

Annotation-Efficient Language Model Alignment via Diverse and Representative Response Texts

Anonymous ACL submission

Abstract

Preference optimization is a standard approach to fine-tuning large language models to align with human preferences. The quantity, diversity, and representativeness of the preference dataset are critical to the effectiveness of preference optimization. However, obtaining a large amount of preference annotations is difficult in many applications. This raises the question of how to use the limited annotation budget to create an effective preference dataset. To this end, we propose Annotation-Efficient Preference Optimization (AEPO). Instead of exhaustively annotating preference over all available response texts, AEPO selects a subset of responses that maximizes diversity and representativeness from the available responses and then annotates preference over the selected ones. In this way, AEPO focuses the annotation budget on labeling preferences over a smaller but informative subset of responses. We evaluate the performance of Direct Preference Optimization (DPO) using AEPO and show that it outperforms models trained using a standard DPO with the same annotation budget. Our code is available at <https://anonymous.4open.science/r/aepo-05B2>.

1 Introduction

Language model alignment aims to address these issues by guiding Large Language Models (LLMs) to generate responses that aligns with human preferences, steering them to generate responses that are informative, harmless, and helpful (Christiano et al., 2017; Ziegler et al., 2020; Stiennon et al., 2020; Bai et al., 2022; Ouyang et al., 2022; Rafailov et al., 2023). The performance of the alignment algorithms is highly dependent on the choice of the preference dataset. However, building a human preference dataset requires expensive human annotations, which is the major bottleneck for constructing a large and high quality preference dataset.

A large number of works have investigated the

synthesis of preference data using a powerful LLM (e.g., GPT-4) to distill the knowledge of human preferences (Dubois et al., 2023; Lee et al., 2024; Ding et al., 2023; Honovich et al., 2023; Cui et al., 2023; Mukherjee et al., 2023; Xu et al., 2024a; Liu et al., 2024a). However, human preferences are known to be diverse and pluralistic, and they are unlikely to be represented by the opinion of a single model (Qiu et al., 2022; Kirk et al., 2023; Wan et al., 2023; Cao et al., 2023b; Zhou et al., 2024; Sorensen et al., 2024a; Rao et al., 2024; Xu et al., 2024b; Sorensen et al., 2024b; Kirk et al., 2024; Shen et al., 2024a; Chakraborty et al., 2024; Pistilli et al., 2024). Several papers have pointed out that LLMs may exhibit bias toward aligning with people from a particular background (Santurkar et al., 2023; Naous et al., 2024; Adilazuarda et al., 2024). For example, Cao et al. (2023b) reports that ChatGPT has a strong alignment with American culture, but adapts less effectively to other cultural contexts. In addition to cultural biases, previous work suggests that even a highly capable model (e.g., GPT-4) still has biases such as length bias (Jain et al., 2024; Dubois et al., 2024), style bias (Gudibande et al., 2024), and positional bias (Zheng et al., 2023). Thus, human annotation is desirable to align and personalize an LLM with diverse and unbiased human preferences (Greene et al., 2023; Jang et al., 2023; Kirk et al., 2023).

The efficiency of annotation is critical to making LLMs accessible and useful to people from diverse backgrounds, who may have only a small amount of preference feedback data to work with. To this end, we investigate how to generate an effective preference dataset with a limited human annotation budget. Previous work has shown that the following three features are desirable for a preference dataset to be effective (Liu et al., 2024c,a):

1. *Quantity and Diversity of instructions.* Greater quantity and diversity are desirable for

the instruction set (Askell et al., 2021; Wang et al., 2023; Ding et al., 2023; Honovich et al., 2023; Cao et al., 2023a; Yuan et al., 2023; Yu et al., 2023; Xu et al., 2024a; Zhang et al., 2024a; Ge et al., 2024).

2. *Diversity of responses.* A set of responses with higher diversity is desirable (Cui et al., 2023; Lu et al., 2024; Yuan et al., 2023; Song et al., 2024; Wang et al., 2024; Dubey et al., 2024).
3. *Representativeness of responses.* Responses that represent the behavior of the training model are more desirable (Guo et al., 2024; Tajwar et al., 2024; Tang et al., 2024a; Xu et al., 2024d)

To achieve all three desiderata with a limited annotation budget, it is desirable to annotate preference over diverse and representative responses with a minimum amount of annotation per instruction.

To this end, we propose **Annotation-Efficient Preference Optimization (AEPO)**, a preference optimization with a preprocessing step on the preference dataset to reduce the required amount of annotation (Figure 1). Instead of annotating the preference over all N responses, AEPO selects $k(< N)$ responses from N responses. We deploy a sophisticated method to select a set of response texts with high diversity and representativeness. It then annotates the preference for the selected k responses. In this way, AEPO uses all N samples to select a subset of responses with high diversity and representativeness, while requiring only an annotation over a subset of responses.

The strength of AEPO is threefold. First, it is applicable to human feedback data. Compared to Reinforcement Learning from AI Feedback (RLAIF) (Lee et al., 2024), our approach can be applied to both human and AI feedback. RLAIF is a scalable approach in terms of both instructions and annotations, but it is known that the feedback from existing language models is biased in various ways (Cao et al., 2023b; Zheng et al., 2023; Jain et al., 2024; Gudibande et al., 2024; Dubois et al., 2024). Second, it is scalable with additional computational resources. By generating a larger amount of responses, AEPO can find more diverse and representative responses to annotate, resulting in a more effective preference dataset with a fixed amount of annotation (Figure 3). Third, less annotation is required to generate an effective preference dataset.

Method	Human feedback	Scalable	Annotation cost
Human feedback	✓	✗	✗
RLAIF	✗	✓	✓
West-of-N	✓	✓	✗
AEPO	✓	✓	✓

Table 1: Comparison of annotation strategies for preference dataset.

Unlike an exhaustive annotation strategy which requires a large annotation effort (e.g., West-of-N strategy, Xu et al. 2023; Yuan et al. 2024b; Pace et al. 2024), AEPO can reduce the annotation cost through the subsampling process.

We evaluate the performance of DPO using AEPO on the AlpacaFarm, hh-rlhf, and JCommonsMorality datasets in Section 4 (Bai et al., 2022; Dubois et al., 2023; Takeshita et al., 2023). With a fixed annotation budget, the performance of vanilla DPO degrades as the number of responses per instruction increases above a certain threshold (Figure 3). In contrast, AEPO scales with the number of responses under a fixed annotation budget, outperforming vanilla DPO when a large number of responses are available. The result shows that AEPO is a promising algorithm for efficient preference optimization, especially when annotation cost is the bottleneck of the alignment process.

2 Background

Preference Optimization. Let $\mathcal{D}_p$ be a pairwise preference dataset $\mathcal{D}_p = \{(x, y_c, y_r)\}$, where x is an instruction ($x \in \mathcal{X}$), y_c is the chosen response, and y_r is the rejected response, that is, y_c is preferred to y_r ($y_c, y_r \in \mathcal{Y}$). One of the popular algorithms for learning from the preference dataset is **Direct Preference Optimization (DPO)** (Rafailov et al., 2023). DPO trains the language model to directly align with the human preference data over the responses without using reward models. The objective function of the DPO is the following:

$$\pi_{\text{DPO}} = \arg \max_{\pi} \mathbb{E}_{(x, y_c, y_r) \sim \mathcal{D}_p} \left[\log \sigma \left(\beta \log \frac{\pi(y_c|x)}{\pi_{\text{ref}}(y_c|x)} - \beta \log \frac{\pi(y_r|x)}{\pi_{\text{ref}}(y_r|x)} \right) \right], \quad (1)$$

where σ is the sigmoid function and β is a hyperparameter that controls the proximity to the SFT model π_{ref} .

