncan Luce. *Individual choice behavior*, volume 4. Wiley New York, 1959a.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. Direct Preference Optimization for Neural Machine Translation with Minimum Bayes Risk Decoding. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 391–398, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.34. URL https://aclanthology.org/2024.naacl-short.34.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual Generalization through Multitask Finetuning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.891. URL https://doi.org/10.18653/v1/2023.acl-long.891.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: Learning Robust Metrics for Text Generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.704. URL https://doi.org/10.18653/v1/2020.acl-main.704.

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct Preference Optimization with an Offset. *CoRR*, abs/2402.10571, 2024a. doi: 10.48550/ARXIV.2402.10571. URL https://doi.org/10.48550/arXiv.2402.10571.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. *CoRR*, abs/2009.01325, 2020. URL https://arxiv.org/abs/2009.01325.

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct Preference Optimization with an Offset. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 9954–9972. Association for Computational Linguistics, 2024b. URL https://aclanthology.org/2024.findings-acl.592.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling Large Language Models with Pairwise Comparison and Generative Fusion. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, 2023.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling Length from Quality in Direct Preference Optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 4998–5017. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-acl.297.

Angelica Chen, Sadhika Malladi, Lily H. Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. Preference Learning Algorithms Do Not Learn Preference Rankings. *CoRR*, abs/2405.19534, 2024. doi: 10.48550/ARXIV.2405.19534. URL https://doi.org/10.48550/arXiv.2405.19534.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple Preference Optimization with a Reference-Free Reward. *CoRR*, abs/2405.14734, 2024. doi: 10.48550/ARXIV.2405.14734. URL https://doi.org/10.48550/arXiv.2405.14734.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model Alignment as Prospect Theoretic Optimization. *CoRR*, abs/2402.01306, 2024. doi: 10.48550/ARXIV.2402.01306. URL https://doi.org/10.48550/arXiv.2402.01306.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Rémi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A General Theoretical Paradigm to Understand Learning from Human Preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pages 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Vincent Dumoulin, Daniel D. Johnson, Pablo Samuel Castro, Hugo Larochelle, and Yann N. Dauphin. A density estimation perspective on learning from pairwise human preferences. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=YH3oERVYjF.

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. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gu3nacA9AH.

Ian Hamilton, Nick Tawn, and David Firth. The many routes to the ubiquitous Bradley-Terry model, 2023. URL https://arxiv.org/abs/2312.13619.

Ralph A. Bradley and John J. Gart. The Asymptotic Properties of ML Estimators when Sampling from Associated Populations. *Biometrika*, 49(1/2):205–214, 1962. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2333482.

R Duncan Luce. *Individual choice behavior*, volume 4. Wiley New York, 1959b.

H. A. David. *The Method of Paired Comparisons*. Number No. 41 in Griffin's Statistical Monographs and Courses. Charles Griffin and Company Ltd., London, 2nd edition, 1988.

Roger R. Davidson and Peter H. Farquhar. A Bibliography on the Method of Paired Comparisons. *Biometrics*, 32(2):241–252, 1976. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529495.

Yuhao Zhou, Ruijie Wang, Yi-Cheng Zhang, An Zeng, and Matúš Medo. Improving Pagerank using sports results modeling. *Knowledge-Based Systems*, 241:108168, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.108168. URL https://www.sciencedirect.com/science/article/pii/S0950705122000314.

Junshan Qiu Roland A. Matsouaka Yu-Wei Chang Jiuzhou Wang Gaohong Dong, David C. Hoaglin and Marc Vandemeulebroecke. The Win Ratio: On Interpretation and Handling of Ties. 12(1):99–106, 2020. doi: 10.1080/19466315.2019.1575279. URL https://doi.org/10.1080/19466315.2019.1575279.

Hoang Tran, Chris Glaze, and Braden Hancock. Iterative DPO Alignment. Technical report, Snorkel AI, 2023.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.213. URL https://doi.org/10.18653/v1/2020.emnlp-main.213.

Matt Post. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319.

A Limitations

The effect of accommodating ties in preference learning can be further investigated using human-annotated tied pairs. However, at the time of writing, there is no substantial preference dataset with annotated ties; notably, current annotation guidelines are typically written to explicitly exclude ties. We note that this enforcement of win/lose judgments has likely conditioned the generative process of human preference towards the Bradley-Terry model. A meaningful extension of this work would be to assess the effectiveness of DPO-RK and DPO-D on preference datasets where the annotators are asked to identify ties. As explained in Sec 2.2, the hyper-parameter ν_{RK} and ν_D can be tuned which would require either grid search or estimation given ground-truth preference/tie probabilities. We find that the choice of $\nu_{RK} = 3$ and $\nu_D = 1$ as motivated in Sec 2.2 works well and we did not need to tune the parameter to obtain good performance. It is likely that better performance and more efficient frontiers can be obtained by tuning ν to better fit the underlying preference generative process for both DPO-RK and DPO-D. Given our focus on accommodating ties from a modeling perspective, we leave performance optimization to future works concerning applications.

B Mathematical Derivations

B.1 Rao-Kupper and Davidson Preference and Tie Probabilities

We derive the win and tie probabilities as functions of the reward margin $d_\theta(x, y_w, y_l) = r_\theta(x, y_w) - r_\theta(x, y_l)$ (Eq 2) under the Rao-Kupper (Eq 10, 11) and Davidson formulations (Eq 12, 13).

