Regularized Best-of-N Sampling to Mitigate Reward Hacking for Language Model Alignment

Yuu Jinnai [1] **Tetsuro Morimura** [1] **Kaito Ariu** [1] **Kenshi Abe** [1]

Abstract

Best-of-N (BoN) sampling with a reward model has been shown to be an effective strategy for aligning Large Language Models (LLMs) to human preferences at the time of decoding. BoN sampling is susceptible to a problem known as *reward hacking*. Because the reward model is an imperfect proxy for the true objective, over-optimizing its value can compromise its performance on the true objective. A common solution to prevent reward hacking in preference learning techniques is to optimize a reward using proximity regularization (e.g., KL regularization), which ensures that the language model remains close to the reference model. In this research, we propose Regularized Best-of-N (RBoN), a variant of BoN that aims to mitigate reward hacking by incorporating a proximity term in response selection, similar to preference learning techniques. We evaluate RBoN on the AlpacaFarm and Anthropic's hh-rlhf datasets and find that it outperforms BoN. As an application of RBoN, we use RBoN to generate a pairwise preference learning dataset. Experimental results show that a DPO model trained on a dataset generated with RBoN outperforms a DPO model generated with vanilla BoN. Our code is available at https://github.com/CyberAgentAILab/regularized-bon.

1. Introduction

While Large language models (LLMs) trained on massive datasets have been shown remarkable ability at next-token prediction, these models often produce misleading, harmful, and unhelpful outputs (Bai et al., 2022; Lin et al., 2022; Touvron et al., 2023; Casper et al., 2023; Guan et al., 2024). The challenge for the field is thus to *align* the behavior of the LLMs with human preferences, steering the models to generate informative, harmless, and helpful responses (Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022).

Alignment strategies can be divided into two categories: preference learning and decoding time alignment. Reinforcement learning from human feedback (RLHF) and Direct Preference Optimization (DPO) are popular strategies for preference learning (Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023). RLHF begins by training a reward model that reflects human preferences for the model's responses. It then trains the model to optimize the responses to maximize the score according to the reward model. DPO trains the language model to align directly with human preference data about responses without training a reward model. Best-of-N (BoN) sampling is widely used to align the LLM at decoding time (Stiennon et al., 2020; Nakano et al., 2022). BoN samples N responses from the language model and selects the best response according to the proxy reward model as the output of the system.

However, these alignment methods are known to suffer from the *reward hacking* problem (Amodei et al., 2016; Ziegler et al., 2020; Stiennon et al., 2020; Skalse et al., 2022; Gao et al., 2023). The reward hacking problem occurs because of reward misspecification (Pan et al., 2022; Lambert & Calandra, 2024); the proxy reward trained from human preferences does not perfectly reflect true human preferences. As a result, optimizing for the reward model does not always optimize for the preference of the true intended objective.

A common approach to mitigate reward hacking in preference learning (RLHF and DPO) is to use proximity regularization, typically by a KL divergence term, to keep the trained model close to the reference model. Previous work in BoN has shown that reducing the number of samples N mitigates the reward hacking (Nakano et al., 2022; Pan et al., 2022; Lambert & Calandra, 2024). This approach successfully reduces the KL divergence from the reference policy (Nakano et al., 2022; Beirami et al., 2024) but at the expense of diminished improvement obtained by the method.

To this end, we propose Reguralized Best-of-N (RBoN), a method that introduces proximity regularization into the BoN to mitigate the reward hacking problem. Instead of

[1]CyberAgent, Tokyo, Japan. Correspondence to: Yuu Jinnai <jinnai_yu@cyberagent.co.jp>.

ICML 2024 Workshop on Models of Human Feedback for AI Alignment, Vienna, Austria. 2024.

optimizing the raw reward score, we optimize a sum of the reward score and a regularization term. RBoN can tune the regularization strength by the hyperparameter β, similar to the proximity regularization in RLHF and DPO. We propose two variants of RBoN, RBoN_{KL} and RBoN_{WD}. RBoN_{KL} is simple to implement, but its improvement over BoN is limited, while RBoN_{WD} significantly improves over BoN.

We evaluate the performance of RBoN on the AlpacaFarm (Dubois et al., 2023) and Anthropic's hh-rlhf datasets (Bai et al., 2022) and show that it outperforms the performance of vanilla BoN in a wide range of settings. We also use RBoN to generate a pairwise preference learning dataset and show that a DPO model trained on a dataset generated with RBoN outperforms a DPO model trained on a dataset generated with vanilla BoN.

2. Background

First, we give an overview of preference learning algorithms including RLHF and DPO. Then we introduce the decoding-time alignment algorithm, BoN sampling.

2.1. Preference Learning

Let $\mathcal{D}$ be a dataset of comparisons $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}$. RLHF uses the learned reward function to train the language model. Typically, the RL process is formulated as the following optimization problem:

$$\arg\max_{\pi} \mathop{\mathbb{E}}_{x\sim\mathcal{D}} \mathop{\mathbb{E}}_{y\sim\pi(y|x)} [R(x,y)] - \beta \mathbb{D}_{\text{KL}}[\pi(\cdot|x)||\pi_{\text{ref}}(\cdot|x)], \quad (1)$$

where β is a hyperparameter that controls the proximity to the base reference model π_{ref}. The regularization term $\mathbb{D}_{\text{KL}}$ is important to prevent the model from deviating too far from the base model. Since the objective is not differentiable, reinforcement learning algorithms are used for optimization (Schulman et al., 2017; Stiennon et al., 2020; Bai et al., 2022; Ouyang et al., 2022; Zheng et al., 2023).

DPO trains the language model to align directly with the human preference data over the responses, so it doesn't need a separate reward model (Rafailov et al., 2023). Although DPO is based on supervised learning rather than reinforcement learning, it uses essentially the same loss function under the Bradley-Terry model (Bradley & Terry, 1952). The objective function of the DPO is the following:

$$\arg\max_{\pi} \mathop{\mathbb{E}}_{(x,y_w,y_l)\sim\mathcal{D}} [\log \sigma(\beta \log \frac{\pi(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi(y_l|x)}{\pi_{\text{ref}}(y_l|x)})], \quad (2)$$

where σ is the sigmoid function. Several variants of DPO also use KL divergence as proximity regularization (Azar et al., 2023; Liu et al., 2024). (Wang et al., 2024) investigate the use of f-divergence as a proximity regularization instead of KL divergence to promote the diversity of the resulting model. (Zhao et al., 2023) uses the cross-entropy loss instead of the KL term because it does not require a copy of the reference model to compute.

Thus, both lines of work in preference optimization have proximity regularization in common, often as a form of KL divergence, to keep the model π close to the reference model π_{ref}.

2.2. Best-of-N (BoN) Sampling

While many methods have been proposed for learning human preferences, a simple, popular, and well-performing method for preference optimization remains Best-of-N (BoN) sampling (Stiennon et al., 2020; Nakano et al., 2022). Let x be an input prompt to the language model π_{ref}. Let Y_{ref} be N responses drawn from $\pi_{\text{ref}}(\cdot|x)$. BoN sampling selects the response with the highest reward score according to the proxy reward model R:

$$y_{\text{BoN}}(x) = \arg\max_{y\in Y_{\text{ref}}} R(x,y). \quad (3)$$

The advantages of BoN over preference learning methods are as follows. First, BoN is simple. It does not require any additional training of the language model. While learning-based alignment methods need to train the LLM, BoN can be applied on the fly. Every time human preferences are updated, learning-based methods must retrain the LLM to adapt to them. On the other hand, BoN only requires an update of the reward model and does not require the training of the LLM, which is the most expensive process. Second, BoN is an effective strategy in its own right. Several previous works have shown that BoN sampling can outperform learning-based alignment methods (Gao et al., 2023; Eisenstein et al., 2023; Mudgal et al., 2024; Gui et al., 2024). Third, BoN is applicable to a black-box model where fine-tuning is not available. BoN does not require access to the model itself and is applicable using the output sequences from the black-box model. In summary, BoN is a practical and efficient alignment strategy that complements the shortcomings of learning-based strategies and is worthy of investigation.

3. Regularized Best-of-N (RBoN) Sampling

We propose Regularized Best-of-N (RBoN), a variant of BoN with a proximity regularization to mitigate the reward hacking. In particular, we present two instances of RBoN, KL regularized BoN (RBoN_{KL}) and Wasserstein Distance regularized BoN (RBoN_{WD}).

3.1. Kullback–Leibler divergence Regularized Best-of-N (RBoN_{KL})

We derive the objective function of RBoN_{KL} directly from Eq. (1) as follows:

$$y_{\text{RBoN}_{\text{KL}}}(x) = \arg\max_{y \in Y_{\text{ref}}} R(x,y) - \beta \mathbb{D}_{\text{KL}}[\mathbb{1}_y || \pi_{\text{ref}}(\cdot|x)], \quad (4)$$

where $\mathbb{1}_y$ is an indicator function with $\mathbb{1}_y(y') = 1$ for $y' = y$ and $\mathbb{1}_y(y') = 0$ for $y' \neq y$. $\mathbb{1}_y$ represents the policy of choosing y with a probability of 1, which is the policy RBoN will end up with if it chooses y as the output. Thus, $\mathbb{D}_{\text{KL}}[\mathbb{1}_y || \pi_{\text{ref}}(\cdot|x)]$ represents the KL divergence between the resulting policy and the reference policy. Intuitively, RBoN_{KL} optimizes the same objective as Eq. (1) but with modifications to make it available at decoding time. Eq. (4) is derived from Eq. (1) by computing the optimal response for a given x instead of computing the optimal policy.

The tradeoff between the reward and the proximity to the reference model is controlled by the hyperparameter β. With a small β, the output is more aligned with the proxy reward model. With $\beta = 0$, the vanilla BoN is restored. With larger β, the output is closer to the behavior of the reference model π_{ref}, where $\beta = +\infty$ selects the response with the highest model probability, recovering the maximum a posteriori (MAP) decoding (Stahlberg & Byrne, 2019; Eikema & Aziz, 2020; Holtzman et al., 2020).

Previous work in BoN has shown that the deviation from the reference model can be reduced by using a smaller number of samples N (Nakano et al., 2022; Pan et al., 2022; Beirami et al., 2024). However, reducing the number of samples also reduces the alignment to the reward model. RBoN_{KL} does not need to reduce the number of samples to enforce proximity to the reference model, while using all generated samples to find the best response according to the objective function.

3.2. Wasserstein Distance Regularized Best-of-N (RBoN_{WD})

Although RBoN_{KL} is simple and straightforward, the performance of RBoN_{KL} is sensitive to the choice of β (Appendix B). As such, the performance improvement of RBoN_{KL} over BoN is limited. In addition, RBoN_{KL} is not applicable to black-box models as it requires access to the probability of the sampled outputs.

To this end, we propose RBoN_{WD}, a regularized BoN based on Wasserstein Distance (WD) (Peyré & Cuturi, 2020; Villani, 2021) instead of KL divergence. WD is a distance function defined between probability distributions. It is also known as the earth mover's distance because it is often expressed as the amount of dirt moved times the distance moved. Let $\mathcal{Y}$ be the set of all possible responses for any input x. For a pair of probability distributions over $\mathcal{Y}$, WD is defined as follows:

$$WD(P,Q) = \min_{\{\mu_{i,j}\}_{i,j} \in \mathcal{J}} \sum_{i=1}^{|\mathcal{Y}|} \sum_{j=1}^{|\mathcal{Y}|} \mu_{i,j} C(y_i, y_j), \quad (5)$$

where C is the cost function that represents the dissimilarity of the inputs. $\mathcal{J}$ is a set of all couplings $\{\mu_{i,j}\}_{i,j}$ (Villani, 2021):

$$\mathcal{J} = \Big\{ \{\mu_{i,j}\}_{i,j} : \sum_{i=1}^{|\mathcal{Y}|} \mu_{i,j} = Q(y_j), \sum_{j=1}^{|\mathcal{Y}|} \mu_{i,j} = P(y_i), \mu_{i,j} \geq 0 \Big\}. \quad (6)$$

RBoN_{WD} uses WD instead of KL divergence as the proximity regularizer:

$$y^*_{\text{RBoN}_{\text{WD}}}(x) = \arg\max_{y \in Y_{\text{ref}}} R(x,y) - \beta WD(\mathbb{1}_y, \pi_{\text{ref}}(\cdot|x)), \quad (7)$$

where β is a hyperparameter to adjust the strength of the regularization. The WD term serves as a proximity regularizer to ensure that the resulting policy is close to the reference policy π_{ref}.