Preference Dataset. The performance of preference optimization largely depends on the choice

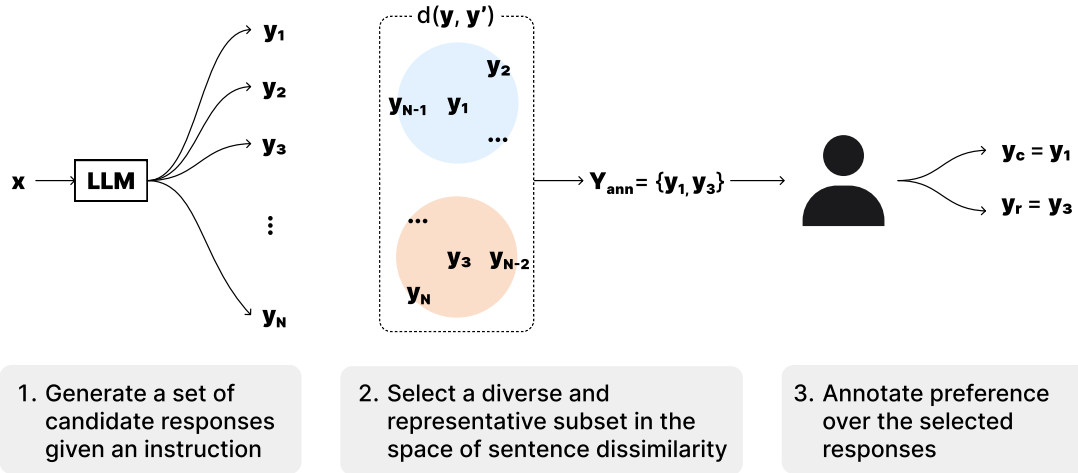


Figure 1: Annotation-Efficient Preference Optimization (AEPO) is a process for generating a preference dataset with diverse and representative responses with fewer annotations. See Section 3 for details. Here we set $k = 2$ and select two responses from the candidate responses to annotate.

of the preference dataset $\mathcal{D}_p$. Aside from synthesizing, several papers have investigated annotation-efficient learning by reducing the number of instructions rather than synthesizing more (Cohn et al., 1994; Settles, 2009; Su et al., 2023; Zhou et al., 2023; Chen et al., 2024).

Regarding the selection of the response texts, several works have proposed to use the **West-of-N (WoN) strategy** (Xu et al., 2023; Yuan et al., 2024b; Pace et al., 2024). The WoN strategy randomly samples N responses $\{y_i\}_{i=1}^N$ for each instruction x . Then, it annotates the preference *over all N responses*. The response with the highest preference is labeled as chosen (win) y_c and the one with the lowest preference is labeled as rejected (lose) y_r to construct $\mathcal{D}_p$:

$$y_c \leftarrow \arg \max_{y \in \{y_i\}_{i=1}^N} R(x, y), \quad y_r \leftarrow \arg \min_{y \in \{y_i\}_{i=1}^N} R(x, y). \quad (2)$$

The strategy is shown to be more efficient than random sampling with the same number of instructions. However, it requires N annotations per instruction to run, making it inapplicable when the annotation budget is limited.

3 Annotation-Efficient Preference Optimization (AEPO)

We propose **Annotation-Efficient Preference Optimization (AEPO)**, a method for efficiently learning preferences from a large number of responses *with a limited budget on preference annotations* (Figure 1).

The procedure of AEPO is described in Algo-

Algorithm 1 Annotation-Efficient Preference Optimization (AEPO)

Input: A set of pairs of an instruction and a set of candidate responses $\mathcal{D} = \{(x, Y_{\text{cand}})\}$, a preference annotator R , and an annotation budget per instruction k

- 1: $\mathcal{D}_{AE} = \emptyset$
 - 2: **for** $(x, Y_{\text{cand}}) \in \mathcal{D}$ **do**
 - 3: $Y_{\text{ann}} \leftarrow \arg \max_{Y \subseteq Y_{\text{cand}}, |Y|=k} f_{\text{rep}}(Y) + \lambda f_{\text{div}}(Y)$ (See Eq. 10)
 - 4: $y_c \leftarrow \arg \max_{y \in Y_{\text{ann}}} R(x, y)$
 - 5: $y_r \leftarrow \arg \min_{y \in Y_{\text{ann}}} R(x, y)$
 - 6: $\mathcal{D}_{AE} \leftarrow \mathcal{D}_{AE} \cup \{(x, y_c, y_r)\}$
 - 7: **end for**
 - 8: **return** $\mathcal{D}_{AE}$
-

rithm 1. We assume that a set of N candidate responses is available for each instruction: $\mathcal{D} = \{(x, Y_{\text{cand}})\}$, where $Y_{\text{cand}} := \{y_i\}_{i=1}^N$. Instead of annotating the preference over all responses in Y_{cand} , AEPO subsamples k responses (e.g., $k = 2$) from Y_{cand} according to the objective function (Eq. 10) that heuristically maximizes the accuracy of reward value estimates (line 3). We explain the objective function later. Then, it deploys the WoN strategy (Eq. 2) on the subsampled subset of responses Y_{ann} instead of all N responses Y_{cand} . It annotates the preference over Y_{ann} to select the best and the worst responses as the chosen and the rejected responses, respectively (lines 4, 5). In this way, we can allocate the annotation budget only to labeling k responses. AEPO achieves to build a

preference dataset with diverse and representative responses using a small amount of annotation effort, which is exactly the characteristics desired for the preference annotation methodology we discussed in Section 1.

Choosing the Responses to Annotate Preference (Line 3 in Algorithm 1). The performance of the procedure is highly dependent on how we subsample a subset Y_{ann} from the candidate set of responses Y_{cand} . Ideally, one wants to identify Y_{ann} that the reward values $R(x, y)$ of $y \in Y_{\text{cand}}$ can be estimated using the preference annotation over Y .

We deploy two heuristics to derive the objective function for choosing Y_{ann} . Let $d(y_1, y_2)$ be a cost function that represents the dissimilarity of the two response texts: $D : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$, where $d(y_1, y_2) = 0$ if $y_1 = y_2$. For example, d can be implemented by the cosine distance of the embedding of the sentences:

$$d(y_1, y_2) = 1 - \cos(\text{emb}(y_1), \text{emb}(y_2)), \quad (3)$$

where $\cos$ is the cosine function and emb is the sentence embedding function.

3.1 Representativeness Heuristic

Heuristic 1 *The preference annotation over Y_{ann} is more likely to be informative in estimating $R(x, y)$ if Y_{ann} is closer to y . That is, if*

$$\sum_{y_i \in Y} d(y, y_i) \leq \sum_{y_i \in Y'} d(y, y_i), \quad (4)$$

then the estimate of $R(x, y)$ is more likely to be accurate using Y than Y' .

Figure 2 illustrates the intuition behind the heuristic. Intuitively, similar texts are more likely to have similar reward values. Thus, knowing the preference values of y_1 and y_2 are likely more informative than those of y'_1 and y'_2 to estimate the reward value of y .

From Eq. 4, we are motivated to choose a subset Y_{ann} so that they are closer to y to estimate $R(x, y)$. Thus, to find Y_{ann} closer to all $y \in Y_{\text{cand}}$, the objective is to maximize the following:

$$f_{\text{rep}}(Y) := - \sum_{y \in Y_{\text{cand}}} f_{\text{rep}}(y; Y), \quad \text{where} \quad (5)$$

$$f_{\text{rep}}(Y; y) := - \frac{1}{N} \sum_{y_i \in Y} d(y, y_i).$$

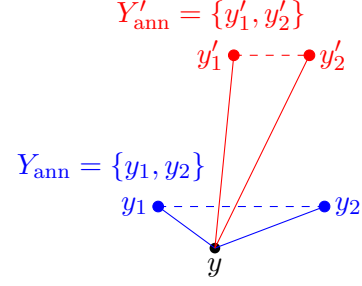


Figure 2: An illustrative example of response subsets for annotating preference. Our algorithm is based on the heuristic that the subset Y that is more diverse and closer to y is more likely to be informative than Y' to infer the value of y .

An alternative explanation of $f_{\text{rep}}(Y)$ is that it quantifies the representativeness of the subset Y for the entire sample set Y_{cand} .

$$f_{\text{rep}}(Y) = \sum_{y \in Y} \underbrace{\left(-\frac{1}{N} \sum_{y' \in Y_{\text{cand}} \setminus \{y\}} d(y, y') \right)}_{\text{Representativeness of } y} \quad (6)$$

where the bracketed term can be interpreted as the representativeness of y , the average distance from y to all other samples. That is, it shows the closeness to the mean of the sample set. Thus, $f_{\text{rep}}(Y)$ represents the objective to select a subset Y that is closer to the center of the samples, making it more representative of the generated samples.

3.2 Diversity Heuristic

Heuristic 2 *The preference over Y_{ann} is likely to be more informative in estimating $R(x, y)$ if each pair of samples in Y_{ann} is more distinct. That is, if*

$$\sum_{y_1 \in Y} \sum_{y_2 \in Y \setminus \{y_1\}} d(y_1, y_2) \geq \sum_{y_1 \in Y'} \sum_{y_2 \in Y' \setminus \{y_1\}} d(y_1, y_2), \quad (7)$$

then, the estimate of $R(x, y)$ is more likely to be accurate using Y than Y' .

An example of high and low diversity subsamples (Y_{ann} and Y'_{ann}) is shown in Figure 2. If the selected samples are too similar (e.g., Y'_{ann}), then it will be difficult to estimate $R(x, y)$ as y is roughly as similar to y'_1 as to y'_2 . On the other hand, if the selected samples are distinct enough (e.g., Y_{ann}), then we expect it to be easier to estimate the value of $R(x, y)$. Thus, we assume that it is difficult to estimate the value of $R(x, y)$ when $|d(y, y_1) - d(y, y_2)|$ is small since y is roughly as close to y_1 as it is as to y_2 .