The Rao-Kupper win and tie probabilities can be obtained by substituting $\lambda_w = e^{r_\theta(x,y_w)}$, $\lambda_l = e^{r_\theta(x,y_l)}$ and $\nu_{RK} = e^{\alpha_{RK}}$ into Eq 6 and Eq 7, respectively:

$$
\begin{aligned}
p_\theta^{RK}(y_w \succ y_l) &= \frac{\lambda_w}{\lambda_w + \nu_{RK}\lambda_l} = \frac{e^{r_\theta(x,y_w)}}{e^{r_\theta(x,y_w)} + \nu_{RK} e^{r_\theta(x,y_l)}} \\
&= \frac{1}{1 + e^{r_\theta(x,y_l) - r_\theta(x,y_w) + \alpha_{RK}}} = \sigma(d_\theta(x, y_w, y_l) - \alpha_{RK}) \\
p_\theta^{RK}(y_w \sim y_l) &= \frac{(\nu_{RK}^2 - 1)\lambda_w\lambda_l}{(\lambda_w + \nu_{RK}\lambda_l)(\lambda_l + \nu_{RK}\lambda_w)} = \frac{(\nu_{RK}^2 - 1) e^{r_\theta(x,y_w)} e^{r_\theta(x,y_l)}}{(e^{r_\theta(x,y_w)} + \nu_{RK} e^{r_\theta(x,y_l)})(e^{r_\theta(x,y_l)} + \nu_{RK} e^{r_\theta(x,y_w)})} \\
&= (\nu_{RK}^2 - 1)\Big(\frac{e^{r_\theta(x,y_l)}}{e^{r_\theta(x,y_l)} + \nu_{RK} e^{r_\theta(x,y_w)}}\Big)\Big(\frac{e^{r_\theta(x,y_w)}}{e^{r_\theta(x,y_w)} + \nu_{RK} e^{r_\theta(x,y_l)}}\Big) \\
&= (\nu_{RK}^2 - 1)\, \sigma(-d_\theta(x, y_w, y_l) - \alpha_{RK})\, \sigma(d_\theta(x, y_w, y_l) - \alpha_{RK}) \\
&= (\nu_{RK}^2 - 1)\, \sigma(-d_\theta(x, y_w, y_l) - \alpha_{RK})\, p_\theta^{RK}(y_w \succ y_l)
\end{aligned}
$$

The Davidson win and tie probabilities can be obtained with the same substitution into Eq 8 and Eq 9, respectively:

$$
\begin{aligned}
p_\theta^{D}(y_w \succ_x y_l) &= \frac{\lambda_w}{\lambda_w + \lambda_l + 2\nu_D\sqrt{\lambda_w\lambda_l}} = \frac{e^{r_\theta(x,y_w)}}{e^{r_\theta(x,y_w)} + e^{r_\theta(x,y_l)} + 2\nu_D\sqrt{e^{r_\theta(x,y_w) + r_\theta(x,y_l)}}} \\
&= \frac{1}{1 + e^{-d_\theta(x,y_w,y_l)} + 2\nu_D e^{-d_\theta(x,y_w,y_l)/2}} \\
p_\theta^{D}(y_w \sim_x y_l) &= \frac{2\nu_D\sqrt{\lambda_w\lambda_l}}{\lambda_w + \lambda_l + 2\nu_D\sqrt{\lambda_w\lambda_l}} = (2\nu_D\lambda_w^{-\frac{1}{2}}\lambda_l^{\frac{1}{2}})\frac{\lambda_w}{\lambda_w + \lambda_l + 2\nu_D\sqrt{\lambda_w\lambda_l}} \\
&= 2\,\nu_D\, e^{-\frac{1}{2}(r_\theta(x,y_w) - r_\theta(x,y_l))}\, p_\theta^{D}(y_w \succ_x y_l) \\
&= 2\,\nu_D\, e^{-d_\theta(x,y_w,y_l)/2}\, p_\theta^{D}(y_w \succ_x y_l)
\end{aligned}
$$

In Figure 5 we plot the preference and tie probabilities as a function of reward margin d_θ under Bradley-Terry (as used in DPO), Rao-Kupper (as used in DPO-RK), and Davidson (as used in DPO-D).

B.2 Gradients for DPO-RK and DPO-D

The gradients of the Rao-Kupper log probabilities (Eq 18, 19) are as follows. For convenience, we use the short-hand d_θ for $d_\theta(x, y_w, y_l)$.

$$\begin{aligned}
\nabla \log p_\theta^{RK}(y_w \succ_x y_l) &= \nabla_\theta \log \sigma(d_\theta - \alpha_{RK}) \\
&= \sigma(\alpha_{RK} - d_\theta)\nabla_\theta d_\theta(x, y_w, y_l) \\
&= \underbrace{\sigma(\alpha_{RK} - d_\theta)}_{\Delta_{win}^{RK}(d_\theta)}\Big[\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x)\Big] \\
&= \Delta_{win}^{RK}(d_\theta)\nabla_\theta \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} \\
\nabla_\theta \log p_\theta^{RK}(y_w \sim_x y_l) &= \nabla_\theta \Big[\log \sigma(-d_\theta - \alpha_{RK}) + \log \sigma(d_\theta - \alpha_{RK})\Big] \\
&= -\sigma(d_\theta + \alpha_{RK})\,\nabla_\theta d_\theta + \sigma(-d_\theta + \alpha_{RK})\,\nabla_\theta d_\theta \\
&= \underbrace{\Big(\sigma(\alpha_{RK} - d_\theta) - \sigma(\alpha_{RK} + d_\theta)\Big)}_{\Delta_{tie}^{RK}(d_\theta)}\big[\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x)\big] \\
&= \Delta_{tie}^{RK}(d_\theta)\nabla_\theta \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)}
\end{aligned}$$

The gradients of the Davidson log-probabilities (Eq 20, 21) follow similarly.