Computing $WD(\mathbb{1}_y, \pi_{\text{ref}}(\cdot|x))$ exactly is intractable, since it requires computing the cost $C(y, y')$ for every possible response y' with $\pi_{\text{ref}}(y'|x) > 0$. Thus, we approximate π_{ref} with an empirical distribution using Y_{ref}. Let $\hat{\pi}_{\text{ref}}$ be the empirical distribution computed from a set of samples Y_{ref}: $\hat{\pi}_{\text{ref}}(y|x) = \frac{1}{N} \sum_{y' \in Y_{\text{ref}}} \mathbb{I}(y = y')$. $WD(\mathbb{1}_y, \hat{\pi}_{\text{ref}}(\cdot|x))$ is computable because the support of $\hat{\pi}_{\text{ref}}$ is less than or equal to the number of samples $|Y_{\text{ref}}|$. Then, RBoN_{WD} using the empirical distribution $\hat{\pi}_{\text{ref}}$ is as follows:

$$y_{\text{RBoN}_{\text{WD}}}(x) = \arg\max_{y \in Y_{\text{ref}}} R(x,y) - \beta WD(\mathbb{1}_y, \hat{\pi}_{\text{ref}}(\cdot|x)). \quad (8)$$

$WD(\mathbb{1}_y, \hat{\pi}_{\text{ref}}(\cdot|x))$ of Eq. 8 can be computed as follows:

$$WD(\mathbb{1}_y, \hat{\pi}_{\text{ref}}(\cdot|x)) = \sum_{y' \in Y_{\text{ref}}} \frac{1}{N} C(y, y'), \quad (9)$$

See Appendix A for the derivation of Eq. (9).

The advantage of RBoN_{WD} is that the WD term is a useful target for text generation on its own, in addition to being a proximal regularizer. The WD term is used as an objective function in Minimum Bayes Risk (MBR) decoding (Kumar & Byrne, 2002; 2004; Eikema & Aziz, 2022), which generates a sequence with minimum Bayes risk (hence maximum expected utility):

$$y_{\text{MBR}}(x) = \arg\max_{y \in Y_{\text{ref}}} \sum_{y' \in Y_{\text{ref}}} \frac{1}{N} U(y, y') \quad (10)$$

Minimizing Bayes risk is equivalent to minimizing the expected cost (Eq. 9), and thus minimizing $WD(\mathbb{1}_y, \hat{\pi}_{\text{ref}}(\cdot|x))$. In fact, RBoN$_{\text{WD}}$ (Eq. 9) with $\beta = +\infty$ recovers the MBR decoding with $U = -C$ as a utility function:

$$\begin{aligned} & y_{\text{RBoN}_{\text{WD}}}(x) \\ & \xrightarrow{\beta \to +\infty} \arg\max_{y \in Y_{\text{ref}}} -WD(\mathbb{1}_y, \hat{\pi}_{\text{ref}}(\cdot|x)) \\ & = \arg\max_{y \in Y_{\text{ref}}} \sum_{y' \in Y_{\text{ref}}} -\frac{1}{N} C(y, y') \\ & = \arg\max_{y \in Y_{\text{ref}}} \sum_{y' \in Y_{\text{ref}}} \frac{1}{N} U(y, y') \\ & = y_{\text{MBR}}(x). \end{aligned} \tag{11}$$

The MBR objective is shown to be effective and outperforms MAP decoding in a variety of text generation tasks (Suzgun et al., 2023; Bertsch et al., 2023; Li et al., 2024). As such, RBoN$_{\text{WD}}$ (Eq. 7) is a combination of two effective objectives. Therefore, we expect the performance of RBoN$_{\text{WD}}$ to be less sensitive to the choice of β.

We use a cosine distance between the sentence embeddings of the sequences as the cost function of WD:

$$C(y, y') = 1 - \cos(\text{emb}(y), \text{emb}(y')), \tag{12}$$

where emb represents the embedding function (e.g. sentence BERT (Reimers & Gurevych, 2019)). Using this cost function, RBoN$_{\text{WD}}$ is computed as follows:

$$\begin{aligned} y_{\text{RBoN}_{\text{WD}}}(x) = & \arg\max_{y \in Y_{\text{ref}}} R(x, y) \\ & - \frac{\beta}{N} \sum_{y' \in Y_{\text{ref}}} (1 - \cos(\text{emb}(y), \text{emb}(y'))). \end{aligned} \tag{13}$$

Similar to RBoN$_{\text{KL}}$, the hyperparameter β controls the trade-off between the reward and proximity to the reference model. Using a small β makes the output more aligned to the proxy reward, with $\beta = 0$ recovering the vanilla BoN. With a larger β, the output is closer to the behavior of the reference model π_{ref}, with $\beta = +\infty$ recovering the MBR decoding.

The computational overhead of RBoN$_{\text{WD}}$ is marginal as the computation of the WD (Eq. 9) for all $y \in Y_{\text{ref}}$ is computable in $O(N)$ time by reference aggregation (Vamvas & Sennrich, 2024).

4. Experiments

We evaluate the performance of RBoN for two use cases. First, we evaluate the performance of RBoN for decoding time alignment (Section 4.1). Then, we evaluate RBoN as a sampling strategy to generate a preference learning dataset to be used for DPO (Section 4.2).

Table 1: Average Spearman's rank correlation coefficient of the proxy reward models to the gold reference reward model (Eurus). For each instruction in the AlpacaFarm, 128 responses are generated using Mistral. Spearman's rank correlation for each instruction is computed and then averaged over the set of instructions.

Proxy reward	Correlation Coefficient
SHP-Large	0.32
SHP-XL	0.39
OASST	0.40

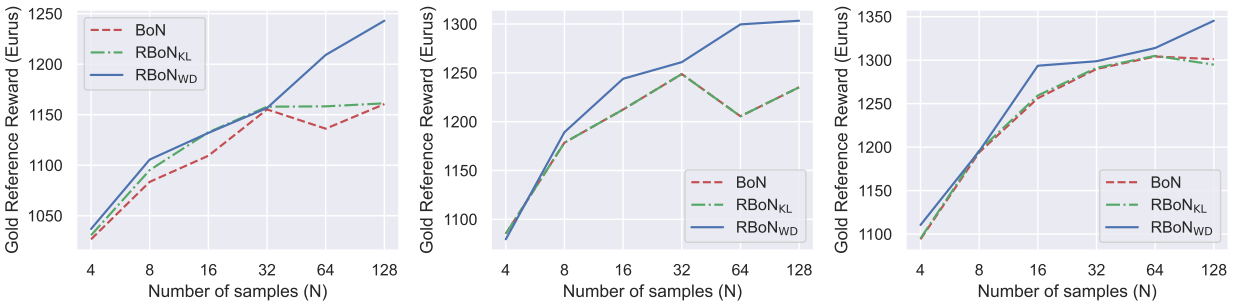
4.1. RBoN for Decoding-Time Alignment

Setup. The evaluation is conducted using the AlpacaFarm (Dubois et al., 2023) and Anthropic's hh-rlhf datasets (Bai et al., 2022). For the AlpacaFarm dataset, we use the first 1000 entries of the train split (`alpaca_human_preference`) as the development set and the whole evaluation split (`alpaca_farm_evaluation`) (805 instructions) as a test dataset. For Anthropic's datasets, we conduct experiments on the `helpful-base` (Helpfulness) and `harmless-base` (Harmlessness) subsets separately. For each subset, we use the first 1000 entries of the train split as the development set and the first 1000 entries of the test split as a test dataset. We use `mistral-7b-sft-beta` (Mistral) and `dolly-v2-3b` (Dolly) as the language models (Jiang et al., 2023a; Tunstall et al., 2023; Conover et al., 2023).

To evaluate RBoN under various conditions, we use SHP-Large, SHP-XL (Ethayarajh et al., 2022), and OASST (Köpf et al., 2023) as proxy reward models. We use Eurus as a gold reference reward model as it is one of the most accurate reward models according to the RewardBench (Lambert et al., 2024) and is open-source which makes the experiments reproducible. The results using other reward models as a gold reference are reported in Appendix B. The average Spearman's rank correlation coefficient ρ (Spearman, 1904) to the gold reference reward (Eurus) is reported in Table 1.

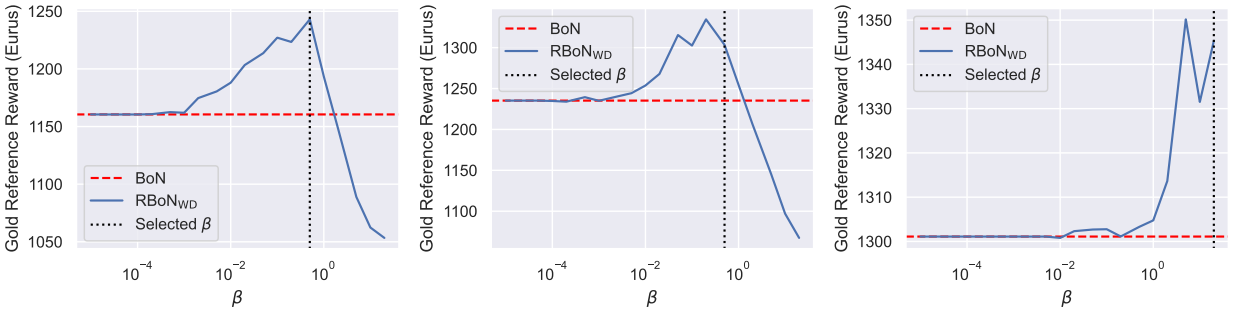
We compare the performance of BoN, RBoN$_{\text{KL}}$, and RBoN$_{\text{WD}}$. We sample up to $N = 128$ responses per instruction using nucleus sampling and select the output using the algorithms. We set the top-p to be $p = 0.9$ and the temperature to be $T = 1.0$ for the nucleus sampling (Holtzman et al., 2020). For a fair comparison, we use the same set of N responses for all algorithms. We use the Sentence BERT model (Reimers & Gurevych, 2019) based on MPNet (Song et al., 2020) to compute the sentence embedding for RBoN$_{\text{WD}}$.

RBoN$_{\text{KL}}$ and RBoN$_{\text{WD}}$ use the development set to select



(a) Proxy: SHP-Large (b) Proxy: SHP-XL (c) Proxy: OASST

Figure 1: Evaluation of the RBoN using Mistral on the AlpacaFarm dataset. The proxy rewards are SHP-Large, SHP-XL, and OASST. The gold reference reward is Eurus.



(a) Proxy: SHP-Large (b) Proxy: SHP-XL (c) Proxy: OASST

Figure 2: Evaluation of the RBoN using Mistral on the AlpacaFarm dataset with varying regularization strength β. The proxy rewards are SHP-Large, SHP-XL, and OASST. The gold reference reward is Eurus. The number of samples is $N = 128$.

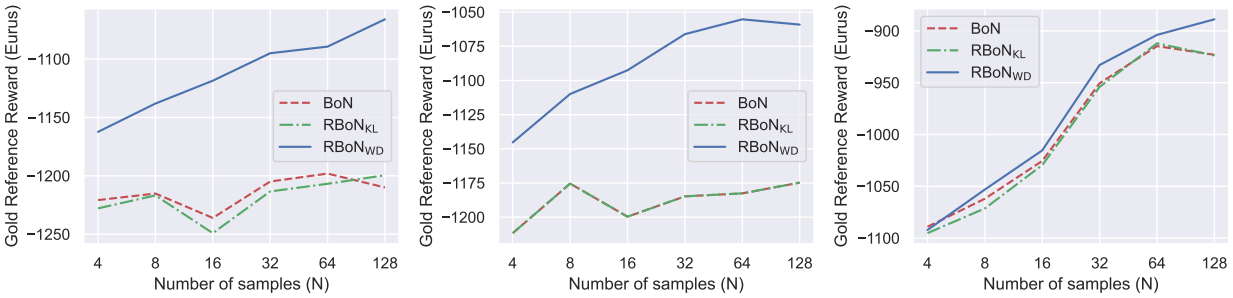
the optimal β. For each pair of a proxy reward and a gold reference reward, we run RBoN$_{KL}$ and RBoN$_{WD}$ with $\beta \in \{10^{-6}, 2\cdot10^{-6}, 5\cdot10^{-6}, 10^{-5}, ..., 2\cdot10^{1}\}$ and pick the best performing β. We pick the β with $N = 128$ and use the same β for all N in evaluation. See Appendix B for the ablation study on the regularization strength β.