Motivated by the heuristic, we propose the following objective function f_{div} as the diversity objective:

$$f_{\text{div}}(Y) := \frac{1}{|Y|} \sum_{y_1 \in Y} \sum_{y_2 \in Y \setminus \{y_1\}} d(y_1, y_2). \quad (8)$$

Another intuition for f_{div} is that it is an upper bound on the sum of the distance differences between the sample pairs in Y , assuming d is a metric. Here, $d(y_1, y_2)$ is an upper bound of $|d(y, y_1) - d(y, y_2)|$ from the triangle inequality. Therefore,

$$\begin{aligned} & \frac{1}{|Y|} \sum_{y \in Y_{\text{cand}}} \sum_{y_1 \in Y} \sum_{y_2 \in Y \setminus \{y_1\}} |d(y, y_1) - d(y, y_2)| \\ & \leq \frac{1}{|Y|} \sum_{y_1 \in Y} \sum_{y_2 \in Y \setminus \{y_1\}} d(y_1, y_2) = f_{\text{div}}(Y). \end{aligned} \quad (9)$$

Thus, ensuring f_{div} to be large is a desirable property to make $|d(y, y_1) - d(y, y_2)|$ large enough, which is likely to contribute to estimating $R(x, y)$ accurately. Note that the cost and utility functions used in NLP are often not precisely a metric, as many of them are based on neural networks (e.g., COMET, Metric-X, and LLM-as-a-Judge; Rei et al. 2020, 2022; Juraska et al. 2024; Zheng et al. 2023). Eq. 9 is intended to be an intuitive explanation of the diversity objective f_{div} rather than a theoretical guarantee of the objective.

3.3 Objective Function for Selecting Y_{ann}

Based on the two heuristics, we propose to optimize the following objective to select the subsample Y to annotate from a set of responses Y_{cand} :

$$Y_{\text{ann}} = \arg \max_{\substack{Y \subseteq Y_{\text{cand}} \\ |Y|=k}} f_{\text{rep}}(Y) + \lambda f_{\text{div}}(Y), \quad (10)$$

where λ is a hyperparameter to control the trade-off between the two objectives.

We use the cosine distance of the embedding computed by all-mpnet-base-v2 sentence BERT model, which has been shown to be effective for a variety of sentence embedding tasks (Reimers and Gurevych, 2019, 2020; Song et al., 2020).

Computing the optimal solution for Eq. 10 requires $O(|Y_{\text{cand}}|^k)$ time in the worst case. In the following experiments, we use $k = 2$ and compute the optimal solutions for Y_{ann} by enumerating all pairs of samples. In a case where k is large, we can compute Y approximately using a greedy

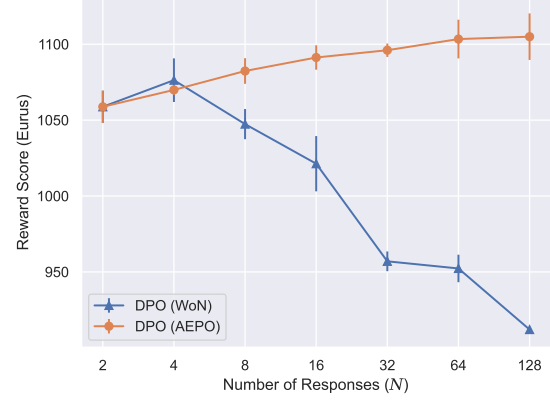


Figure 3: Evaluation of AEPO and West-of-N for DPO with an annotation budget fixed to 2 times the number of instructions on AlpacaFarm. The line represents the average reward score and the bar shows the standard deviation over three runs.

search algorithm that finds the y that maximizes the objective at each step until it selects k responses.

The use of the representativeness and diversity is shown to be useful for text generation algorithms in prior work (Vijayakumar et al., 2016; Eikema and Aziz, 2022; Jinnai et al., 2024; Li et al., 2024a). In fact, the objective function Eq. 10 corresponds to the decoding objective of Diverse Minimum Bayes Risk decoding algorithm (Jinnai et al., 2024). These objectives are also often used in active learning in NLP (Zhang et al. 2022; See Section 5), including the application to select instruction set for supervised fine tuning (Zhou et al., 2023; Li et al., 2024b). The contribution of the study is on importing these objectives to the problem of selecting a set of responses to annotate from a set of candidate responses for preference optimization, which is critical when aligning LLMs with languages, communities, and tasks with limited annotations.

4 Experiments

We first conduct simulated experiments using reward models instead of human annotation to evaluate the performance of AEPO in depth with ablation studies. Then, we evaluate AEPO on a more realistic setting of learning cultural commonsense morality (Awad et al., 2020; Hendrycks et al., 2021) where human annotations are difficult to obtain.

4.1 Simulated Experiment

For the purpose of the method, it is ideal to use human annotations to evaluate the performance of the algorithms. However, human annotations

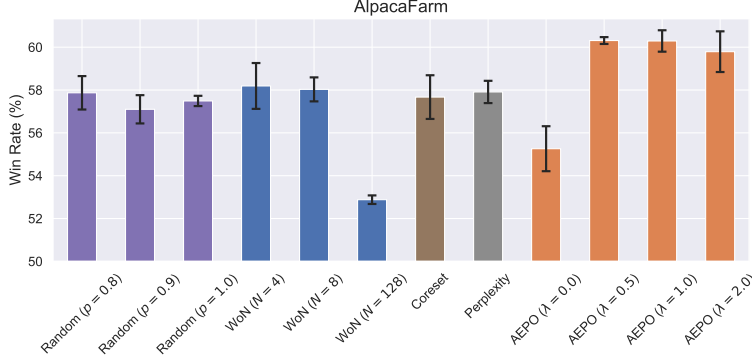


Figure 4: Evaluation of preference annotation strategies for DPO on AlpacaFarm using Mistral under the annotation budget fixed to 2 times the number of instructions. The win rate against the SFT model is evaluated. The bar represents the mean, and the error bar indicates the standard deviation of three runs.

Method	#Insts	#Annots
SFT (Mistral)	0	0
Random ($p = 0.8$)	$ \mathcal{D} $	$2 \mathcal{D} $
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $
Random ($p = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $
WoN ($N = 128$)	$ \mathcal{D} /64$	$2 \mathcal{D} $
Coreset	$ \mathcal{D} $	$2 \mathcal{D} $
Perplexity	$ \mathcal{D} $	$2 \mathcal{D} $
AEPO ($\lambda = 0$)	$ \mathcal{D} $	$2 \mathcal{D} $
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $

Table 2: The number of instructions (#Insts) and annotations (#Annots) used by the preference annotation strategies in Figures 4, 5, and 7.

are expensive and difficult to reproduce. To this end, we first evaluate the proposed method on a simulated scenario where the annotations are from open source reward models.

Datasets and models. We evaluate the performance of AEPO on DPO using the AlpacaFarm (Dubois et al., 2023) and hh-rlhf (Bai et al., 2022) datasets. We use mistral-7b-sft-beta (Mistral) (Jiang et al., 2023a; Tunstall et al., 2024) as the language model. See D.2 for the results using dolly-v2-3b (Conover et al., 2023) as the language model.

AEPO and baseline strategies. We generate up to $N = 128$ responses per instruction with nucleus sampling ($p = 0.9$) (Holtzman et al., 2020) to be used for the subsampling strategies. The temperature of the sampling algorithm is set to 1.0 for all experiments. Both AEPO and baseline strategies use the same set of responses to ensure a fair comparison.

For AEPO, the size of Y_{ann} is set to $k = 2$ and the diversity hyperparameter is set to $\lambda \in \{0.0, 0.5, 1.0, 2.0\}$ for AlpacaFarm and $\lambda \in \{0.5, 1.0, 2.0\}$ for the rest of the datasets.

As baselines, we evaluate two active learning strategies, Coreset and Perplexity, in addition to random sampling and WoN strategy. Coreset strategy (Sener and Savarese, 2018) is inspired from the coreset selection problem which chooses a subset of the dataset such that the model trained on the selected subset will perform as closely as possible to the model trained on the entire dataset (Feldman, 2020). Perplexity strategy selects a pair of responses with the highest and lowest perplexity.

See Appendix B for the details of the algorithms.

Since WoN strategy uses $N/2$ times more annotations per instruction than AEPO with $k = 2$, we reduce the number of instructions for WoN to $2/N$ so that the number of required annotations is the same as for AEPO.