$$\begin{aligned}
\nabla_\theta \log p_\theta^{D}(y_w \succ_x y_l) &= \frac{\nabla_\theta p_\theta^{D}(y_w \succ_x y_l)}{p_\theta^{D}(y_w \succ_x y_l)} \\
&= \frac{\nabla_\theta (1 + e^{-d_\theta} + 2\nu e^{-d_\theta/2})^{-1}}{p_\theta^{D}(y_w \succ_x y_l)} \\
&= (-1)\frac{(1 + e^{-d_\theta} + 2\nu e^{-d_\theta/2})^{-2}}{p_\theta^{D}(y_w \succ_x y_l)}(-e^{d_\theta} - \nu e^{d_\theta/2})\nabla_\theta d_\theta \\
&= \frac{p_\theta^{D}(y_w \succ_x y_l)^2}{p_\theta^{D}(y_w \succ_x y_l)}(e^{-d_\theta} + \nu e^{-d_\theta/2})\nabla_\theta d_\theta \\
&= p_\theta^{D}(y_w \succ_x y_l)(e^{-d_\theta} + \nu e^{-d_\theta/2})\nabla_\theta d_\theta \\
&= \underbrace{\frac{e^{-d_\theta} + \nu e^{-d_\theta/2}}{1 + e^{-d_\theta} + 2\nu e^{-d_\theta/2}}}_{\Delta_{win}^{D}(d_\theta)}\Big[\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x)\Big] \\
&= \Delta_{win}^{D}(d_\theta)\nabla_\theta \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} \\
\nabla_\theta \log p_\theta^{D}(y_w \sim_x y_l) &= \nabla_\theta \log\big(2\nu e^{-d_\theta/2} p_\theta^{D}(y_w \succ_x y_l)\big) = \nabla_\theta \Big[\log p_\theta^{D}(y_w \succ_x y_l) - d_\theta/2\Big] \\
&= \Big[\frac{e^{-d_\theta} + \nu e^{-d_\theta/2}}{1 + e^{-d_\theta} + 2\nu e^{-d_\theta/2}} - \frac{1}{2}\Big]\nabla_\theta d_\theta \\
&= \underbrace{\Big[\Delta_{win}^{D}(d_\theta) - \frac{1}{2}\Big]}_{\Delta_{tie}^{D}(d_\theta)}\Big[\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_w|x)\Big] \\
&= \Delta_{tie}^{D}(d_\theta)\nabla_\theta \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)}
\end{aligned}$$

For illustration, we plot Δ_{win} and Δ_{tie} as a function of reward margin d_θ in Figure 6.

The quantities $\nabla_\theta \mathcal{L}_D(\pi_\theta; \pi_{ref})$ and $\nabla_\theta \mathcal{L}_{RK}(\pi_\theta; \pi_{ref})$ follow by substituting the above results into the gradient of Eq 14

$$\nabla_\theta \mathcal{L}(\pi_\theta; \pi_{ref}) = -\nabla_\theta \mathbb{E}_{x,y_w,y_l,t=0} \log p_\theta(y_w \succ_x y_l) - \nabla_\theta \mathbb{E}_{x,y_w,y_l,t=1} \log p_\theta(y_w \sim_x y_l) \quad (24)$$

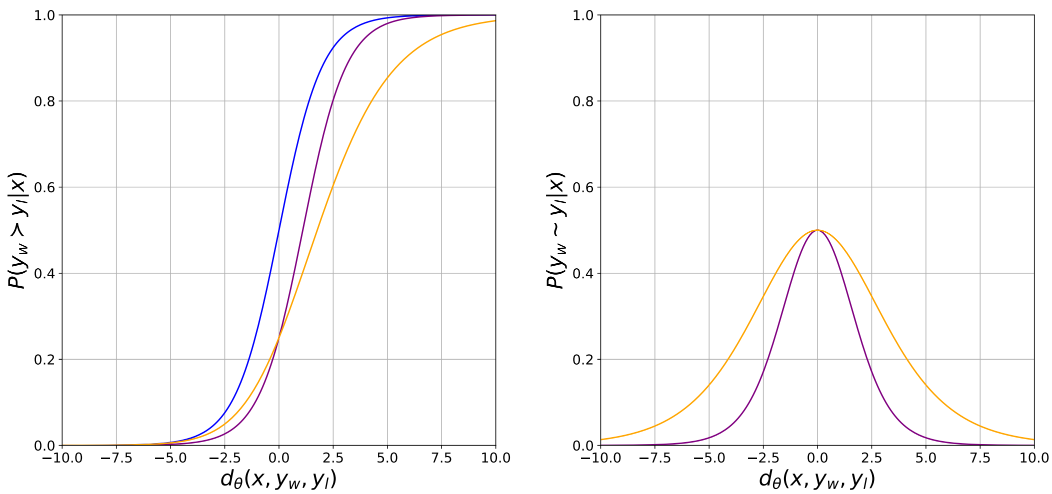


Figure 5: The clear preference probabilities $P(y_w \succ y_l|x)$ (left) and tie probabilities $P(y_w \sim y_l|x)$ (right) as a function of reward margins $d_\theta(x, y_w, y_l)$ for Bradley-Terry (as used in DPO) (blue), Rao-Kupper (purple) (as used in DPO-RK), and Davidson (orange) (as used in DPO-D). $\alpha_{RK} = \log 3$ and $\nu_D = 1$ are used in producing these plots.