Results. Figures 1 and 3 show the performance of BoN, RBoN$_{KL}$, and RBoN$_{WD}$ on the AlpacaFarm dataset using Mistral and Dolly as a language model, evaluated by Eurus score. RBoN$_{WD}$ outperforms BoN on all the settings. RBoN$_{KL}$ significantly outperforms BoN using Dolly and SHP reward models, but is on par with BoN in other settings. Figures 2 and 4 show the performance of RBoN$_{WD}$ with $N = 128$ and with varying regularization strength β. The vertical line shows the β selected using the development set. Overall, RBoN$_{WD}$ outperforms BoN in a wide range of β and is relatively robust to the choice of the β.

Figures 5 and 6 show the results on the Helpfulness and Harmlessness subsets of the Anthropic's hh-rlhf dataset. RBoN$_{WD}$ consistently outperforms BoN in both datasets, showing that the method is competitive in a wide range of tasks.

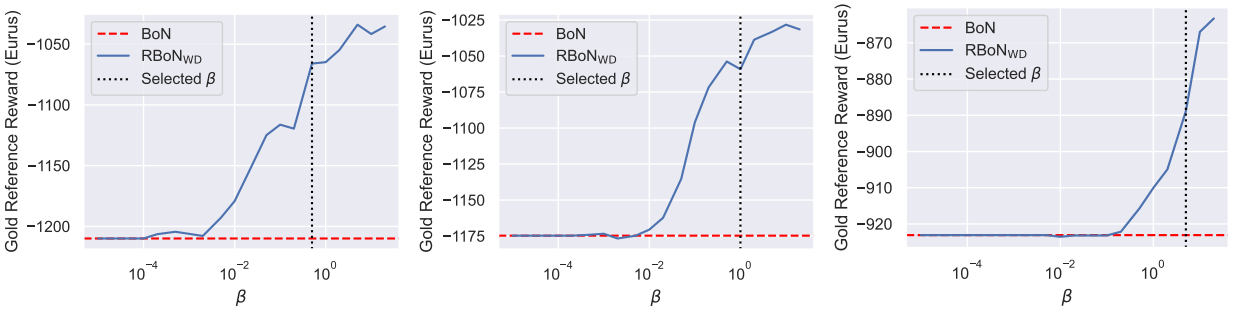
As expected, we observe that RBoN$_{WD}$ and RBoN$_{KL}$ have lower proxy reward scores than BoN (See Appendix B). The regularization term effectively mitigates the reward hacking of the BoN, resulting in a higher score in the gold reference score (Eurus).

Choice of Regularization strength. Table 2 summarizes the regularization strength β picked using the development set. The optimal value of β depends on the choice of the language model, dataset, and proxy reward model, which requires the use of the development set to tune the hyperparameter β. Note that the computational cost of tuning the hyperparameter for RBoN is marginal compared to that of RLHF or DPO as it does not involve any training of the language model or reward model. Running RBoN with different β only requires the computation of a multiplication of a matrix of size $N \times N$ and a vector of size N (Eq. 4



(a) Proxy: SHP-Large (b) Proxy: SHP-XL (c) Proxy: OASST

Figure 3: Evaluation of the RBoN using Dolly on the AlpacaFarm dataset. The proxy rewards are SHP-Large, SHP-XL, and OASST. The gold reference reward is Eurus. Note that the RBoN$_{KL}$ is overlapping with BoN when $\beta = 0$ is chosen as the optimal β for RBoN$_{KL}$ (Table 2).



(a) Proxy: SHP-Large (b) Proxy: SHP-XL (c) Proxy: OASST

Figure 4: Evaluation of the RBoN using Mistral on the AlpacaFarm dataset with varying regularization strength β. The proxy rewards are SHP-Large, SHP-XL, and OASST. The gold reference reward is Eurus. The number of samples is $N = 128$.

and 13).

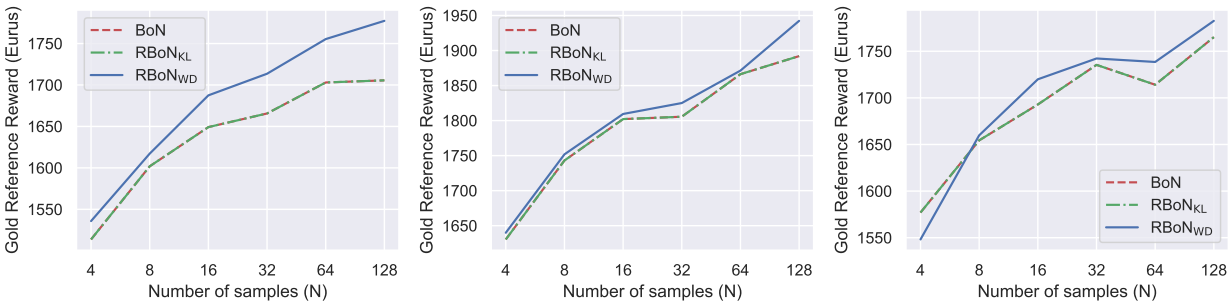
4.2. RBoN for Generating Pairwise Preference Learning Dataset

Previous work has shown that BoN sampling is an effective strategy for generating an efficient preference dataset (Xu et al., 2023; Yuan et al., 2024b; Pace et al., 2024). They show that the efficiency of pairwise preference learning is improved by using the best and worst responses according to the reward model.

We evaluate the performance of RBoN$_{WD}$ compared to BoN in generating a preference dataset for DPO (Rafailov et al., 2023). We use the instructions from the `alpaca_human_preference` subset of the AlpacaFarm dataset as the training dataset. We sample 128 responses for each instruction and use the response selected by RBoN$_{WD}$ or BoN as the chosen response and the response with the lowest reward as the rejected response. We use Mistral as the language model to generate the pairwise preference dataset and train it using the generated dataset (Jiang et al., 2023a; Tunstall et al., 2023).

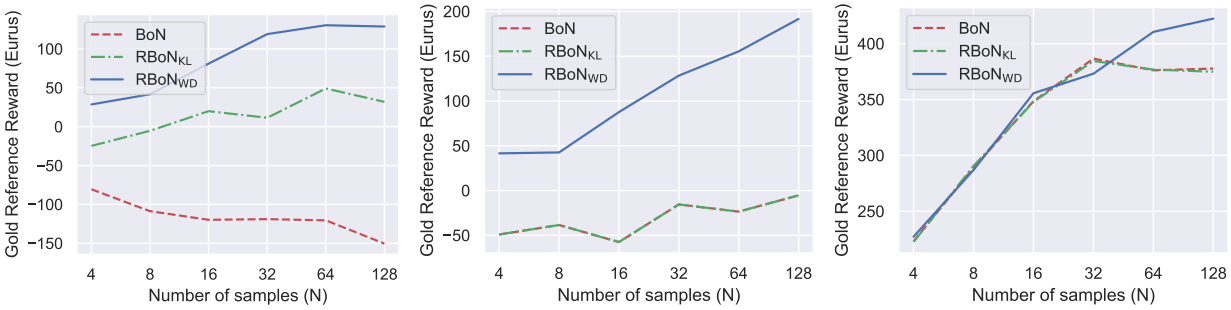
OASST is used as the proxy reward model and Eurus is used for evaluation (Köpf et al., 2023; Yuan et al., 2024a). We train a model with DPO using Low-Rank Adaptation (LoRA) (Hu et al., 2022; Sidahmed et al., 2024). The trained models are evaluated using the evaluation split of the AlpacaFarm dataset. For the evaluation, we generate a response from the trained model by a nucleus sampling with $p = 0.9$ (Holtzman et al., 2020). Other hyperparameters are described in Appendix D.

Figure 7 shows the performance of models trained using RBoN$_{WD}$ and BoN to generate a pairwise preference dataset. The models trained with RBoN$_{WD}$ outperform a model trained with BoN. The result shows the potential of RBoN$_{WD}$ as a tool for generating pairwise preference datasets for preference learning.



(a) Proxy: SHP-Large (b) Proxy: SHP-XL (c) Proxy: OASST

Figure 5: Evaluation of the RBoN using Mistral on the Helpfulness subset of the hh-rlhf dataset. The proxy rewards are SHP-Large, SHP-XL, and OASST. The gold reference reward is Eurus. Note that the RBoN$_{KL}$ is overlapping with BoN when $\beta = 0$ is chosen as the optimal β for RBoN$_{KL}$ (Table 2).



(a) Proxy: SHP-Large (b) Proxy: SHP-XL (c) Proxy: OASST

Figure 6: Evaluation of the RBoN using Mistral on the Harmlessness subset of the hh-rlhf dataset. The proxy rewards are SHP-Large, SHP-XL, and OASST. The gold reference reward is Eurus.

5. Related Work

Using proximity regularization is not the only way to mitigate the reward hacking problem. Several studies have explored the use of multiple rewards. (Coste et al., 2024; Ramé et al., 2024) propose to ensemble multiple reward functions to mitigate reward hacking. Several studies have investigated training models by the reward functions and combining by interpolating the parameters (Ramé et al., 2023; Jang et al., 2023) or ensembling the model (Mitchell et al., 2024). Our approach is applicable to any proxy reward model, so it can be combined with these methods.

The proximity term of RBoN$_{WD}$ is closely related to the objective function of the Minimum Bayes Risk (MBR) decoding (Kumar & Byrne, 2002; 2004). MBR decoding is a decoding strategy that selects the sequence with the highest similarity to the sequences generated by the probability model. It has been shown to produce high-quality outputs in many text generation tasks, including machine translation, text summarization, and image captioning (Freitag et al., 2023; Jinnai et al., 2024; Suzgun et al., 2023; Bertsch et al., 2023). Li et al. (2024) show that the MBR strategy outperforms sophisticated methods such as chain-of-thought (Wei et al., 2022; Wang et al., 2023).[1] The novelty of our work is to introduce the MBR objective to BoN sampling for language model alignment.

6. Conclusions

We propose RBoN sampling, a variant of BoN sampling with proximity regularization, to mitigate the reward hacking problem. We propose two instances of RBoN, RBoN$_{KL}$ and RBoN$_{WD}$, and evaluate their performance using the AlpacaFarm and Anthropic's hh-rlhf dataset.

RBoN$_{KL}$ simply introduces the KL regularization commonly used in the RLHF and DPO families to the BoN sampling. However, its improvement over BoN is limited, showing that simply introducing the widely used KL term to the decoding-time objective is not sufficient to improve the BoN sampling.

[1] MBR is named Sampling-and-voting in Li et al. (2024).

Table 2: The value of the regularization strength β used by RBoN_{KL} and RBoN_{WD}. Note that $\beta = 0$ indicates that it does not improve over BoN, thus recovers BoN.

		SHP-Large		SHP-XL		OASST	
Model	Dataset	KL	WD	KL	WD	KL	WD
Mistral	AlpacaFarm	10^{-6}	0.5	0	0.5	$5 \cdot 10^{-6}$	20.0
Dolly	AlpacaFarm	$2 \cdot 10^{-6}$	0.5	0	1.0	10^{-4}	5.0
Mistral	Helpfulness	0	0.05	0	0.1	0	20.0
Mistral	Harmlessness	10^{-5}	2.0	0	2.0	$5 \cdot 10^{-6}$	20.0

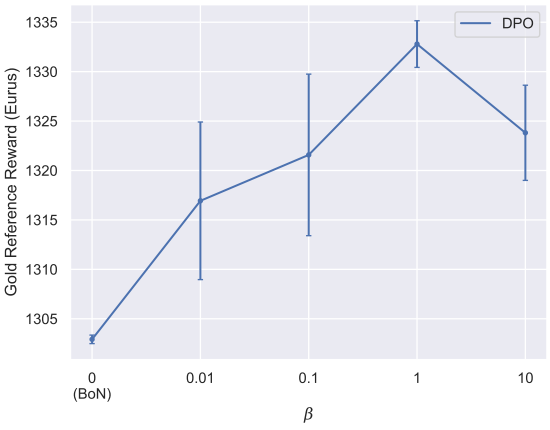


Figure 7: Evaluation of the DPO using RBoN_{WD} to generate the preference dataset. OASST is used as the proxy reward model to generate the preference dataset, and Eurus is used as the gold reference reward. The point represents the average over three runs and the error bar shows the standard error of the mean.