Alignment procedure. We train the same model that generates the responses (Mistral) using DPO with Low-Rank Adaptation (LoRA) (Hu et al., 2022; Sidahmed et al., 2024). We set the LoRA’s $r = 64$ and $\alpha = r/4$. Other hyperparameters for the training process are described in Appendix A. For the AlpacaFarm dataset, we use the alpaca_human_preference subset as the training set and use the alpaca_farm_evaluation subset as the evaluation set. For the hh-rlhf datasets, we use the first 5000 entries of the training set of both the helpful-base and harmless-base subsets as the training set. Then we evaluate the trained model on the first 1000 entries of the test set of the helpful-base (Helpfulness) and harmless-base (Harmlessness) subsets. For WoN, we reduce the number of instructions evenly for the two subsets so that the dataset always has the same number of instructions from the two subsets.

Results. We evaluate the quality of the trained models by sampling a response using nucleus sampling ($p = 0.7$). The model output is evaluated using Eurur-RM-7B (Eurur; Yuan et al. 2024a) as it is open source and shown to have a high correlation with human annotations in RewardBench (Lambert et al., 2024).

Figure 3 shows the Eurur score of the DPO

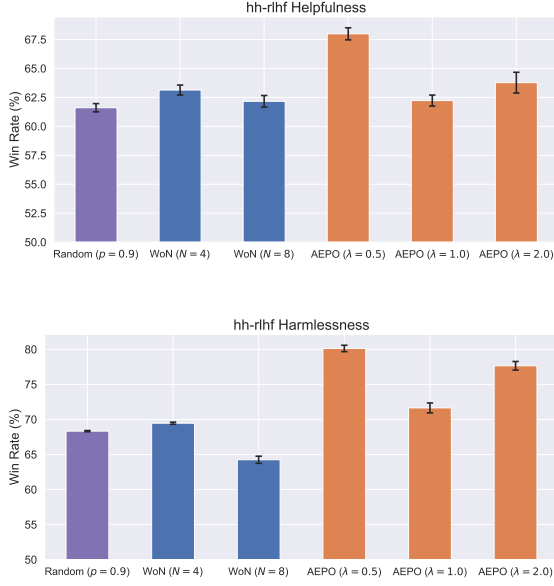


Figure 5: Evaluation of preference dataset annotation strategies for DPO on hh-rlhf’s Helpfulness and Harmlessness dataset using Mistral under the annotation budget. The win rate against the SFT model is evaluated. The bar represents the mean, and the error bar indicates the standard deviation of three runs.

models on AlpacaFarm using AEPO ($\lambda = 1.0$) and WoN with different numbers of responses. WoN with $N = 4$ outperforms the random sampling baselines (i.e., WoN with $N = 2$), even though it uses only half of the available instructions, which is consistent with the results of Song et al. (2024). However, WoN’s score drops significantly for $N \geq 8$ as the number of instructions decreases. In contrast, AEPO scales with the number of responses N and outperforms WoN (Figure 3).

Figures 4 and 5 show the win rate of the DPO models with $N = 128$ under a fixed annotation budget. The win rate is computed against the SFT model using Eurús as a reference reward model. See Appendix H for the evaluation using other reward models. In all three datasets, AEPO outperforms the baseline algorithms except for when λ is set to 0 so that no diversity is assured.

The ablation study of AEPO is described in Appendix D where we evaluate AEPO on a smaller LLM, out-of-domain tasks, using varying LoRA hyperparameters, and using varying loss functions. The result shows that AEPO consistently outperforms the baselines in a wide range of settings.

AEPO generates a diverse and representative preference dataset. We evaluate the diversity, representativeness, and quality of the preference

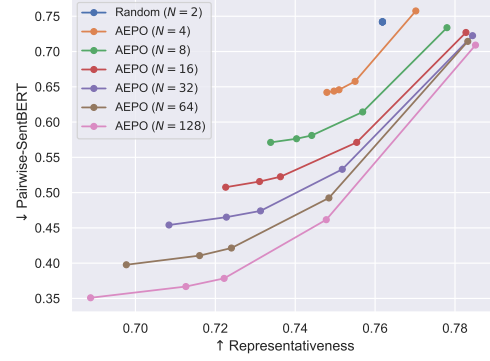


Figure 6: Diversity ($\downarrow$ Pairwise-SentBERT) and representativeness of the responses of the preference datasets $\mathcal{D}_{AE}$ generated by AEPO with a varying number of input responses (N). The number of selected responses (k) is fixed at 2. AEPO successfully generates datasets with better diversity-representativeness trade-offs without requiring additional annotations.

dataset generated by AEPO. To measure the semantic and lexical diversity of the responses, we use pairwise Sentence BERT and distinct-n (Li et al., 2016). We use the same Sentence BERT model (all-mpnet-base-v2) as AEPO to evaluate the average cosine similarity between the selected pairs of responses. Distinct-n counts the number of distinct n-grams in a sentence divided by the length of the sentence. The representativeness is measured by $-f_{\text{rep}}(Y)/|Y_{\text{cand}}|$ which is the average similarity of Y_{ann} to Y_{cand} . The quality of the responses is measured by the average reward score of the selected responses.

The result is shown in Figure 6. By using a larger number of responses N , AEPO manages to generate more diverse and representative response pairs than a random sampling with the same number of annotations. Interestingly, AEPO also results in higher-quality texts being selected than random sampling (Figure 14 in Appendix H). This aligns with prior work reporting that diversity and representativeness objectives can improve the quality of the output texts (Vijayakumar et al., 2016, 2018; Eikema and Aziz, 2022; Jinnai et al., 2024). See Appendix E for examples of the preference data generated by AEPO. We observe similar trends in the results on distinct-n, as well as the results on the hh-rlhf datasets (Figures 14, 15, and 16 in Appendix H).

Both diversity and representativeness of the preference dataset are important for preference learning. The question is what contributes to the

improved performance of AEPO. In Figures 4 and 9, AEPO with moderate size of λ outperforms AEPO with higher or lower λ . The result indicates that both the diversity and the representativeness of responses are important for the preference dataset, which is consistent with the observations in previous work (Mukherjee et al., 2023; Chen et al., 2024; Liu et al., 2024c; Song et al., 2024).

4.2 Realistic Experiment

We evaluate AEPO on a more realistic setting using JCommonsenseMorality (JCM) dataset (Takeshita et al., 2023). JCM is a collection of texts labeled with whether a text contains a morally wrong statement according to the commonsense morality of people in Japanese culture. Because commonsense morality is culturally dependent and requires annotation by the members of the community (Durmus et al., 2024; Shen et al., 2024b), it is difficult to collect a large number of annotations. Therefore, we consider the task of learning Japanese commonsense morality to be a suitable benchmark for evaluating AEPO in a realistic application.

We use 800 entries ($|\mathcal{D}| = 800$) from the train split for training and 500 entries from the test split for evaluation. The preference annotation is done semi-automatically; we use Gemma 2 to evaluate if the generated text is aligned with the human annotation in the dataset (Zheng et al., 2023; Team et al., 2024). See Appendix G for the evaluation procedure. We train a Japanese LLM (calm2-7b-chat; Ishigami 2023). All the other hyperparameters follow Section 4.1. The results are summarized in Figure 7. Overall, AEPO outperforms the baselines within the same annotation budget constraint. The result shows the potential of AEPO in tasks where the available annotations are limited.

5 Related Work

Active learning. Annotation-efficient learning has long been a challenge in natural language processing (Zhang et al., 2022). Active learning is an approach that aims to achieve training with fewer training labels by proactively selecting the data to be annotated and used for learning (Cohn et al., 1994; Settles, 2009; Houlsby et al., 2011). Active learning methods in NLP are often categorized in two strategies (Zhang et al., 2022). One uses the informativeness of the data instances, such as uncertainty and disagreement of the models (Lewis and Gale, 1994; Engelson and Dagan, 1996; Siddhant

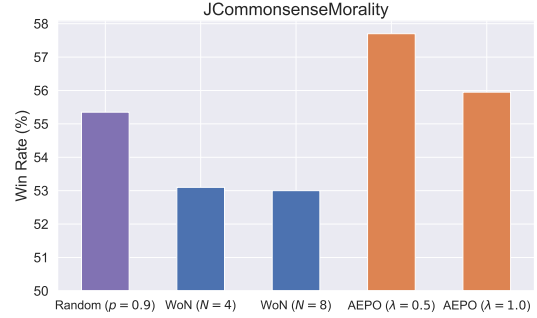


Figure 7: Evaluation of preference annotation strategies for DPO on the JCommonsenseMorality (JCM) dataset using calm2-7b-chat under a fixed annotation budget. The win rate against the SFT model is evaluated.

and Lipton, 2018; Huang et al., 2024; Dwaracherla et al., 2024). This approach has proven to be efficient in many text classification tasks. The other strategy is based on the representativeness of the data instances (McCallum and Nigam, 1998; Settles and Craven, 2008; Zhao et al., 2020; Chen and Wang, 2024). The strategy annotates instances with high average similarity to all the other instances so that it can cover a large portion of the dataset with few annotations. Another approach is to select instances that maximize the diversity of labeled instances (Eck et al., 2005; Zeng et al., 2019; Bloodgood and Callison-Burch, 2010). Our approach is related to these approaches as our objective is a combination of representative and diversity measures designed to maximize the information gain.