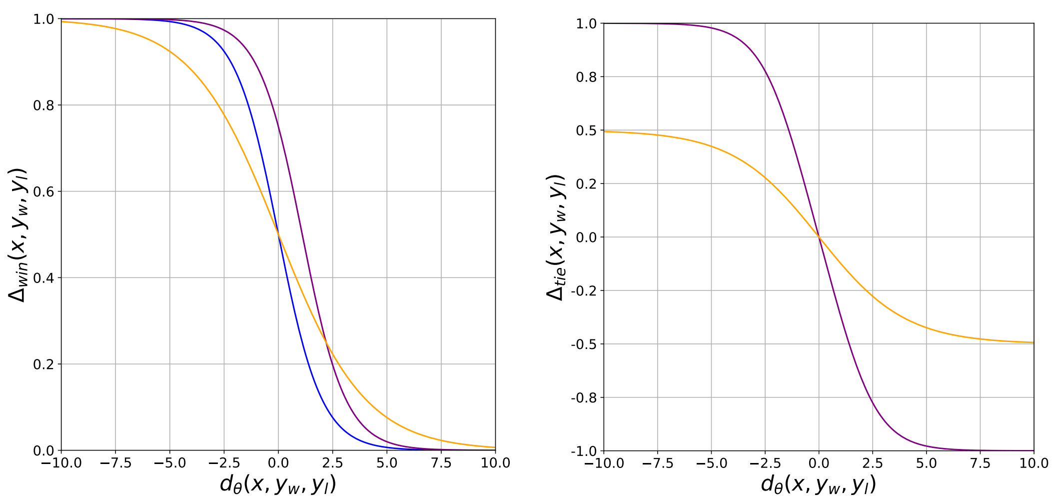


Figure 6: The gradient scale factors for DPO (blue) and DPO-RK (purple) and DPO-D (orange) as a function of reward margins $d_\theta(x, y_w, y_l)$ on clear preference pairs (left) and tied pairs (right).$\alpha_{RK} = \log 3$ and $\nu_D = 1$ are used in producing these plots.

C Experimental Details and Full Results

We provide additional details of our experiments on Neural Machine Translation and Summarization with respect to the SFT models, the training configurations, and the decoding procedures. All experiments are run with the random seed set to 0.

C.1 Neural Machine Translation

We largely follow Yang et al. (2024b) in our experimental setup for NMT where the preference dataset is obtained via sampling and BLEURT-based ranking as explained in Sec.3.1.

SFT Models On WMT-21 ZH-EN, we performed supervised fine-tuning on the BLOOMZ-mt-7b Muennighoff et al. (2023) using previous WMT test sets to obtain the SFT model from which we train with DPO/DPO-RK/DPO-D. The clear preference pairs and tied pairs are generated by sampling from this SFT model. On IWSLT-17 FR-EN, we use the pretrained BLOOMZ-mt-7b model directly in sampling clear preferences and tied pairs and in DPO fine-tuning, as we find further SFT leads to repetitive generation.

Training Details We use the RMSProp optimizer with the learning rate set to $5e^{-7}$ and the number of warm-up steps set to 150. All NMT experiments are run on two Nvidia A100-80G GPUs with an effective batch size of 4. We used FP32 for training the policy. The log-probabilities from the reference model are pre-computed with FP32 precision. Each training run takes ≈ 2 hours on WMT20 ZH-EN CP+TP data and ≈ 1 hour on IWSLT17 FR-EN data.

Decoding Following Yang et al. (2024b), we use beam search with a beam size = 4 to decode all models.

Held-out Clear Preference Pairs and Tied Pairs As explained in Sec.3.1, we curate held-out sets by generating translations by sampling on the WMT18 ZH-EN test set. Clear Preference Pairs and Tied Pairs are identified using their rankings under BLEURT exactly as done for WMT21 ZH-EN (Sec.3.2.2). This gives 3980 CPs and 3980 TPs for held-out evaluation.

C.2 Summarization

We follow Amini et al. (2024a) in experimental setups. The preference dataset is obtained via sampling and ranking with a DPO model without requiring an external reward model as explained in Sec.3.1.

SFT Model We follow Amini et al. (2024a) to supervise-finetune a Pythia-2.8B model Biderman et al. (2023) on the chosen responses in TL;DR train split for one epoch to obtain the initial checkpoint for preference learning. We use the summarization prompt provided in Appendix D.2 by Rafailov et al. (2023).

Training Details We use the RMSProp optimizer with the learning rate set to $5e^{-7}$ and the number of warm-up steps set to 150. All summarization experiments are run on two Nvidia A100-40G GPUs with an effective batch size of 64. We used FP32 for the policy and FP16 for the reference model. Each training run takes ≈ 7 hours on TL;DR CP+TP data.

Decoding We use greedy decoding for all models as we find it performs on-par or better than temperature sampling (Appendix C.3).

C.3 PairRM as a Proxy Evaluator for GPT-4

PairRM (Jiang et al., 2023) is a strong reward model that has been shown to be effective in curating preference datasets for iterative DPO training (Tran et al., 2023). In our experiments on TL;DR summarization, we use the PairRM reward model instead of GPT-4 for comparing generated summaries against human references. In this appendix, we show that win-rate as judged by PairRM is a good proxy for GPT4-0613 (OpenAI et al., 2024) win-rate on the TL;DR dataset Stiennon et al. (2020).

We generate summaries from SFT pythia-2.8B model by sampling at temperature $T = [0.0, 0.5, 1.0]$ and the DPO model ($\beta = 0.1$) trained on TL;DR's full training set at temperature $T = [0.0, 0.25, 0.5, 0.75, 1.0]$. Their win-rates against the 256 human-written summaries in the TL;DR valid-2 split as judged by GPT-4 and PairRM are tabulated in Table 3. We find that the win-rates by GPT-4 and PairRM are similar and that system rankings are generally preserved. We opt to use PairRM as our evaluation metric which enables us to conduct experiments faster and at lower costs.

System	GPT-4	PairRM
DPO		
T=1.0	23.4%	27.3%
T=0.75	40.2%	40.6%
T=0.5	52.3%	54.7%
T=0.25	46.9%	51.6%
T=0.0	50.4%	55.5%
SFT		
T=1.0	22.3%	23.0%
T=0.5	37.5%	38.7%
T=0.0	36.7%	39.8%

Table 3: Win-rate of Pythia-2.8B model SFT/DPO on TL;DR train against 256 human-written summaries as judged by GPT4-0613 and PairRM.