RBoN_{WD} uses Wasserstein distance as proximity regularization. Since Wasserstein distance is an effective optimization objective on its own, it is relatively robust to the choice of β, resulting in significant improvement over BoN in a wide range of settings.

As an application of the decoding strategies, we evaluate the performance of RBoN_{WD} for generating a pairwise preference dataset for preference learning. We generate a preference dataset using RBoN_{WD} and train a DPO model using the generated dataset. The experimental result shows that DPO models trained using a preference dataset generated by RBoN_{WD} outperform a DPO model trained on a dataset generated by BoN.

The empirical result shows that RBoN is an effective decoding time alignment method and is also useful for generating a dataset for a preference learning method. We believe that RBoN will be a practical choice for future decoding-time alignment methods because of its applicability and performance improvements.

7. Limitations

The drawback of the proposed method is that it requires a development set to tune the hyperparameter. Given that there is no clear strategy to pick the β parameter even for RLHF and DPO, we speculate that it would be challenging to develop a strategy to find an effective β automatically. Still, the hyperparameter tuning of RBoN_{WD} is much more computation efficient than that of RLHF and DPO as it does not involve any training procedures.

We use automated evaluation metrics to evaluate the models. Although we use one of the most accurate publicly available reward models (Eurus) to evaluate the performance of the models (Yuan et al., 2024a; Lambert et al., 2024), it would be desirable to perform a human evaluation.

Our experiments on preference learning are limited to the evaluation of DPO. Evaluation of RBoN_{WD} for other preference optimization algorithms is future work (Azar et al., 2023; Liu et al., 2024; Ethayarajh et al., 2024; Xu et al., 2024; Morimura et al., 2024; Hong et al., 2024; Meng et al., 2024; Park et al., 2024).

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions. Kaito Ariu's research is supported by JSPS KAKENHI Grant No. 23K19986.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T.,

El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Beirami, A., Agarwal, A., Berant, J., D'Amour, A., Eisenstein, J., Nagpal, C., and Suresh, A. T. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*, 2024.

Bertsch, A., Xie, A., Neubig, G., and Gormley, M. It's MBR all the way down: Modern generation techniques through the lens of minimum Bayes risk. In Elazar, Y., Ettinger, A., Kassner, N., Ruder, S., and A. Smith, N. (eds.), *Proceedings of the Big Picture Workshop*, pp. 108–122, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bigpicture-1.9. URL https://aclanthology.org/2023.bigpicture-1.9.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T. T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Biyik, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification.

Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., and Xin, R. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Coste, T., Anwar, U., Kirk, R., and Krueger, D. Reward model ensembles help mitigate overoptimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dcjtMYkpXx.

Dubois, Y., Li, C. X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P. S., and Hashimoto, T. B. Alpacafarm: A simulation framework for methods that learn from human feedback. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 30039–30069. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5fc47800ee5b30b8777fdd30abcaaf3b-Paper-Conference.pdf.

Eikema, B. and Aziz, W. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.398. URL https://aclanthology.org/2020.coling-main.398.

Eikema, B. and Aziz, W. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10978–10993, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.754. URL https://aclanthology.org/2022.emnlp-main.754.

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

Ethayarajh, K., Choi, Y., and Swayamdipta, S. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Freitag, M., Ghorbani, B., and Fernandes, P. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9198–9209, Singapore, December

2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.617. URL https://aclanthology.org/2023.findings-emnlp.617.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., and Zhou, T. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2024.

Gui, L., Gârbacea, C., and Veitch, V. BoNBoN alignment for large language models and the sweetness of best-of-n sampling. *arXiv preprint arXiv:2406.00832*, 2024.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Hong, J., Lee, N., and Thorne, J. ORPO: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jang, J., Kim, S., Lin, B. Y., Wang, Y., Hessel, J., Zettlemoyer, L., Hajishirzi, H., Choi, Y., and Ammanabrolu, P. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b. *arXiv*, 2023a.

Jiang, D., Ren, X., and Lin, B. Y. 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)*, 2023b.

Jinnai, Y., Morimura, T., Honda, U., Ariu, K., and Abe, K. Model-based minimum bayes risk decoding. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 21–27 Jul 2024.

Kumar, S. and Byrne, W. Minimum Bayes-risk word alignments of bilingual texts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pp. 140–147. Association for Computational Linguistics, July 2002. doi: 10.3115/1118693.1118712. URL https://aclanthology.org/W02-1019.

Kumar, S. and Byrne, W. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/N04-1022.

Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., and Mattick, A. Openassistant conversations – democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*, 2023.

Lambert, N. and Calandra, R. The alignment ceiling: Objective mismatch in reinforcement learning from human feedback. *arXiv preprint arXiv:2311.00168*, 2024.

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., Smith, N. A., and Hajishirzi, H. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Li, J., Zhang, Q., Yu, Y., Fu, Q., and Ye, D. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229.

Liu, T., Qin, Z., Wu, J., Shen, J., Khalman, M., Joshi, R., Zhao, Y., Saleh, M., Baumgartner, S., Liu, J., Liu, P. J., and Wang, X. LiPO: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*, 2024.

Meng, Y., Xia, M., and Chen, D. SimPO: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

Mitchell, E., Rafailov, R., Sharma, A., Finn, C., and Manning, C. D. An emulator for fine-tuning large language models using small language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Eo7kv0sllr.

Morimura, T., Sakamoto, M., Jinnai, Y., Abe, K., and Ariu, K. Filtered direct preference optimization. *arXiv preprint arXiv:2404.13846*, 2024.

Mudgal, S., Lee, J., Ganapathy, H., Li, Y., Wang, T., Huang, Y., Chen, Z., Cheng, H.-T., Collins, M., Strohman, T., Chen, J., Beutel, A., and Beirami, A. Controlled decoding from language models. *arXiv preprint arXiv:2310.17022*, 2024.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2022.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Pace, A., Mallinson, J., Malmi, E., Krause, S., and Severyn, A. West-of-n: Synthetic preference generation for improved reward modeling. *arXiv preprint arXiv:2401.12086*, 2024.

Pan, A., Bhatia, K., and Steinhardt, J. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JYtwGwIL7ye.

Park, R., Rafailov, R., Ermon, S., and Finn, C. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

Peyré, G. and Cuturi, M. Computational optimal transport. *arXiv preprint arXiv:1803.00567*, 2020.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Ramé, A., Couairon, G., Dancette, C., Gaya, J., Shukor, M., Soulier, L., and Cord, M. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in neural information processing systems*, 2023.

Ramé, A., Vieillard, N., Hussenot, L., Dadashi, R., Cideron, G., Bachem, O., and Ferret, J. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*, 2024.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sidahmed, H., Phatale, S., Hutcheson, A., Lin, Z., Chen, Z., Yu, Z., Jin, J., Komarytsia, R., Ahlheim, C., Zhu, Y., Chaudhary, S., Li, B., Ganesh, S., Byrne, B., Hoffmann, J., Mansoor, H., Li, W., Rastogi, A., and Dixon, L. Perl: Parameter efficient reinforcement learning from human feedback. *arXiv preprint arXiv:2403.10704*, 2024.

Skalse, J. M. V., Howe, N. H. R., Krasheninnikov, D., and Krueger, D. Defining and characterizing reward gaming. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=yb3HOXO3lX2.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 1904.

Stahlberg, F. and Byrne, B. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3356–3362, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1331. URL https://aclanthology.org/D19-1331.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Suzgun, M., Melas-Kyriazi, L., and Jurafsky, D. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4265–4293, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.262. URL https://aclanthology.org/2023.findings-acl.262.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Vamvas, J. and Sennrich, R. Linear-time minimum bayes risk decoding with reference aggregation. *arXiv preprint arXiv:2402.04251*, 2024.

Villani, C. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Wang, C., Jiang, Y., Yang, C., Liu, H., and Chen, Y. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2cRzmWXK9N.

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Durme, B. V., Murray, K., and Kim, Y. J. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024.

Xu, J., Lee, A., Sukhbaatar, S., and Weston, J. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*, 2023.

Yuan, L., Cui, G., Wang, H., Ding, N., Wang, X., Deng, J., Shan, B., Chen, H., Xie, R., Lin, Y., Liu, Z., Zhou, B., Peng, H., Liu, Z., and Sun, M. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*, 2024a.

Yuan, W., Pang, R. Y., Cho, K., Li, X., Sukhbaatar, S., Xu, J., and Weston, J. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024b.

Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M., and Liu, P. J. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Zheng, R., Dou, S., Gao, S., Hua, Y., Shen, W., Wang, B., Liu, Y., Jin, S., Liu, Q., Zhou, Y., Xiong, L., Chen, L., Xi, Z., Xu, N., Lai, W., Zhu, M., Chang, C., Yin, Z., Weng, R., Cheng, W., Huang, H., Sun, T., Yan, H., Gui,

T., Zhang, Q., Qiu, X., and Huang, X. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*, 2023.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2020.

A. Derivation of Equation 9

We show the derivation of Eq. (9). From the definition of Wasserstein distance with $p = 1$ (Peyré & Cuturi, 2020; Villani, 2021), we get the following:

$$WD(\mathbb{1}_y, \hat{\pi}_{\text{ref}}(\cdot|x)) = \min_{\{\mu_{i,j}\}_{i,j} \in \mathcal{J}} \sum_{i=1}^{|\mathcal{Y}|} \sum_{j=1}^{|\mathcal{Y}|} \mu_{i,j} C(y_i, y_j), \tag{14}$$

where $\mathcal{J}$ is a set of all couplings $\{\mu_{i,j}\}_{i,j}$ (Villani, 2021):

$$\mathcal{J} = \Big\{ \{\mu_{i,j}\}_{i,j} : \sum_{i=1}^{|\mathcal{Y}|} \mu_{i,j} = \hat{\pi}_{\text{ref}}(y_j|x), \sum_{j=1}^{|\mathcal{Y}|} \mu_{i,j} = \mathbb{1}_y(y_i), \mu_{i,j} \geq 0 \Big\}. \tag{15}$$

Because $\mathbb{1}_y(y_i) = 0$ for all $y_i \neq y$ and $\mu_{i,j} \geq 0$, we get $\mu_{i,j} = 0$ for all $y_i \neq y$. Thus,

$$(14) = \min_{\mathcal{J}} \sum_{j=1}^{|\mathcal{Y}|} \mu_{y,j} C(y, y_j) \tag{16}$$

Using $\mu_{i,j} = 0$ for all $i \neq y$ and $\sum_{i=1}^{|\mathcal{Y}|} \mu_{i,j} = \hat{\pi}_{\text{ref}}(y_j|x)$, we get $\mu_{y,j} = \hat{\pi}_{\text{ref}}(y_j|x)$. Thus,

$$(16) = \min_{\mathcal{J}} \sum_{j=1}^{|\mathcal{Y}|} \hat{\pi}_{\text{ref}}(y_j|x) C(y, y_j) \tag{17}$$

$$= \sum_{j=1}^{|\mathcal{Y}|} \hat{\pi}_{\text{ref}}(y_j|x) C(y, y_j). \tag{18}$$

Because $\hat{\pi}_{\text{ref}}(y_j|x)$ is an empirical distribution from the set of samples Y_{ref}, $\hat{\pi}_{\text{ref}}(y_j|x) = \sum_{y' \in Y_{\text{ref}}} \frac{1}{N} \mathbb{1}_y(y')$. Thus,

$$(18) = \sum_{y' \in Y_{\text{ref}}} \frac{1}{N} C(y, y'). \tag{19}$$

Thus, we get Eq. (9).