Minimum Bayes risk decoding. See Appendix C.

6 Conclusions

We propose Annotation-Efficient Preference Optimization (AEPO), an annotation-efficient dataset subsampling strategy for language model alignment. AEPO selects response texts to annotate from candidate responses, maximizing the representativeness and diversity. By focusing the annotation effort on the selected responses, AEPO achieves efficient preference optimization under a limited annotation budget. The experimental results and ablation studies (Appendix D) show that AEPO consistently outperforms the baseline strategies on a wide range of benchmarks. We believe that AEPO is a critical contribution to promoting preference optimization in low resource settings by addressing the severe obstacle, the cost of creating better preference data.

7 Limitations

Although our method is motivated by the situation where the human annotation is needed to align the language model, part of our experiments (AlpacaFarm and hh-rlhf) are conducted using a proxy reward model to annotate preference on training datasets instead of using human annotation. We use human annotation for the JCM dataset but combined with an LLM to evaluate the agreement of the response text with the human annotation. Evaluation of the method using human annotation directly to the responses would be desirable for future work.

The underlying assumption of the paper is that human annotation is the correct gold reference, preferable to synthesized annotation. However, it is known that human annotation can be wrong and unreliable (Ipeirotis et al., 2010; Clark et al., 2021; Hosking et al., 2024). Developing an annotation strategy that considers the possibility of annotation errors is future work.

Our focus is on developing a method to generate a diverse and representative set of responses. The preparation of diverse and representative instructions is also an important task to generate an efficient dataset (Sanh et al., 2022; Ding et al., 2023; Cui et al., 2023; Liu et al., 2024a; Xu et al., 2024a). Our method is orthogonal to methods for generating high quality instructions and can be combined. Comparing and combining AEPO with methods for generating diverse instructions is future work.

All experiments are performed using LoRA (Hu et al., 2022). The evaluation of AEPO with full parameter fine-tuning is future work. Our experiments are limited to the evaluation on DPO. Evaluating AEPO on variants of DPO (Amini et al., 2024; Gheshlaghi Azar et al., 2024; Tang et al., 2024b; Morimura et al., 2024; Zhang et al., 2024b) and other preference optimization algorithms (Ouyang et al., 2022; Zhao et al., 2023; Ahmadian et al., 2024) is future work.

The performance of AEPO depends on the choice of the hyperparameter λ . We observe that $\lambda = 1.0$ is a good choice throughout the experiments, but developing a strategy to find an effective λ for a given dataset is future work.

We assume that the cost of annotating the preference rank for N responses is linear in N . However, prior work shows that it becomes increasingly difficult to annotate preference rank over a larger set of options (Ganzfried, 2017). Because the assumption only favors WoN over AEPO, we believe it will not

affect the analysis of the paper. Evaluating the human annotation burden of ranking N responses is important future work.

We assume that the cost of annotating the preference rank for N responses is linear in N . However, prior work shows that it becomes increasingly difficult to annotate preference ranks as the number of options increases (Ganzfried, 2017). Because this assumption favors WoN over AEPO, we believe it will not affect the overall analysis presented in this paper. Evaluating the human annotation burden of ranking N responses is an important direction for future work.

8 Ethical Considerations

We believe that this work will have a positive impact by encouraging work on AI systems that work better with a diverse set of people. LLMs would be more useful if they could adapt to the preferences of diverse groups of people, even if little preference annotation is available from their communities.

We foresee our method being useful for personalizing LLMs (Greene et al., 2023; Jang et al., 2023; Kirk et al., 2023). Personalized LLMs could have far-reaching benefits, but also a number of worrisome risks, such as the propagation of polarized views. We refer to Kirk et al. (2023) for a discussion of potential risks and countermeasures for personalized LLMs.

References

- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. *Towards measuring and modeling “culture” in LLMs: A survey*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.
- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. *Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267, Bangkok, Thailand. Association for Computational Linguistics.
- Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. *Direct preference optimization with an offset*. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9954–9972, Bangkok, Thailand

670	and virtual meeting. Association for Computational	diverse human preferences. In <i>ICML 2024 Workshop</i>	727
671	Linguistics.	on Models of Human Feedback for AI Alignment.	728
672	Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain,	Jonathan D. Chang, Wenhao Zhan, Owen Oertell,	729
673	Deep Ganguli, Tom Henighan, Andy Jones, Nicholas	Kianté Brantley, Dipendra Misra, Jason D. Lee, and	730
674	Joseph, Ben Mann, Nova DasSarma, Nelson El-	Wen Sun. 2024. Dataset reset policy optimization for	731
675	hage, Zac Hatfield-Dodds, Danny Hernandez, Jack-	RLHF. <i>arXiv preprint arXiv:2404.08495</i> .	732
676	son Kernion, Kamal Ndousse, Catherine Olsson,	Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa	733
677	Dario Amodei, Tom Brown, Jack Clark, Sam Mc-	Gunaratna, Vikas Yadav, Zheng Tang, Vijay Sriniv-	734
678	Candlish, Chris Olah, and Jared Kaplan. 2021. A	asan, Tianyi Zhou, Heng Huang, and Hongxia Jin.	735
679	general language assistant as a laboratory for align-	2024. Alpagasus: Training a better Alpaca with	736
680	ment. <i>arXiv preprint arXiv:2112.00861</i> .	fewer data . In <i>The Twelfth International Conference</i>	737
681	Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rah-	on Learning Representations.	738
682	wan, and Jean-François Bonnefon. 2020. Univer-	Ruitao Chen and Liwei Wang. 2024. The power	739
683	sals and variations in moral decisions made in 42	of active multi-task learning in reinforcement	740
684	countries by 70,000 participants. <i>Proceedings of the</i>	learning from human feedback. <i>arXiv preprint</i>	741
685	<i>National Academy of Sciences</i> , 117(5):2332–2337.	<i>arXiv:2405.11226</i> .	742
686	Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda	Paul F Christiano, Jan Leike, Tom Brown, Miljan Mar-	743
687	Askell, Anna Chen, Nova DasSarma, Dawn Drain,	tic, Shane Legg, and Dario Amodei. 2017. Deep	744
688	Stanislav Fort, Deep Ganguli, Tom Henighan,	reinforcement learning from human preferences . In	745
689	Nicholas Joseph, Saurav Kadavath, Jackson Kernion,	<i>Advances in Neural Information Processing Systems</i> ,	746
690	Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac	volume 30. Curran Associates, Inc.	747
691	Hatfield-Dodds, Danny Hernandez, Tristan Hume,	Elizabeth Clark, Tal August, Sofia Serrano, Nikita	748
692	Scott Johnston, Shauna Kravec, Liane Lovitt, Neel	Haduong, Suchin Gururangan, and Noah A. Smith.	749
693	Nanda, Catherine Olsson, Dario Amodei, Tom	2021. All that’s ‘human’ is not gold: Evaluating	750
694	Brown, Jack Clark, Sam McCandlish, Chris Olah,	human evaluation of generated text . In <i>Proceedings</i>	751
695	Ben Mann, and Jared Kaplan. 2022. Training a	<i>of the 59th Annual Meeting of the Association for</i>	752
696	helpful and harmless assistant with reinforcement	<i>Computational Linguistics and the 11th International</i>	753
697	learning from human feedback. <i>arXiv preprint</i>	<i>Joint Conference on Natural Language Processing</i>	754
698	<i>arXiv:2204.05862</i> .	<i>(Volume 1: Long Papers)</i> , pages 7282–7296, Online.	755
699	Amanda Bertsch, Alex Xie, Graham Neubig, and	Association for Computational Linguistics.	756
700	Matthew Gormley. 2023. It’s MBR all the way down:	Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot,	757
701	Modern generation techniques through the lens of	Ashish Sabharwal, Carissa Schoenick, and Oyvind	758
702	minimum Bayes risk . In <i>Proceedings of the Big</i>	Tafjord. 2018. Think you have solved question	759
703	<i>Picture Workshop</i> , pages 108–122, Singapore. Asso-	answering? try ARC, the AI2 Reasoning Challenge.	760
704	ciation for Computational Linguistics.	<i>arXiv preprint arXiv:1803.05457</i> .	761
705	Michael Bloodgood and Chris Callison-Burch. 2010.	David Cohn, Les Atlas, and Richard Ladner. 1994. Im-	762
706	Bucking the trend: Large-scale cost-focused active	proving generalization with active learning. <i>Machine</i>	763
707	learning for statistical machine translation . In <i>Pro-</i>	<i>learning</i> , 15:201–221.	764
708	<i>ceedings of the 48th Annual Meeting of the Associa-</i>	Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie,	765
709	<i>tion for Computational Linguistics</i> , pages 854–864,	Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell,	766
710	Uppsala, Sweden. Association for Computational	Matei Zaharia, and Reynold Xin. 2023. Free Dolly:	767
711	Linguistics.	Introducing the world’s first truly open instruction-	768
712	Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun.	tuned LLM .	769
713	2023a. Instruction mining: When data mining meets	Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao,	770
714	large language model finetuning. <i>arXiv preprint</i>	Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and	771
715	<i>arXiv:2307.06290</i> .	Maosong Sun. 2023. Ultrafeedback: Boosting lan-	772
716	Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min	guage models with high-quality feedback. <i>arXiv</i>	773
717	Chen, and Daniel Hershcovich. 2023b. Assessing	<i>preprint arXiv:2310.01377</i> .	774
718	cross-cultural alignment between ChatGPT and hu-	Javier De la Rosa, Eduardo G Ponferrada, Paulo Ville-	775
719	man societies: An empirical study . In <i>Proceedings of</i>	gas, Pablo Gonzalez de Prado Salas, Manu Romero,	776
720	<i>the First Workshop on Cross-Cultural Considerations</i>	and Maria Grandury. 2022. Bertin: Efficient pre-	777
721	<i>in NLP (C3NLP)</i> , pages 53–67, Dubrovnik, Croatia.	training of a spanish language model using perplexity	778
722	Association for Computational Linguistics.	sampling. <i>arXiv preprint arXiv:2207.06814</i> .	779
723	Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Kop-		
724	pel, Furong Huang, Dinesh Manocha, Amrit Bedi,		
725	and Mengdi Wang. 2024. Maxmin-RLHF: Towards		
726	equitable alignment of large language models with		