C.4 Verifying a Tied Pair Selection Strategy for TL;DR

As explained in Sec. 3.1, we use the reward model associated with the DPO model trained on TL;DR to identify summarizations that are similar in quality. Note that we are performing unsupervised labelling of ties in the DPO training data, which is somewhat more forgiving than the classification task discussed in other sections which requires labelling ties in held-out data not seen in training. We do however assume that the reward model should perform well on the data it was trained on.

To investigate these assumptions, we swap the preferred and the dispreferred responses in all tied pairs to form "reversed Tied Pairs" (rTP). If the responses in TP are truly similar in quality (i.e., it is acceptable to reverse the preference direction), training with DPO(CP+TP) and DPO(CP+rTP) should yield similar performing models. Furthermore, the DPO-RK and DPO-D learning procedures which explicitly model tied pairs should yield better performing model. We conduct experiments on TL;DR. Table 4 Right shows that the performance relation DPO-D(CP+TP) $\sim$ DPO-RK(CP+TP) $\succ$ DPO(CP+TP) $\sim$ DPO(CP+rTP) indeed holds for TL;DR, which suggests that our Tied Pair selection strategy is reasonable.

System	PairRM
DPO(CP+ TP)	58.6%
DPO(CP+rTP)	60.9%
DPO-RK(CP+TP)	68.0%
DPO-D(CP+TP)	68.8%

Table 4: Win-rates of Pythia-2.8B model DPO on TL;DR train against 256 human-written summaries as judged by PairRM. Systems were trained on CP+TP or CP+rTP data with DPO, DPO-RK, or DPO-D at fixed $\beta = 0.3$. For DPO-RK and DPO-D learning, rTP is equivalent to TP as there is no preference direction for ties.

C.5 Tabulated KL-Performance Results on NMT and Summarization

We tabulate the KL-Performance results shown in Figure 1 and Figure 3.

C.5.1 Neural Machine Translation

In addition to KL Divergence and BLEURT, we also provide COMET (Rei et al., 2020) scores, BLEU (Post, 2018) scores and BLEU's Length Ratio.

We observe the "reward hacking" phenomenon identified by Yang et al. (2024b) on both WMT21 ZH-EN and IWSLT17 FR-EN where systems achieve good BLEURT but have large length ratio (>1.5) and lower COMET than the pre-DPO system. These systems learn to generate long, repetitive translations which BLEURT fails to recognize as low-quality. Yang et al. (2024b) find that using small beta values (e.g. 0.1) in DPO training results in reward hacking models. Our results are consistent with their findings and further suggest that large KL divergence from the reference model is a good indicator for reward hacking. On WMT21 ZH-EN, the only model that exhibits reward hacking is trained by DPO(CP) with beta=0.1 which also yields the highest KL divergence (174.13) among all models, greatly exceeding the second-highest KL divergence (68.12). On IWSLT17 FR-EN, Almost all models with KL Divergence > 30 (DPO(CP), $\beta = 0.1$, DPO-RK(CP+TP), $\beta = 0.1$ and DPO-D(CP+TP) $\beta = 0.1, 0.5$) show reward hacking behaviours.

Reward hacking on NMT can be resolved by increasing regularization with respect to the reference model. We find that training with larger beta values or incorporating ties in DPO-RK/DPO-D training can provide such regularization without performance degradation.

System	beta	KL Divergence	BLEU	Length Ratio	COMET	BLEURT
Bloomz-mt-7b1-SFT	-	0	17.6		77.9	61.6
DPO(CP)	0.1	174.13	7.23	3.01	70.2	67.7
DPO(CP)	0.2	68.12	20.8	1.10	80.8	66.2
DPO(CP)	0.3	62.85	20.7	1.13	80.6	66.4
DPO(CP)	0.4	56.02	21.4	1.09	80.7	66.4
DPO(CP)	0.5	50.99	21.2	1.11	80.8	66.5
DPO(CP)	0.6	47.97	21.5	1.09	80.9	66.5
DPO(CP)	0.7	44.08	21.5	1.11	81.0	66.7
DPO(CP)	0.8	41.88	21.3	1.14	80.8	66.7
DPO(CP)	0.9	41.24	21.5	1.14	80.8	66.8
DPO(CP)	1.9	33.69	22.3	1.09	81.2	67.0
DPO(CP)	1.2	32.01	22.4	1.09	81.3	67.1
DPO(CP)	1.5	29.58	21.7	1.13	81.1	67.1
DPO(CP)	1.55	29.01	21.9	1.13	81.1	67.1
DPO(CP+TP)	0.1	51.29	20.3	1.16	80.0	66.0
DPO(CP+TP)	0.2	36.37	18.8	1.30	80.1	66.6
DPO(CP+TP)	0.3	26.15	19.5	1.24	80.2	66.6
DPO(CP+TP)	0.4	18.21	20.6	1.20	80.4	66.6
DPO(CP+TP)	0.5	15.47	21.2	1.15	80.4	66.4
DPO(CP+TP)	0.6	14.74	21.9	1.10	80.6	66.4
DPO(CP+TP)	0.7	13.29	22.1	1.11	80.5	66.4
DPO(CP+TP)	0.8	12.57	22.2	1.10	80.5	66.2
DPO(CP+TP)	0.9	12.10	21.9	1.10	80.5	66.3
DPO(CP+TP)	1.0	11.43	22.0	1.11	80.5	66.2
DPO-RK(CP+TP)	0.1	48.55	19.3	1.22	80.2	66.9
DPO-RK(CP+TP)	0.2	28.61	22.1	1.11	80.9	66.9
DPO-RK(CP+TP)	0.3	20.21	22.5	1.11	81.0	67.1
DPO-RK(CP+TP)	0.4	14.80	22.4	1.12	81.1	67.1
DPO-RK(CP+TP)	0.5	11.66	22.8	1.10	81.0	67.1
DPO-RK(CP+TP)	0.6	9.74	22.2	1.13	80.8	66.8
DPO-RK(CP+TP)	0.7	8.04	22.3	1.12	80.8	66.7
DPO-RK(CP+TP)	0.8	8.10	22.1	1.13	80.8	66.8
DPO-RK(CP+TP)	0.9	7.58	21.8	1.15	80.7	66.8
DPO-RK(CP+TP)	1.0	6.31	22.3	1.11	80.7	66.6
DPO-D(CP+TP)	0.2	42.74	21.4	1.13	80.8	66.6
DPO-D(CP+TP)	0.3	38.56	21.2	1.15	80.2	66.5
DPO-D(CP+TP)	0.4	17.01	22.5	1.11	81.0	67.1
DPO-D(CP+TP)	0.5	20.20	22.7	1.10	81.1	67.1
DPO-D(CP+TP)	0.6	26.85	22.3	1.10	81.1	66.9
DPO-D(CP+TP)	0.7	14.97	22.6	1.11	81.1	67.1
DPO-D(CP+TP)	0.8	13.33	22.7	1.11	81.1	67.1
DPO-D(CP+TP)	1.0	10.05	22.3	1.12	80.9	67.0