B. Effect of the Regularization Strength

To understand the effect of the regularization strength on the performance of RBoN under different pairs of proxy and gold reward models, we evaluate RBoN using SHP-Large, SHP-XL, OASST, and PairRM (Jiang et al., 2023b) as gold reward models. Figure 16 reports the average Spearman's rank correlation coefficient ρ of a pair of reward models (Spearman, 1904). Note that SHP-Large and SHP-XL reward models are highly correlated as they are trained on the same training procedure.

We perform the sampling using one of the reward models as the proxy reward model and evaluate the selected responses using the remaining reward models as the gold reference rewards. We do not use PairRM as a proxy reward model because it is a pairwise reward model that estimates the preference for a pair of responses rather than computing an absolute preference for a response. The use of a pairwise reward model as a proxy reward model for RBoN is future work.

Figure 8 shows the performance of BoN, RBoN_{KL} with the best β, and RBoN_{WD} with varying β with $N = 128$ using Mistral on the AlpacaFarm dataset. See Appendix C for results using other N. RBoN_{WD} outperforms BoN in all settings except when the proxy reward model is highly correlated with the gold reward model (SHP-Large and SHP-XL). Reward hacking isn't a problem when the proxy reward model is a highly accurate representation of the true reward, so BoN ($\beta = 0$) achieves the best performance. RBoN_{KL} also outperforms BoN when the proxy reward has a low correlation to the gold reference reward. Figure 9 shows the performance of RBoN_{KL} with varying β. RBoN_{KL} with correctly chosen β outperforms BoN. However, the performance of RBoN_{KL} significantly drops when choosing β too large and is sensitive to the choice of β.

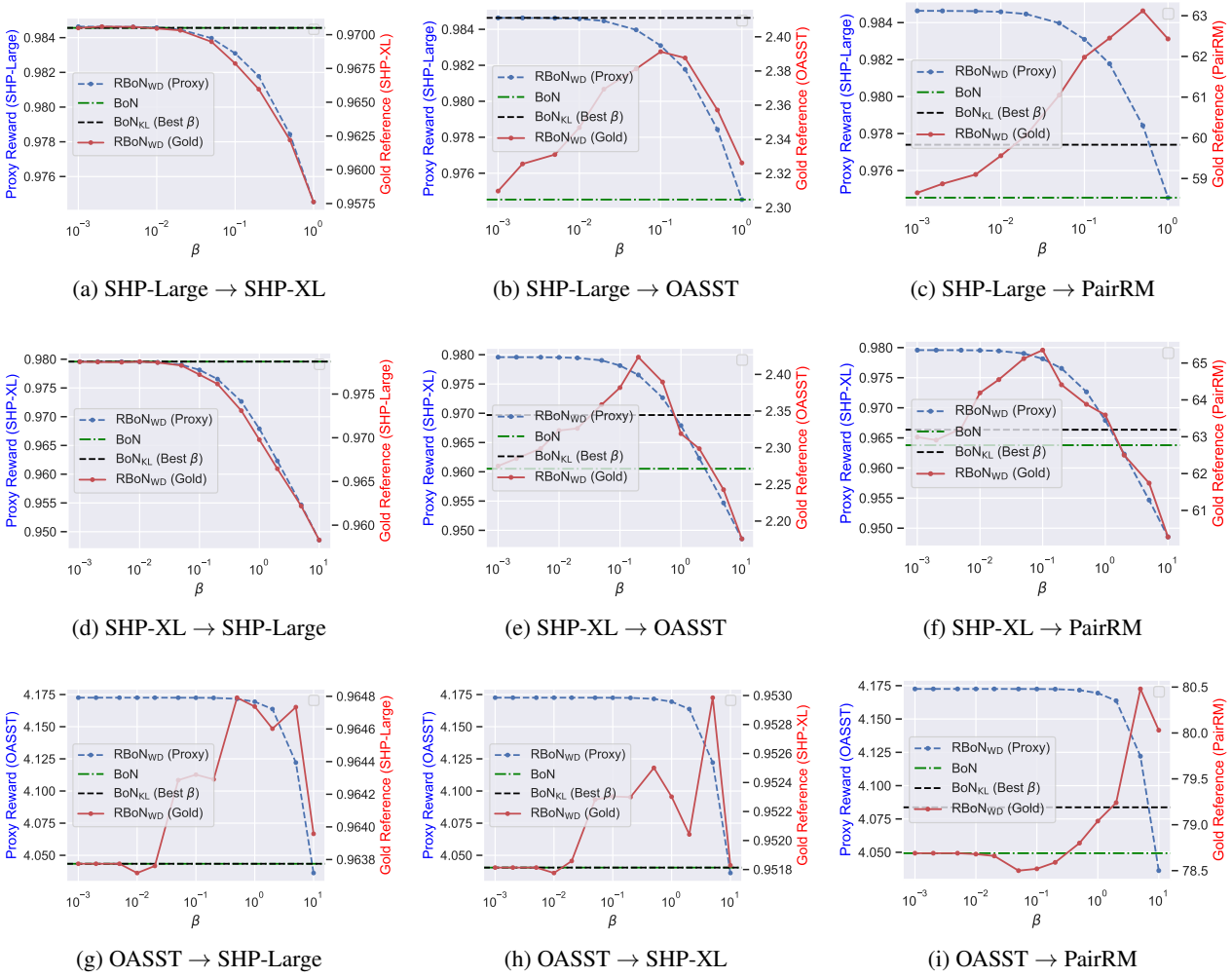
Figure 10 shows the result using Dolly on the AlpacaFarm dataset. Overall, we observe qualitatively the same result as on Mistral. RBoN_{WD} outperforms BoN in a wide range of β when the proxy and gold rewards are not a pair of SHP-Large and SHP-XL.

Figures 12, 13, 14, and 15 show the performance of RBoN_{WD} and RBoN_{KL} on the hh-rlhf dataset. We observe RBoN_{WD} to improve upon BoN in both Helpfulness and Harmlessness subsets. RBoN_{KL} successfully improves upon BoN in the Harmless dataset, but is less effective in the Helpfulness dataset. The result suggests that the KL-divergence term is useful for mitigating harmful behavior but not effective in being helpful.

The experiment shows that the optimal β depends on various factors, but the strength of the correlation between the proxy reward model and the gold reference reward seems to be the key factor. For example, SHP-Large is strongly correlated with SHP-XL ($\rho = 0.66$), so the optimal β is close to 0. In this case, RBoN has little to no advantage over BoN. On the other hand, SHP-Large is only weakly correlated with OASST and PairRM ($\rho = 0.29, 0.20$), where the optimal β for SHP-Large $\rightarrow$ OASST and PairRM is large ($\beta = 0.1 - 1.0$).

C. Additional Figures

Figures 17 and 18 show the performance of BoN ($\beta = 0$), RBoN_{WD}, and MBR decoding ($\beta = +\infty$) with different number of samples N using Mistral and Dolly. We observe qualitatively similar results with smaller N to the result of



(a) SHP-Large → SHP-XL (b) SHP-Large → OASST (c) SHP-Large → PairRM

(d) SHP-XL → SHP-Large (e) SHP-XL → OASST (f) SHP-XL → PairRM

(g) OASST → SHP-Large (h) OASST → SHP-XL (i) OASST → PairRM

Figure 8: The gold reward score and the proxy reward score of the RBoN$_{WD}$ with different regularization strengths and reward models. The captions of the subfigures show the proxy and the gold reward model (Proxy → Gold). The performance of RBoN$_{KL}$ with the best β hyperparameter and BoN is shown in the horizontal lines. The responses are generated by Mistral. The number of samples N is 128.

$N = 128$ in Figures 8 and 10.

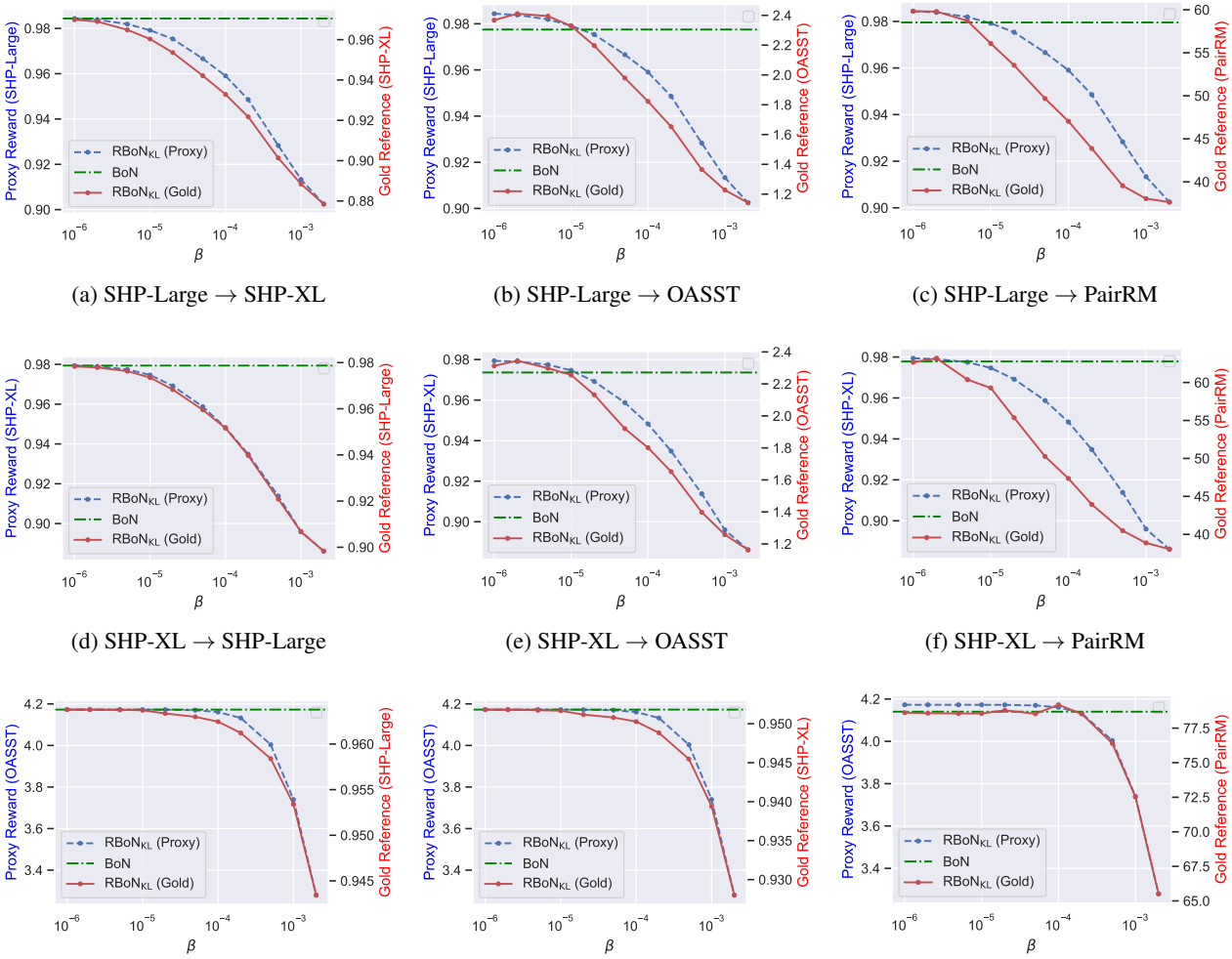
Figure 11 is the result of the RBoN$_{KL}$ with varying β using Dolly. Overall, we observe qualitatively the same results as in Mistral. RBoN$_{KL}$ outperforms BoN except when the proxy reward model is highly correlated with the gold reference reward. The performance of RBoN$_{KL}$ is sensitive to the choice of β.

D. Hyperparameters

Table 3 describes the hyperparameters used to generate responses from the π_{ref}. The parameters are used for both Sections 4.1 and 4.2. Table 4 summarizes the hyperparameters used for training the DPO model in Section 4.2.