780	Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin,	Sean P. Engelson and Ido Dagan. 1996. Minimizing	837
781	Shengding Hu, Zhiyuan Liu, Maosong Sun, and	manual annotation cost in supervised training from	838
782	Bowen Zhou. 2023. Enhancing chat language mod-	corpora . In <i>34th Annual Meeting of the Association</i>	839
783	els by scaling high-quality instructional conversa-	<i>for Computational Linguistics</i> , pages 319–326, Santa	840
784	tions . In <i>Proceedings of the 2023 Conference on</i>	Cruz, California, USA. Association for Computa-	841
785	<i>Empirical Methods in Natural Language Processing</i> ,	tional Linguistics.	842
786	pages 3029–3051, Singapore. Association for Com-		
787	putational Linguistics.		
788	Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang,	Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff,	843
789	Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo,	Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model	844
790	Caiming Xiong, and Tong Zhang. 2024. RLHF work-	alignment as prospect theoretic optimization. <i>arXiv</i>	845
791	flow: From reward modeling to online RLHF . <i>Trans-</i>	<i>preprint arXiv:2402.01306</i> .	846
792	<i>actions on Machine Learning Research</i> .		
793	Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey,	Dan Feldman. 2020. Core-sets: Updated survey. <i>Sam-</i>	847
794	Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman,	<i>pling techniques for supervised or unsupervised tasks</i> ,	848
795	Akhil Mathur, Alan Schelten, Amy Yang, Angela	pages 23–44.	849
796	Fan, et al. 2024. The llama 3 herd of models. <i>arXiv</i>		
797	<i>preprint arXiv:2407.21783</i> .		
798	Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi	Markus Freitag, Behrooz Ghorbani, and Patrick Fernan-	850
799	Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin,	des. 2023. Epsilon sampling rocks: Investigating	851
800	Percy S Liang, and Tatsunori B Hashimoto. 2023.	sampling strategies for minimum Bayes risk decod-	852
801	AlpacaFarm: A simulation framework for methods	ing for machine translation . In <i>Findings of the As-</i>	853
802	that learn from human feedback . In <i>Advances in</i>	<i>sociation for Computational Linguistics: EMNLP</i>	854
803	<i>Neural Information Processing Systems</i> , volume 36,	2023, pages 9198–9209, Singapore. Association for	855
804	pages 30039–30069. Curran Associates, Inc.	Computational Linguistics.	856
805	Yann Dubois, Percy Liang, and Tatsunori Hashimoto.	Sam Ganzfried. 2017. Optimal number of choices in	857
806	2024. Length-controlled AlpacaEval: A simple de-	rating contexts . In <i>Proceedings of the NIPS 2016</i>	858
807	biasing of automatic evaluators . In <i>First Conference</i>	<i>Workshop on Imperfect Decision Makers</i> , volume 58	859
808	<i>on Language Modeling</i> .	of <i>Proceedings of Machine Learning Research</i> , pages	860
809	Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas	61–74. PMLR.	861
810	Schiefer, Amanda Askell, Anton Bakhtin, Carol		
811	Chen, Zac Hatfield-Dodds, Danny Hernandez,	Leo Gao, John Schulman, and Jacob Hilton. 2023a.	862
812	Nicholas Joseph, Liane Lovitt, Sam McCandlish,	Scaling laws for reward model overoptimization . In	863
813	Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared	<i>Proceedings of the 40th International Conference</i>	864
814	Kaplan, Jack Clark, and Deep Ganguli. 2024. To-	<i>on Machine Learning</i> , volume 202 of <i>Proceedings</i>	865
815	wards measuring the representation of subjective	<i>of Machine Learning Research</i> , pages 10835–10866.	866
816	global opinions in language models . In <i>First Confer-</i>	PMLR.	867
817	<i>ence on Language Modeling</i> .		
818	Vikranth Dwaracherla, Seyed Mohammad Asghari, Bo-	Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman,	868
819	tao Hao, and Benjamin Van Roy. 2024. Efficient	Sid Black, Anthony DiPofi, Charles Foster, Laurence	869
820	exploration for LLMs . In <i>Proceedings of the 41st</i>	Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li,	870
821	<i>International Conference on Machine Learning</i> , vol-	Kyle McDonell, Niklas Muennighoff, Chris Ociepa,	871
822	ume 235 of <i>Proceedings of Machine Learning Re-</i>	Jason Phang, Laria Reynolds, Hailey Schoelkopf,	872
823	<i>search</i> , pages 12215–12227. PMLR.	Aviya Skowron, Lintang Sutawika, Eric Tang, An-	873
824	Matthias Eck, Stephan Vogel, and Alex Waibel. 2005.	ish Thite, Ben Wang, Kevin Wang, and Andy Zou.	874
825	Low cost portability for statistical machine transla-	2023b. A framework for few-shot language model	875
826	tion based on n-gram frequency and TF-IDF . In	evaluation .	876
827	<i>Proceedings of the Second International Workshop</i>		
828	<i>on Spoken Language Translation</i> , Pittsburgh, Penn-	Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin	877
829	sylvania, USA.	Tao, Xiaofeng Zhao, Mahong Xia, Zhang Li, Boxing	878
830	Bryan Eikema and Wilker Aziz. 2022. Sampling-based	Chen, Hao Yang, Bei Li, Tong Xiao, and JingBo Zhu.	879
831	approximations to minimum Bayes risk decoding	2024. Clustering and ranking: Diversity-preserved	880
832	for neural machine translation . In <i>Proceedings of</i>	instruction selection through expert-aligned quality	881
833	<i>the 2022 Conference on Empirical Methods in Natu-</i>	estimation . In <i>Proceedings of the 2024 Conference</i>	882
834	<i>ral Language Processing</i> , pages 10978–10993, Abu	<i>on Empirical Methods in Natural Language Process-</i>	883
835	Dhabi, United Arab Emirates. Association for Com-	<i>ing</i> , pages 464–478, Miami, Florida, USA. Associa-	884
836	putational Linguistics.	tion for Computational Linguistics.	885
		Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bi-	886
		ljal Piot, Remi Munos, Mark Rowland, Michal Valko,	887
		and Daniele Calandriello. 2024. A general theoretical	888
		paradigm to understand learning from human pref-	889
		erences . In <i>Proceedings of The 27th International</i>	890
		<i>Conference on Artificial Intelligence and Statistics</i> ,	891
		volume 238 of <i>Proceedings of Machine Learning</i>	892
		<i>Research</i> , pages 4447–4455. PMLR.	893