Table 5: KL-Performance evaluated on WMT-21 ZH-EN.

C.5.2 Summarization

Table 7 shows the KL-PairRM winrate on TL;DR summarization.

System	beta	KL Divergence	BLEU	Length Ratio	COMET	BLEURT
Bloomz-mt-7b1	-		17.6		85.4	74.8
DPO(CP)	0.1	53.60	25.8	1.40	82.3	74.7
DPO(CP)	0.3	30.80	23.7	1.60	83.6	76.5
DPO(CP)	0.5	16.70	36.8	1.00	86.1	76.2
DPO(CP)	0.7	13.80	38.5	1.00	86.4	76.4
DPO(CP)	1.0	12.40	38.6	1.00	86.5	76.5
DPO(CP)	1.2	11.80	38.8	0.98	86.5	76.5
DPO(CP)	1.5	10.70	38.9	0.99	86.5	76.5
DPO(CP+TP)	0.1	35.60	35.8	1.00	85.6	75.5
DPO(CP+TP)	0.3	25.80	35.7	1.10	85.4	75.9
DPO(CP+TP)	0.5	22.00	35.1	1.10	85.8	76.3
DPO(CP+TP)	0.7	17.00	38.7	1.00	86.3	76.3
DPO(CP+TP)	1.0	11.50	38.9	1.00	86.4	76.4
DPO(CP+TP)	1.2	8.50	39.1	0.98	86.5	76.4
DPO(CP+TP)	1.5	6.30	39.0	0.98	86.4	76.3
DPO-RK(CP+TP)	0.1	46.70	23.0	1.60	78.7	76.3
DPO-RK(CP+TP)	0.2	19.51	35.9	1.05	85.9	76.4
DPO-RK(CP+TP)	0.3	15.50	36.1	1.10	86.1	76.5
DPO-RK(CP+TP)	0.5	13.30	31.4	1.20	85.7	76.6
DPO-RK(CP+TP)	0.7	10.90	31.3	1.20	85.8	76.5
DPO-RK(CP+TP)	0.8	10.90	29.9	1.28	85.6	76.5
DPO-RK(CP+TP)	0.9	11.60	27.2	1.40	85.3	76.4
DPO-RK(CP+TP)	1.0	11.60	26.1	1.50	85.1	76.3
DPO-RK(CP+TP)	1.2	11.80	24.4	1.57	84.8	76.3
DPO-D(CP+TP)	0.1	48.60	25.3	1.41	82.6	76.3
DPO-D(CP+TP)	0.3	19.90	35.4	1.07	85.8	76.5
DPO-D(CP+TP)	0.5	51.90	8.4	4.35	75.1	76.1
DPO-D(CP+TP)	0.7	12.80	36.6	1.06	86.2	76.6
DPO-D(CP+TP)	1.0	10.30	37.8	1.03	86.3	76.6
DPO-D(CP+TP)	1.2	10.90	32.1	1.20	85.9	76.6

Table 6: KL-Performance evaluated on IWSLT17 FR-EN

System	beta	KL Divergence	PairRM Winrate
Pythia-2.8B-SFT, Greedy	-	0.00	37.5
DPO(CP)	0.025	97.03	67.9
DPO(CP)	0.05	60.31	70.3
DPO(CP)	0.07	57.14	71.5
DPO(CP)	0.08	38.16	66.4
DPO(CP)	0.10	26.82	62.5
DPO(CP)	0.30	9.97	63.7
DPO(CP)	0.50	5.79	59.0
DPO(CP)	0.70	3.78	57.8
DPO(CP+TP)	0.025	87.66	63.7
DPO(CP+TP)	0.03	119.60	66.8
DPO(CP+TP)	0.04	70.69	69.5
DPO(CP+TP)	0.05	35.39	63.3
DPO(CP+TP)	0.10	17.21	57.4
DPO(CP+TP)	0.30	4.50	58.6
DPO(CP+TP)	0.50	7.61	57.8
DPO(CP+TP)	0.70	2.91	55.9
DPO-RK(CP+TP)	0.04	80.86	65.2
DPO-RK(CP+TP)	0.05	62.57	67.2
DPO-RK(CP+TP)	0.10	40.50	67.6
DPO-RK(CP+TP)	0.20	22.24	67.6
DPO-RK(CP+TP)	0.30	12.45	68.0
DPO-RK(CP+TP)	0.50	6.15	65.6
DPO-RK(CP+TP)	0.70	4.33	61.7
DPO-D(CP+TP)	0.05	82.35	64.8
DPO-D(CP+TP)	0.10	54.06	71.5
DPO-D(CP+TP)	0.20	39.23	66.0
DPO-D(CP+TP)	0.30	22.46	68.8
DPO-D(CP+TP)	0.40	12.57	67.6
DPO-D(CP+TP)	0.50	9.92	67.2
DPO-D(CP+TP)	0.70	6.82	64.8

Table 7: KL-PairRM winrate against 256 human-written summaries on TL;DR summarization

C.6 Empirical Reward Margin Distributions

In Sec.3.2.3, we show that DPO(CP) yields models that often show strong preference for either one of a pair of translations even though the pairs are known to be ties. This is shown by the estimated preference probability $P(y_1 \succ y_2)$ on held-out tied pairs (Figure 4). For completeness, we provide the estimated preference probability of the same models on held-out clear preference pairs in Figure 7.