Table 3: Generation hyperparameters used in Section 4.1 and 4.2

Parameter	Value
Max instruction length	256
Max new tokens	256
Temperature	1.0
Top-p	0.9



(a) SHP-Large → SHP-XL (b) SHP-Large → OASST (c) SHP-Large → PairRM

(d) SHP-XL → SHP-Large (e) SHP-XL → OASST (f) SHP-XL → PairRM

(g) OASST → SHP-Large (h) OASST → SHP-XL (i) OASST → PairRM

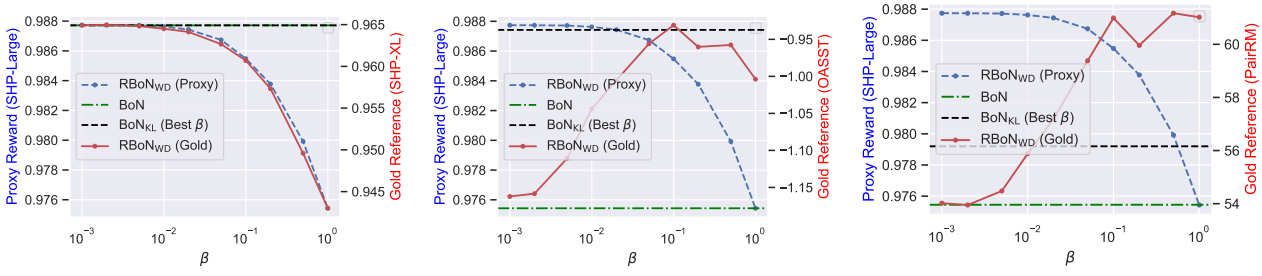
Figure 9: Evaluation of the RBoN$_{KL}$. The gold reward score and the proxy reward score of the RBoN$_{WD}$ with different regularization strengths and reward models are shown. The captions of the subfigures show the proxy and the gold reward model (Proxy → Gold). The responses are generated by Mistral. The number of samples N is 128.

Table 4: DPO hyperparameters used in Section 4.2.

Parameter	Value
Epochs	3
Learning rate	1e-5
Optimizer	AdamW
Batch size	4
Regularization factor (β)	0.1
LoRA r	128
LoRA α	32

E. Reproducibility Statement

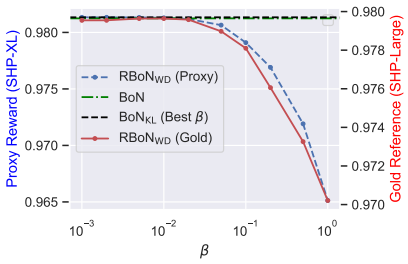
All datasets and models used in the experiments are open source (Table 5). Our code is available at https://github.com/CyberAgentAILab/regularized-bon.



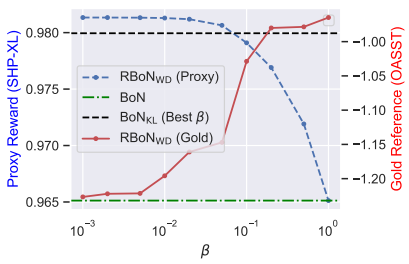
(a) SHP-Large → SHP-XL

(b) SHP-Large → OASST

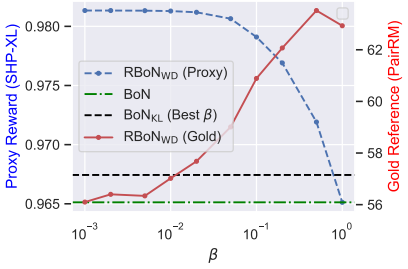
(c) SHP-Large → PairRM



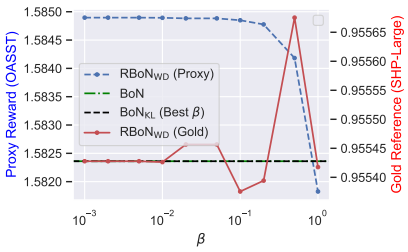
(d) SHP-XL → SHP-Large



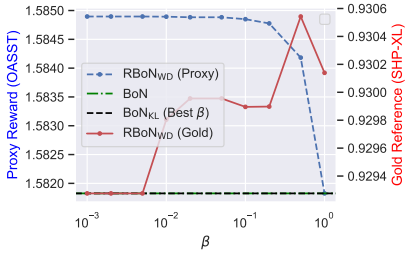
(e) SHP-XL → OASST



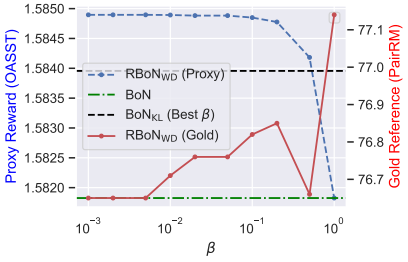
(f) SHP-XL → PairRM



(g) OASST → SHP-Large

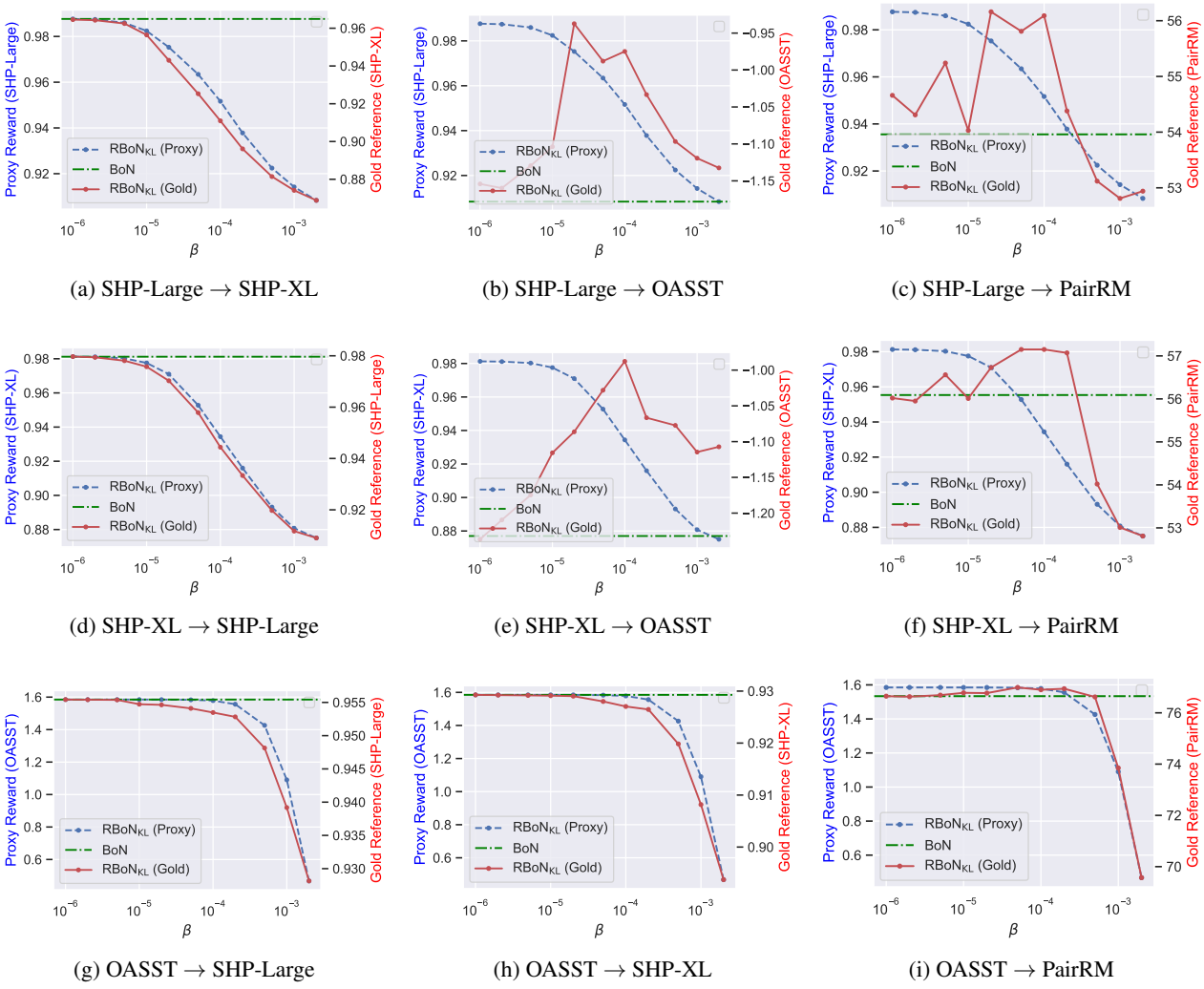


(h) OASST → SHP-XL



(i) OASST → PairRM

Figure 10: Evaluation using Dolly on AlpacaFarm. The gold reward score and the proxy reward score of the RBoN$_{WD}$ with different regularization strengths and reward models. The number of samples N is 128.

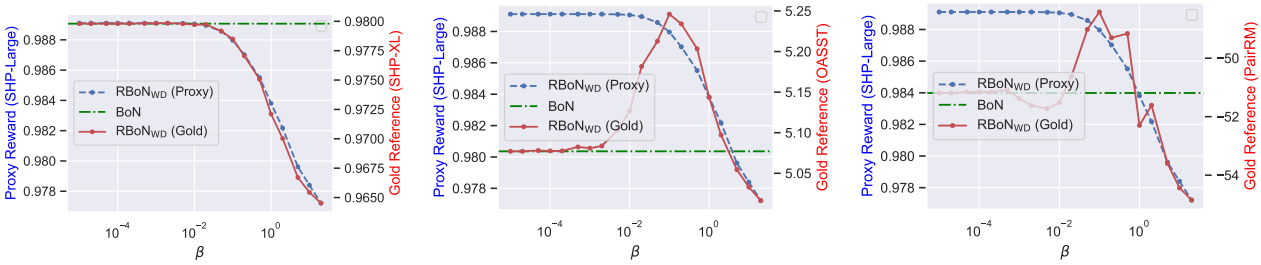


(a) SHP-Large → SHP-XL (b) SHP-Large → OASST (c) SHP-Large → PairRM

(d) SHP-XL → SHP-Large (e) SHP-XL → OASST (f) SHP-XL → PairRM

(g) OASST → SHP-Large (h) OASST → SHP-XL (i) OASST → PairRM

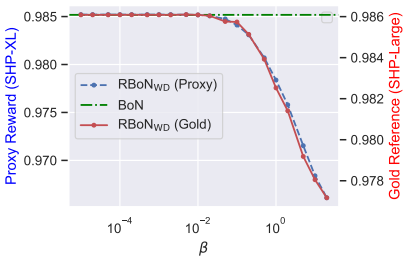
Figure 11: Evaluation of the RBoN$_{KL}$. The gold reward score and the proxy reward score of the RBoN$_{KL}$ with different regularization strengths and reward models. The captions of the subfigures show the proxy and the gold reward model (Proxy → Gold). The responses are generated by Dolly. The number of samples N is 128.