894	Travis Greene, Galit Shmueli, and Soumya Ray. 2023.	<i>on Human Computation</i> , HCOMP '10, page 64–67,	950
895	Taking the person seriously: Ethically aware is re-	New York, NY, USA. Association for Computing	951
896	search in the era of reinforcement learning-based	Machinery.	952
897	personalization. <i>Journal of the Association for Infor-</i>		
898	<i>mation Systems</i> , 24(6):1527–1561.		
899	Arnav Gudibande, Eric Wallace, Charlie Victor Snell,	Ryosuke Ishigami. 2023. cyberagent/calm2-7b-chat .	953
900	Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey		
901	Levine, and Dawn Song. 2024. The false promise	Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchen-	954
902	of imitating proprietary language models . In <i>The</i>	bauer, Hong-Min Chu, Gowthami Somepalli, Brian R.	955
903	<i>Twelfth International Conference on Learning Repre-</i>	Bartoldson, Bhavya Kailkhura, Avi Schwarzschild,	956
904	<i>sentations</i> .	Aniruddha Saha, Micah Goldblum, Jonas Geiping,	957
905		and Tom Goldstein. 2024. NEFTune: Noisy embed-	958
906	Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu,	dings improve instruction finetuning . In <i>The Twelfth</i>	959
907	Misha Khalman, Felipe Llinares, Alexandre Rame,	<i>International Conference on Learning Representa-</i>	960
908	Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret,	<i>tions</i> .	961
909	and Mathieu Blondel. 2024. Direct language model		
910	alignment from online AI feedback. <i>arXiv preprint</i>	Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong	962
911	<i>arXiv:2402.04792</i> .	Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh	963
912	Dan Hendrycks, Collin Burns, Steven Basart, Andrew	Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu.	964
913	Critch, Jerry Li, Dawn Song, and Jacob Steinhardt.	2023. Personalized soups: Personalized large lan-	965
914	2021. Aligning AI with shared human values. <i>Pro-</i>	guage model alignment via post-hoc parameter merg-	966
915	<i>ceedings of the International Conference on Learning</i>	ing. <i>arXiv preprint arXiv:2310.11564</i> .	967
916	<i>Representations (ICLR)</i> .		
917	Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and	Albert Q. Jiang, Alexandre Sablayrolles, Arthur Men-	968
918	Yejin Choi. 2020. The curious case of neural text de-	sch, Chris Bamford, Devendra Singh Chaplot, Diego	969
919	generation . In <i>International Conference on Learning</i>	de las Casas, Florian Bressand, Gianna Lengyel, Guil-	970
920	<i>Representations</i> .	laume Lample, Lucile Saulnier, L��lio Renard Lavaud,	971
921		Marie-Anne Lachaux, Pierre Stock, Teven Le Scao,	972
922	Or Honovich, Thomas Scialom, Omer Levy, and Timo	Thibaut Lavril, Thomas Wang, Timoth��e Lacroix,	973
923	Schick. 2023. Unnatural instructions: Tuning lan-	and William El Sayed. 2023a. Mistral 7b. <i>arXiv</i>	974
924	guage models with (almost) no human labor . In	<i>preprint arXiv:2310.06825</i> .	975
925	<i>Proceedings of the 61st Annual Meeting of the As-</i>		
926	<i>sociation for Computational Linguistics (Volume 1:</i>	Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023b.	976
927	<i>Long Papers)</i> , pages 14409–14428, Toronto, Canada.	LLM-blender: Ensembling large language models	977
928	Association for Computational Linguistics.	with pairwise ranking and generative fusion . In <i>Pro-</i>	978
929		<i>ceedings of the 61st Annual Meeting of the Associa-</i>	979
930	Tom Hosking, Phil Blunsom, and Max Bartolo. 2024.	<i>tion for Computational Linguistics (Volume 1: Long</i>	980
931	Human feedback is not gold standard . In <i>The Twelfth</i>	<i>Papers)</i> , pages 14165–14178, Toronto, Canada. As-	981
932	<i>International Conference on Learning Representa-</i>	sociation for Computational Linguistics.	982
933	<i>tions</i> .		
934		Yuu Jinnai. 2024. Does cross-cultural alignment change	983
935	Neil Houlsby, Ferenc Husz��r, Zoubin Ghahramani, and	the commonsense morality of language models? In	984
936	M��t�� Lengyel. 2011. Bayesian active learning for	<i>Proceedings of the 2nd Workshop on Cross-Cultural</i>	985
937	classification and preference learning. <i>arXiv preprint</i>	<i>Considerations in NLP</i> , pages 48–64, Bangkok, Thai-	986
938	<i>arXiv:1112.5745</i> .	land. Association for Computational Linguistics.	987
939			
940	Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan	Yuu Jinnai, Ukyo Honda, Tetsuro Morimura, and Peinan	988
941	Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and	Zhang. 2024. Generating diverse and high-quality	989
942	Weizhu Chen. 2022. LoRA: Low-rank adaptation of	texts by minimum Bayes risk decoding . In <i>Find-</i>	990
943	large language models . In <i>International Conference</i>	<i>ings of the Association for Computational Linguistics</i>	991
944	<i>on Learning Representations</i> .	<i>ACL 2024</i> , pages 8494–8525, Bangkok, Thailand	992
945		and virtual meeting. Association for Computational	993
946	Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv,	Linguistics.	994
947	and Ido Dagan. 2024. Selective annotation via data	Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and	995
948	allocation: These data should be triaged to experts for	Markus Freitag. 2024. MetricX-24: The Google	996
949	annotation rather than the model . In <i>Findings of the</i>	submission to the WMT 2024 metrics shared task .	997
	<i>Association for Computational Linguistics: EMNLP</i>	In <i>Proceedings of the Ninth Conference on Machine</i>	998
	<i>2024</i> , pages 301–320, Miami, Florida, USA. Associ-	<i>Translation</i> , pages 492–504, Miami, Florida, USA.	999
	ation for Computational Linguistics.	Association for Computational Linguistics.	1000
	Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang.		
	2010. Quality management on amazon mechanical	Hannah Rose Kirk, Bertie Vidgen, Paul R��ttger, and	1001
	turk . In <i>Proceedings of the ACM SIGKDD Workshop</i>	Scott A. Hale. 2023. Personalisation within bounds:	1002
		A risk taxonomy and policy framework for the align-	1003
		ment of large language models with personalised	1004
		feedback. <i>arXiv preprint arXiv:2303.05453</i> .	1005