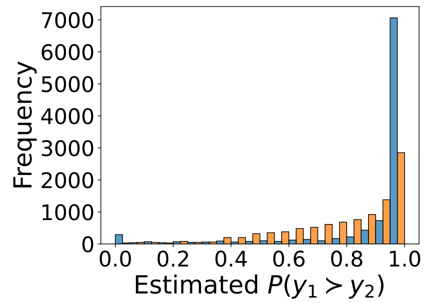
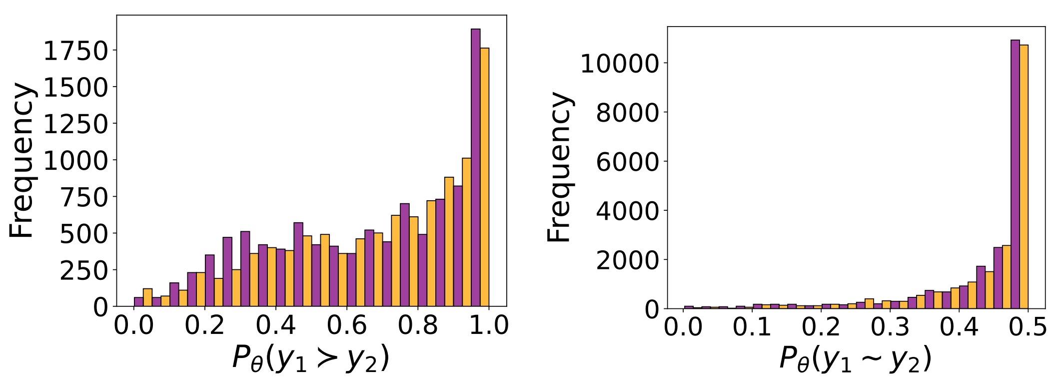


Figure 7: Empirical distribution of clear preference probabilities on clear preference pairs. DPO(CP) in blue, and DPO(CP+TP) in orange.

The DPO(CP) model correctly assigns high preference probability to most of the held-out CPs. This is consistent with its high classification accuracy on clear preference pairs in Table 1. Similar to the estimated preference probability on held-out TPs, the DPO(CP) model tends to give confident, clear preference judgment with > 0.8 probability in either direction. In comparison, the DPO(CP+TP) model is more conservative in making preference judgments, showing a less-sharp preference probability distribution over the held-out CP pairs. These results suggest that incorporating ties in DPO training leads to preference probability distributions that more evenly spread on both CPs and TPs as opposed to one concentrated on the two ends.

For completeness, we also show the clear preference/tie probability distributions produced by models trained with DPO-RK(CP+TP) and DPO-D(CP+TP) on held-out clear preference pairs and tied pairs. Figure 8 show that these distributions are well-behaved in that most of the probability mass are allocated to $P_\theta(y_1 \succ y_2) > 0.5$ on held-out clear preference pairs and to $P_\theta(y_1 \sim y_2) \approx 0.5$ on held-out tied pairs. We note that under our hyper-parameter setting for the Rao-Kupper and Davidson models, the maximal tie probability is 0.5.

All models in this analysis are trained with $\beta = 0.1$.



(a) Preference probability under the models on held-out clear preference pairs.

(b) Tie probability under the models on held-out tied pairs.

Figure 8: DPO-D (orange) and DPO-RK (purple) preference/tie probability on held-out sets under the Davidson and Rao-Kupper models, respectively.

D Simulating the Perfect DPO-Davidson Policy

In Section 3.1.1 we make use of the relationship derived by Chen et al. (2024, Appendix A.2) which specifies the optimal DPO policy to minimize the binary classification loss

$$\min_\pi \mathbb{P}(y_1 \succ_x y_2) \log \pi(y_1 \succ_x y_2) + (1 - \mathbb{P}(y_1 \succ_x y_2)) \log(1 - \pi(y_1 \succ_x y_2))$$

where $\mathbb{P}(y_1 \succ_x y_2)$ is the human ground truth preference distribution.

We extend the analysis of Chen et al. (2024) to include the Davidson model, noting that the binary maximum likelihood objective becomes ternary. We assume we have the ground-truth human preference distributions $\mathbb{P}(y_1 \succ_x y_2)$, $\mathbb{P}(y_2 \succ_x y_1)$, and $\mathbb{P}(y_1 \sim_x y_2)$ needed to define the objective. The resulting Theorem 1 can be viewed as a generalization of Theorem 3 of Chen et al. (2024) that allows for the observations of ties. Where ties are not allowed (i.e. $\nu_D = 0$), the Davidson model simplifies to the Bradley-Terry model and Theorem 3 of Chen et al. (2024) is recovered as a special case of Theorem 1.