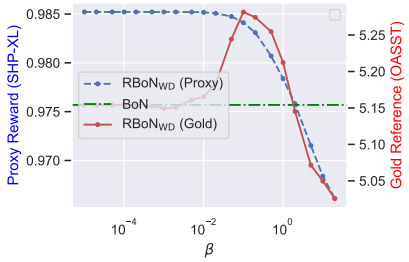
(a) SHP-Large → SHP-XL

(b) SHP-Large → OASST

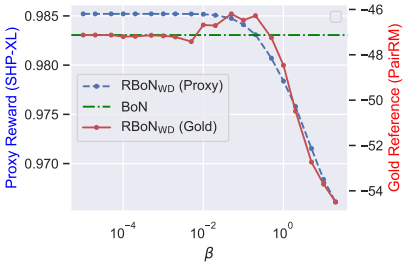
(c) SHP-Large → PairRM



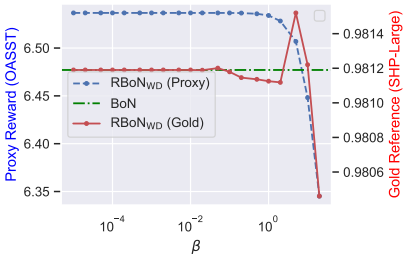
(d) SHP-XL → SHP-Large



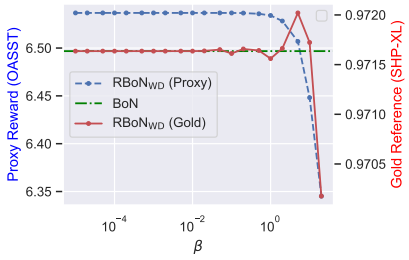
(e) SHP-XL → OASST



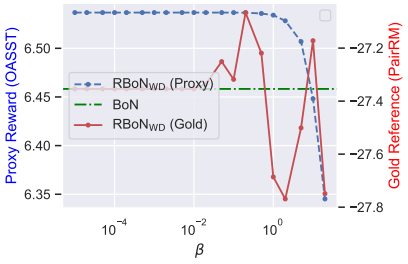
(f) SHP-XL → PairRM



(g) OASST → SHP-Large

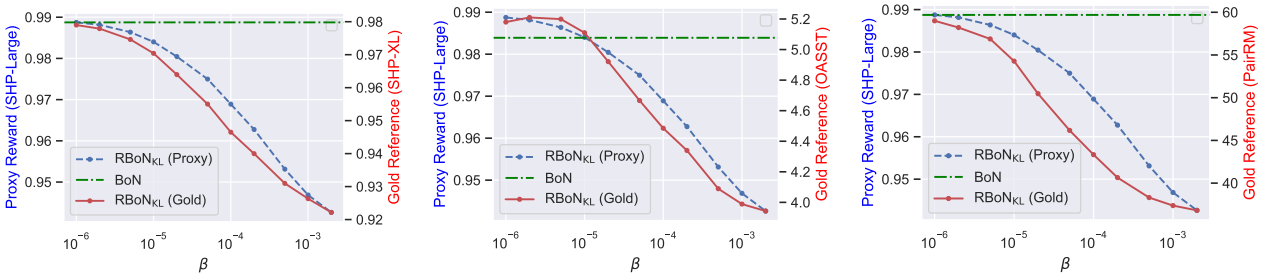


(h) OASST → SHP-XL



(i) OASST → PairRM

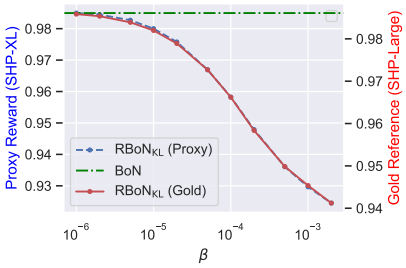
Figure 12: Evaluation on the Anthropic's Helpfulness dataset. The gold reward score and the proxy reward score of the RBoN_{WD} with different regularization strengths and reward models. The captions of the subfigures show the proxy and the gold reward model (Proxy → Gold). The performance of RBoN_{KL} with the best β hyperparameter and BoN is shown in the horizontal lines. The responses are generated by Mistral. The number of samples N is 128.



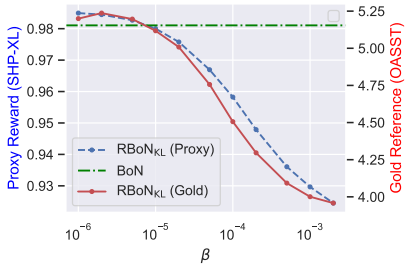
(a) SHP-Large → SHP-XL

(b) SHP-Large → OASST

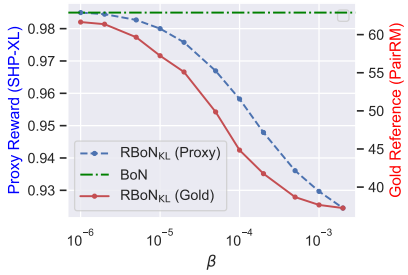
(c) SHP-Large → PairRM



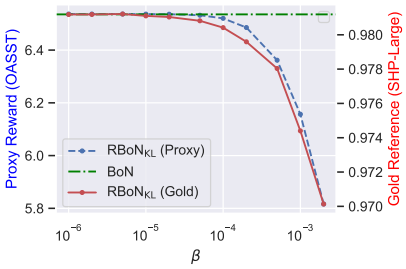
(d) SHP-XL → SHP-Large



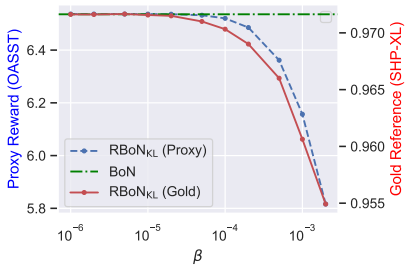
(e) SHP-XL → OASST



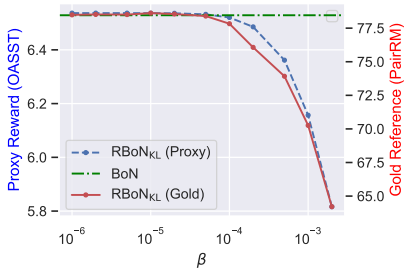
(f) SHP-XL → PairRM



(g) OASST → SHP-Large

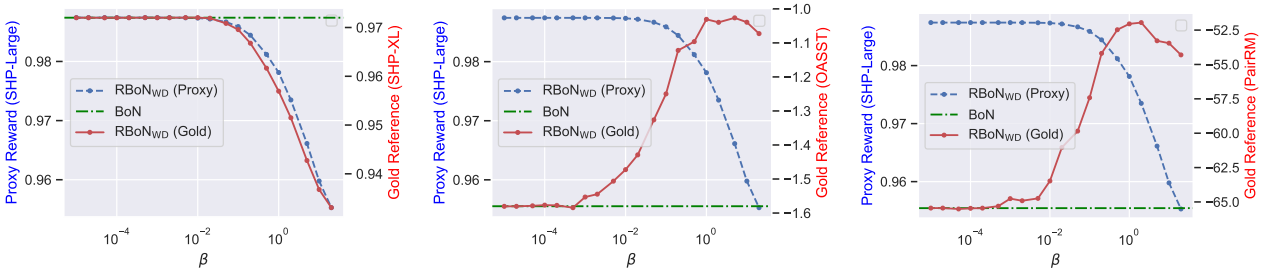


(h) OASST → SHP-XL



(i) OASST → PairRM

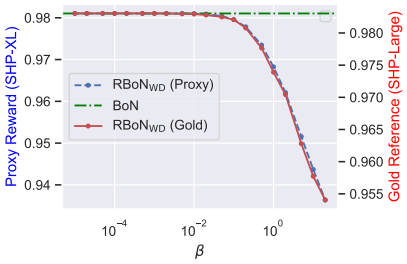
Figure 13: Evaluation of the RBoN$_{KL}$ on the Anthropic's Helpfulness dataset. The gold reward score and the proxy reward score of the RBoN$_{WD}$ with different regularization strengths and reward models are shown. The captions of the subfigures show the proxy and the gold reward model (Proxy → Gold). The responses are generated by Mistral. The number of samples N is 128.



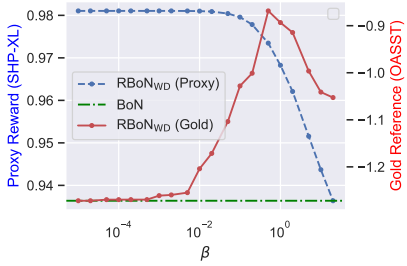
(a) SHP-Large → SHP-XL

(b) SHP-Large → OASST

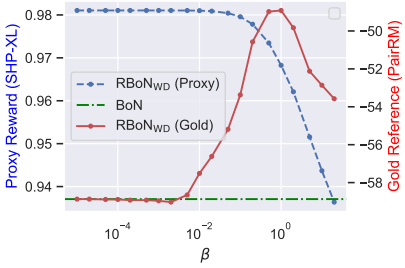
(c) SHP-Large → PairRM



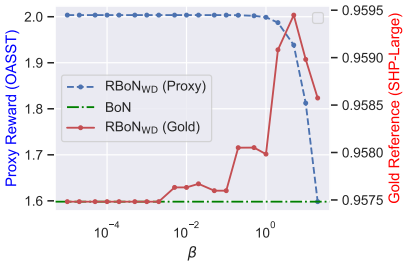
(d) SHP-XL → SHP-Large



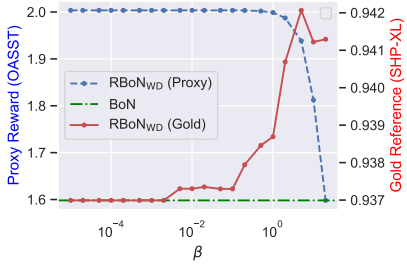
(e) SHP-XL → OASST



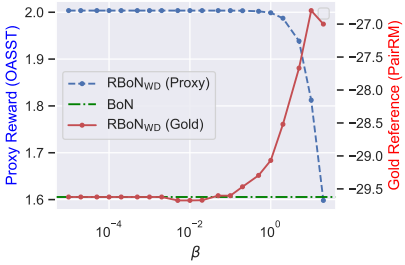
(f) SHP-XL → PairRM



(g) OASST → SHP-Large

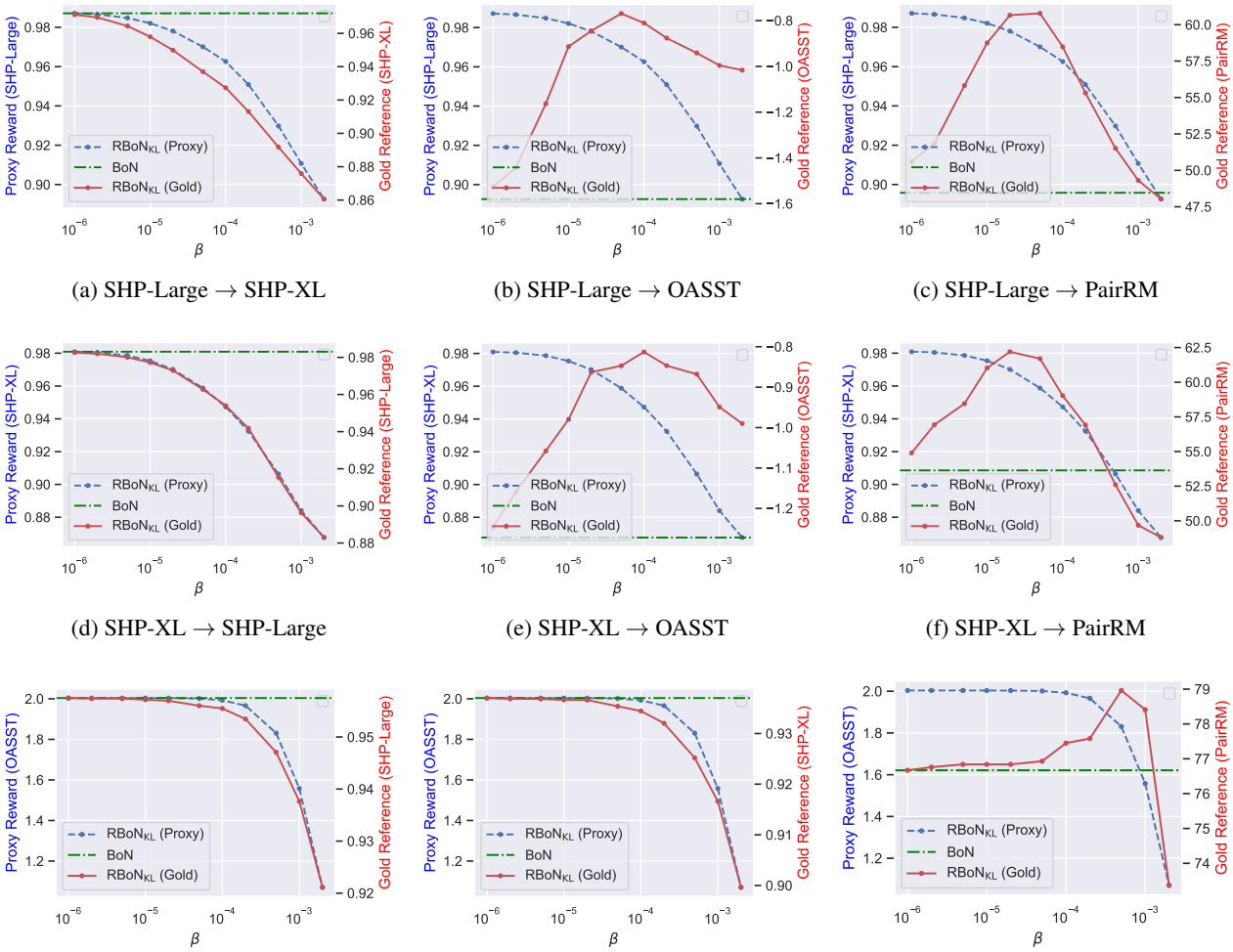


(h) OASST → SHP-XL



(i) OASST → PairRM

Figure 14: Evaluation on the Anthropic's Harmlessness dataset. The gold reward score and the proxy reward score of the RBoN_{WD} with different regularization strengths and reward models. The captions of the subfigures show the proxy and the gold reward model (Proxy → Gold). The performance of RBoN_{KL} with the best β hyperparameter and BoN is shown in the horizontal lines. The responses are generated by Mistral. The number of samples N is 128.



(a) SHP-Large → SHP-XL (b) SHP-Large → OASST (c) SHP-Large → PairRM

(d) SHP-XL → SHP-Large (e) SHP-XL → OASST (f) SHP-XL → PairRM

(g) OASST → SHP-Large (h) OASST → SHP-XL (i) OASST → PairRM

Figure 15: Evaluation of the RBoN_{KL} on the Anthropic's Harmlessness dataset. The gold reward score and the proxy reward score of the RBoN_{WD} with different regularization strengths and reward models are shown. The captions of the subfigures show the proxy and the gold reward model (Proxy → Gold). The responses are generated by Mistral. The number of samples N is 128.

Table 5: List of datasets and models used in the experiments.

Name	Reference
AlpacaFarm	(Dubois et al., 2023) https://huggingface.co/datasets/tatsu-lab/alpaca_farm
Anthropic's hh-rlhf	(Bai et al., 2022) https://huggingface.co/datasets/Anthropic/hh-rlhf
mistral-7b-sft-beta (Mistral)	(Jiang et al., 2023a; Tunstall et al., 2023) https://huggingface.co/HuggingFaceH4/mistral-7b-sft-beta
dolly-v2-3b (Dolly)	(Conover et al., 2023) https://huggingface.co/databricks/dolly-v2-3b
SHP-Large	(Ethayarajh et al., 2022) https://huggingface.co/stanfordnlp/SteamSHP-flan-t5-large
SHP-XL	(Ethayarajh et al., 2022) https://huggingface.co/stanfordnlp/SteamSHP-flan-t5-xl
OASST	(Köpf et al., 2023) https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2
PairRM	(Jiang et al., 2023b) https://huggingface.co/llm-blender/PairRM
Eurus	(Yuan et al., 2024a) https://huggingface.co/openbmb/Eurus-RM-7b
MPNet	(Song et al., 2020) https://huggingface.co/sentence-transformers/all-mpnet-base-v2

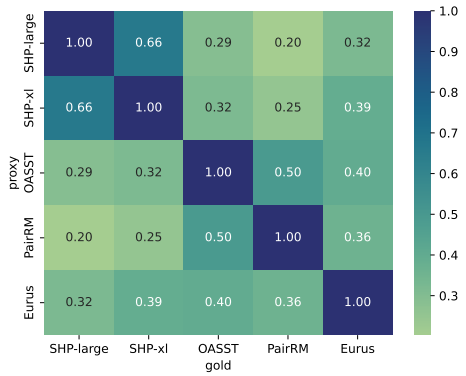
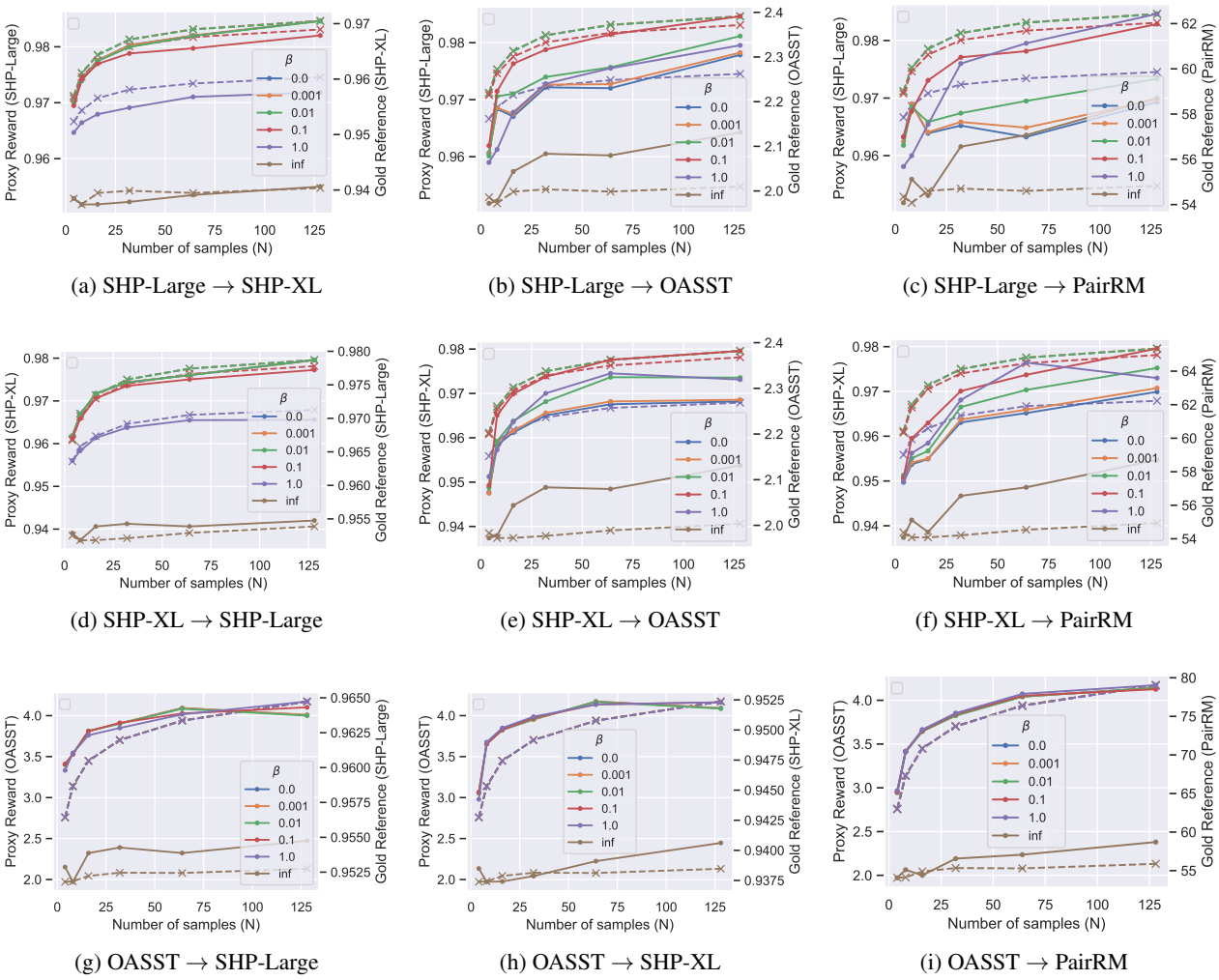


Figure 16: Average Spearman's rank correlation coefficient of the reward models in the evaluation split of the AlpacaFarm dataset for the responses generated by Mistral. 128 responses are used to compute Spearman's rank correlation for each instruction, averaged over the 805 instructions.

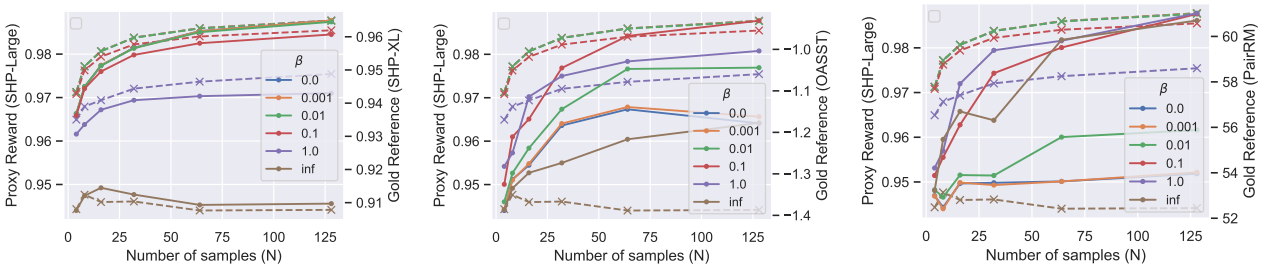


(a) SHP-Large → SHP-XL (b) SHP-Large → OASST (c) SHP-Large → PairRM

(d) SHP-XL → SHP-Large (e) SHP-XL → OASST (f) SHP-XL → PairRM

(g) OASST → SHP-Large (h) OASST → SHP-XL (i) OASST → PairRM

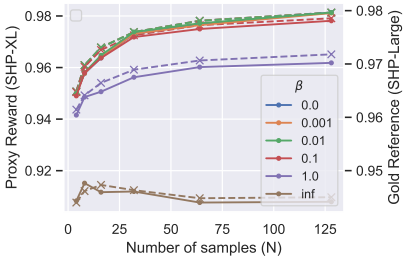
Figure 17: Evaluation of RBoN using Mistral on AlpacaFarm. The gold reward score and the proxy reward score of the RBoN_{WD} with different regularization strengths and reward models. The captions of the subfigures show the proxy and the gold reward model (Proxy → Gold).



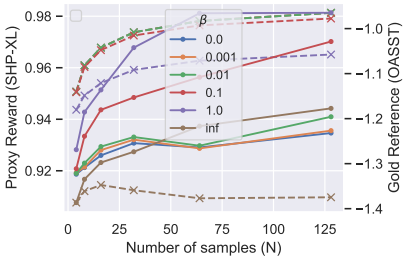
(a) SHP-Large → SHP-XL

(b) SHP-Large → OASST

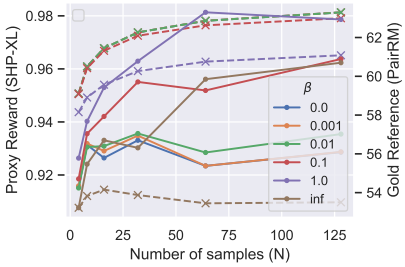
(c) SHP-Large → PairRM



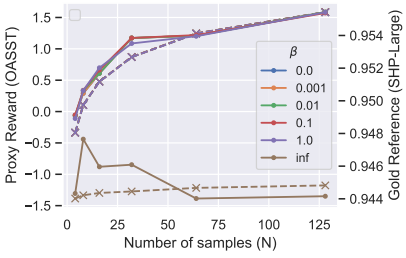
(d) SHP-XL → SHP-Large



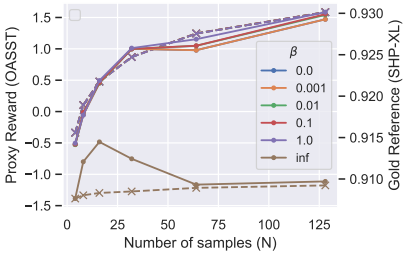
(e) SHP-XL → OASST



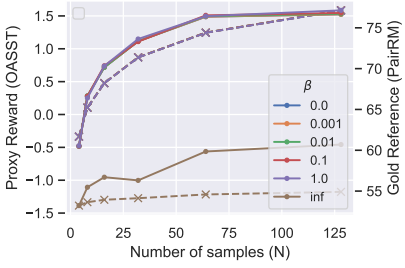
(f) SHP-XL → PairRM



(g) OASST → SHP-Large



(h) OASST → SHP-XL



(i) OASST → PairRM

Figure 18: Evaluation of RBoN using Dolly on AlpacaFarm. The gold reward score and the proxy reward score of the RBoN_{WD} with different regularization strengths and reward models. The captions of the subfigures show the proxy and the gold reward model (Proxy → Gold).