Theorem 1 (Simulating Perfect DPO-D Policy)**.** *Assume we are given an aggregated comparison datapoint (x, y_1, y_2) and human ground-truth preference probabilities $\mathbb{P}(y_1 \succ_x y_2)$, $\mathbb{P}(y_1 \succ_x y_2)$, and $\mathbb{P}(y_1 \sim_x y_2)$ which obey the Davidson model with hyper-parameter ν_D. Let the reference model be π_{ref}. It follows that the perfect DPO-Davidson policy π^* on this aggregated comparison datapoint satisfies*

$$\frac{\pi^*(y_1|x)}{\pi^*(y_2|x)} = \frac{\pi_{ref}(y_1|x)}{\pi_{ref}(y_2|x)} \Big(\frac{\mathbb{P}(y_1 \succ_x y_2)}{\mathbb{P}(y_2 \succ_x y_1)} \Big)^{1/\beta} \tag{25}$$

or equivalently

$$\frac{\pi^*(y_1|x)}{\pi^*(y_2|x)} = \frac{\pi_{ref}(y_1|x)}{\pi_{ref}(y_2|x)} \Big(2\nu_D \frac{\mathbb{P}(y_1 \succ_x y_2)}{\mathbb{P}(y_1 \sim_x y_2)} \Big)^{2/\beta} \tag{26}$$

Proof. The DPO-D policy objective optimizes the following three-way classification loss:

$$\min_{\pi} \mathbb{P}(y_1 \succ_x y_2) \log \pi(y_1 \succ_x y_2) + \mathbb{P}(y_2 \succ_x y_1) \log \pi(y_2 \succ_x y_1) + \mathbb{P}(y_1 \sim_x y_2) \log \pi(y_1 \sim_x y_2)$$

Let θ^* denotes a set of parameters such that π_{θ^*} is an optimal policy for the above loss, then π_{θ^*} satisfies:

$$\begin{aligned} \pi_{\theta^*}(y_1 \succ_x y_2) &= \mathbb{P}(y_1 \succ_x y_2) \\ \pi_{\theta^*}(y_2 \succ_x y_1) &= \mathbb{P}(y_2 \succ_x y_1) \\ \pi_{\theta^*}(y_1 \sim_x y_2) &= \mathbb{P}(y_1 \sim_x y_2) \end{aligned}$$

Expressing the policy probability $\pi_{\theta^*}(y_w \succ_x y_l)$ and $\pi_{\theta^*}(y_l \succ_x y_w)$ in terms of the reward margins $d_{\theta^*}(x, y_w, y_l)$:

$$\begin{aligned} \mathbb{P}(y_1 \succ_x y_2) &= \frac{1}{1 + e^{-d_{\theta^*}(x,y_w,y_l)} + 2\nu_D e^{-d_{\theta^*}(x,y_w,y_l)/2}} \\ \mathbb{P}(y_2 \succ_x y_1) &= \frac{e^{-d_{\theta^*}(x,y_1,y_2)}}{1 + e^{-d_{\theta^*}(x,y_1,y_2)} + 2\nu_D e^{-d_{\theta^*}(x,y_1,y_2)/2}} \end{aligned}$$

Rearranging, we have

$$\frac{\mathbb{P}(y_2 \succ_x y_1)}{\mathbb{P}(y_1 \succ_x y_2)} = \exp\big(-d_{\theta^*}(x, y_1, y_2)\big) = \exp\Big(\beta \log \frac{\pi_{\theta^*}(y_2|x)}{\pi_{ref}(y_2|x)} - \beta \log \frac{\pi_{\theta^*}(y_1|x)}{\pi_{ref}(y_1|x)}\Big)$$

Taking logarithms on both side and divide by β.

$$\frac{1}{\beta} \log \frac{\mathbb{P}(y_1 \succ_x y_2)}{\mathbb{P}(y_2 \succ_x y_1)} = \log \frac{\pi_{\theta^*}(y_2|x)\pi_{ref}(y_1|x)}{\pi_{ref}(y_2|x)\pi_{\theta^*}(y_1|x)}$$

Exponentiating both sides gives

$$\frac{\pi_{\theta^*}(y_2|x)}{\pi_{\theta^*}(y_1|x)} = \frac{\pi_{ref}(y_2|x)}{\pi_{ref}(y_1|x)} \Big(\frac{\mathbb{P}(y_2 \succ_x y_1)}{\mathbb{P}(y_1 \succ_x y_2)} \Big)^{1/\beta}$$

Taking the inverse yields Eq 25.

To see the equivalence between Eq 25 and Eq 26, note that the ground-truth preference and tie probabilities which obey the Davidson model satisfy the following relation:

$$\mathbb{P}(y_1 \sim_x y_2) = 2\nu_D \sqrt{\mathbb{P}(y_1 \succ_x y_2)\mathbb{P}(y_2 \succ_x y_1)}$$

Rearranging Eq 25:

$$\begin{aligned}
\frac{\pi^*(y_1|x)}{\pi^*(y_2|x)} &= \frac{\pi_{ref}(y_1|x)}{\pi_{ref}(y_2|x)}\Big(\frac{\mathbb{P}(y_1 \succ_x y_2)}{\mathbb{P}(y_2 \succ_x y_1)}\Big)^{1/\beta} \\
&= \frac{\pi_{ref}(y_1|x)}{\pi_{ref}(y_2|x)}\Big(\sqrt{\frac{\mathbb{P}(y_1 \succ_x y_2)}{\mathbb{P}(y_2 \succ_x y_1)}}\Big)^{2/\beta} \\
&= \frac{\pi_{ref}(y_1|x)}{\pi_{ref}(y_2|x)}\Big(\frac{\mathbb{P}(y_1 \succ_x y_2)}{\sqrt{\mathbb{P}(y_1 \succ_x y_2)\mathbb{P}(y_2 \succ_x y_1)}}\Big)^{2/\beta} \\
&= \frac{\pi_{ref}(y_1|x)}{\pi_{ref}(y_2|x)}\Big(2\nu_D\frac{\mathbb{P}(y_1 \succ_x y_2)}{\mathbb{P}(y_1 \sim_x y_2)}\Big)^{2/\beta}
\end{aligned}$$

which is Eq 26. □