1006	Hannah Rose Kirk, Alexander Whitefield, Paul Röttger,	Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and	1062
1007	Andrew Bean, Katerina Margatina, Juan Ciro, Rafael	Deheng Ye. 2024a. More agents is all you need.	1063
1008	Mosquera, Max Bartolo, Adina Williams, He He,	<i>Transactions on Machine Learning Research.</i>	1064
1009	Bertie Vidgen, and Scott A. Hale. 2024. The PRISM		
1010	alignment project: What participatory, representa-	Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang	1065
1011	tive and individualised human feedback reveals about	Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and	1066
1012	the subjective and multicultural alignment of large	Jing Xiao. 2024b. From quantity to quality: Boosting	1067
1013	language models. <i>arXiv preprint arXiv:2404.16019.</i>	LLM performance with self-guided data selection	1068
		for instruction tuning. In <i>Proceedings of the 2024</i>	1069
1014	Andreas Köpf, Yannic Kilcher, Dimitri von Rütte,	<i>Conference of the North American Chapter of the</i>	1070
1015	Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens,	<i>Association for Computational Linguistics: Human</i>	1071
1016	Abdullah Barhoum, Duc Minh Nguyen, Oliver	<i>Language Technologies (Volume 1: Long Papers),</i>	1072
1017	Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri,	pages 7602–7635, Mexico City, Mexico. Association	1073
1018	David Alexandrovich Glushkov, Arnav Varma Dan-	for Computational Linguistics.	1074
1019	tuluri, Andrew Maguire, Christoph Schuhmann, Huu		
1020	Nguyen, and Alexander Julian Mattick. 2023. Ope-	Stephanie Lin, Jacob Hilton, and Owain Evans. 2022.	1075
1021	nAssistant conversations - democratizing large lan-	TruthfulQA: Measuring how models mimic human	1076
1022	guage model alignment. In <i>Thirty-seventh Con-</i>	falsehoods. In <i>Proceedings of the 60th Annual Meet-</i>	1077
1023	<i>ference on Neural Information Processing Systems</i>	<i>ing of the Association for Computational Linguistics</i>	1078
1024	<i>Datasets and Benchmarks Track.</i>	<i>(Volume 1: Long Papers),</i> pages 3214–3252, Dublin,	1079
		Ireland. Association for Computational Linguistics.	1080
1025	Shankar Kumar and William Byrne. 2002. Minimum	Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe	1081
1026	Bayes-risk word alignments of bilingual texts. In <i>Pro-</i>	Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi	1082
1027	<i>ceedings of the 2002 Conference on Empirical Meth-</i>	Yang, Denny Zhou, and Andrew M. Dai. 2024a. Best	1083
1028	<i>ods in Natural Language Processing (EMNLP 2002),</i>	practices and lessons learned on synthetic data. In	1084
1029	pages 140–147. Association for Computational Lin-	<i>First Conference on Language Modeling.</i>	1085
1030	guistics.		
1031	Shankar Kumar and William Byrne. 2004. Minimum	Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman,	1086
1032	Bayes-risk decoding for statistical machine transla-	Mohammad Saleh, Peter J Liu, and Jialu Liu. 2024b.	1087
1033	tion. In <i>Proceedings of the Human Language Tech-</i>	Statistical rejection sampling improves preference op-	1088
1034	<i>nology Conference of the North American Chapter</i>	timization. In <i>The Twelfth International Conference</i>	1089
1035	<i>of the Association for Computational Linguistics:</i>	<i>on Learning Representations.</i>	1090
1036	<i>HLT-NAACL 2004,</i> pages 169–176, Boston, Mas-		
1037	sachusetts, USA. Association for Computational Lin-	Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and	1091
1038	guistics.	Junxian He. 2024c. What makes good data for align-	1092
		ment? a comprehensive study of automatic data se-	1093
1039	Nathan Lambert, Valentina Pyatkin, Jacob Morrison,	lection in instruction tuning. In <i>The Twelfth Interna-</i>	1094
1040	LJ Miranda, Bill Yuchen Lin, Khyathi Chandu,	<i>tional Conference on Learning Representations.</i>	1095
1041	Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi,		
1042	Noah A. Smith, and Hannaneh Hajishirzi. 2024. Re-	Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Jun-	1096
1043	wardBench: Evaluating reward models for language	yang Lin, Chuanqi Tan, Chang Zhou, and Jingren	1097
1044	modeling. <i>arXiv preprint arXiv:2403.13787.</i>	Zhou. 2024. #instag: Instruction tagging for analyz-	1098
1045	Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas	ing supervised fine-tuning of large language models.	1099
1046	Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop,	In <i>The Twelfth International Conference on Learning</i>	1100
1047	Ethan Hall, Victor Carbune, Abhinav Rastogi, et al.	<i>Representations.</i>	1101
1048	2024. RLAIF vs. RLHF: Scaling reinforcement learn-		
1049	ing from human feedback with AI feedback. In <i>Forty-</i>	Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex	1102
1050	<i>first International Conference on Machine Learning.</i>	Wang, Marzieh Fadaee, and Sara Hooker. 2023.	1103
		When less is more: Investigating data pruning	1104
1051	David D. Lewis and William A. Gale. 1994. A sequen-	for pretraining LLMs at scale. <i>arXiv preprint</i>	1105
1052	tial algorithm for training text classifiers. In <i>SIGIR</i>	<i>arXiv:2309.04564.</i>	1106
1053	'94, pages 3–12, London. Springer London.		
1054	Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao,	Andrew McCallum and Kamal Nigam. 1998. Employ-	1107
1055	and Bill Dolan. 2016. A diversity-promoting ob-	ing em and pool-based active learning for text classifi-	1108
1056	jective function for neural conversation models. In	cation. In <i>Proceedings of the Fifteenth International</i>	1109
1057	<i>Proceedings of the 2016 Conference of the North</i>	<i>Conference on Machine Learning, ICML '98,</i> page	1110
1058	<i>American Chapter of the Association for Computa-</i>	350–358, San Francisco, CA, USA. Morgan Kauf-	1111
1059	<i>tional Linguistics: Human Language Technologies,</i>	mann Publishers Inc.	1112
1060	pages 110–119, San Diego, California. Association		
1061	for Computational Linguistics.	Tetsuro Morimura, Mitsuki Sakamoto, Yuu Jinnai, Ken-	1113
		shi Abe, and Kaito Ariu. 2024. Filtered direct prefer-	1114
		ence optimization. In <i>Proceedings of the 2024 Con-</i>	1115
		<i>ference on Empirical Methods in Natural Language</i>	1116
		<i>Processing,</i> pages 22729–22770, Miami, Florida,	1117
		USA. Association for Computational Linguistics.	1118

1119	Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawa-	2022. COMET-22: Unbabel-IST 2022 submission	1176
1120	har, Sahaj Agarwal, Hamid Palangi, and Ahmed	for the metrics shared task . In <i>Proceedings of the</i>	1177
1121	Awadallah. 2023. Orca: Progressive learning from	<i>Seventh Conference on Machine Translation (WMT)</i> ,	1178
1122	complex explanation traces of GPT-4. <i>arXiv preprint</i>	pages 578–585, Abu Dhabi, United Arab Emirates	1179
1123	<i>arXiv:2306.02707</i> .	(Hybrid). Association for Computational Linguistics.	1180
1124	Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu.	Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon	1181
1125	2024. Having beer after prayer? measuring cultural	Lavie. 2020. COMET: A neural framework for MT	1182
1126	bias in large language models . In <i>Proceedings of the</i>	evaluation . In <i>Proceedings of the 2020 Conference</i>	1183
1127	<i>62nd Annual Meeting of the Association for Computa-</i>	<i>on Empirical Methods in Natural Language Process-</i>	1184
1128	<i>tational Linguistics (Volume 1: Long Papers)</i> , pages	<i>ing (EMNLP)</i> , pages 2685–2702, Online. Association	1185
1129	16366–16393, Bangkok, Thailand. Association for	for Computational Linguistics.	1186
1130	Computational Linguistics.		
1131	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	Nils Reimers and Iryna Gurevych. 2019. Sentence-	1187
1132	Carroll Wainwright, Pamela Mishkin, Chong Zhang,	BERT: Sentence embeddings using Siamese BERT-	1188
1133	Sandhini Agarwal, Katarina Slama, Alex Ray, John	networks . In <i>Proceedings of the 2019 Conference on</i>	1189
1134	Schulman, Jacob Hilton, Fraser Kelton, Luke Miller,	<i>Empirical Methods in Natural Language Processing</i>	1190
1135	Maddie Simens, Amanda Askell, Peter Welinder,	<i>and the 9th International Joint Conference on Natu-</i>	1191
1136	Paul F Christiano, Jan Leike, and Ryan Lowe. 2022.	<i>ral Language Processing (EMNLP-IJCNLP)</i> , pages	1192
1137	Training language models to follow instructions with	3982–3992, Hong Kong, China. Association for Com-	1193
1138	human feedback . In <i>Advances in Neural Information</i>	putational Linguistics.	1194
1139	<i>Processing Systems</i> , volume 35, pages 27730–27744.		
1140	Curran Associates, Inc.	Nils Reimers and Iryna Gurevych. 2020. Making	1195
1141	Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebas-	monolingual sentence embeddings multilingual us-	1196
1142	tian Krause, and Aliaksei Severyn. 2024. West-of-N:	ing knowledge distillation . In <i>Proceedings of the</i>	1197
1143	Synthetic preference generation for improved reward	<i>2020 Conference on Empirical Methods in Natural</i>	1198
1144	modeling . In <i>ICLR 2024 Workshop on Navigating</i>	<i>Language Processing</i> . Association for Computational	1199
1145	<i>and Addressing Data Problems for Foundation Mod-</i>	Linguistics.	1200
1146	<i>els</i> .		
1147	Giada Pistilli, Alina Leidingner, Yacine Jernite, Atoosa	Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavat-	1201
1148	Kasirzadeh, Alexandra Sasha Luccioni, and Margaret	ula, and Yejin Choi. 2021. Winogrande: An adver-	1202
1149	Mitchell. 2024. Civics: Building a dataset for exam-	sarial Winograd schema challenge at scale. <i>Commu-</i>	1203
1150	ining culturally-informed values in large language	<i>nications of the ACM</i> , 64(9):99–106.	1204
1151	models . <i>Proceedings of the AAAI/ACM Conference</i>		
1152	<i>on AI, Ethics, and Society</i> , 7(1):1132–1144.	Victor Sanh, Albert Webson, Colin Raffel, Stephen	1205
1153	Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin	Bach, Lintang Sutawika, Zaid Alyafeai, Antoine	1206
1154	Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. Val-	Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey,	1207
1155	uenet: A new dataset for human value driven di-	M Saiful Bari, Canwen Xu, Urmish Thakker,	1208
1156	alogue system . In <i>Thirty-Sixth AAAI Conference</i>	Shanya Sharma Sharma, Eliza Szczechla, Taewoon	1209
1157	<i>on Artificial Intelligence, AAAI 2022, Thirty-Fourth</i>	Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti	1210
1158	<i>Conference on Innovative Applications of Artificial</i>	Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han	1211
1159	<i>Intelligence, IAAI 2022, The Twelveth Symposium</i>	Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong,	1212
1160	<i>on Educational Advances in Artificial Intelligence,</i>	Harshit Pandey, Rachel Bawden, Thomas Wang, Tr-	1213
1161	<i>EAAI 2022 Virtual Event, February 22 - March 1,</i>	ishala Neeraj, Jos Rozen, Abheesht Sharma, And-	1214
1162	2022, pages 11183–11191. AAAI Press.	rea Santilli, Thibault Fevry, Jason Alan Fries, Ryan	1215
1163	Rafael Rafailov, Archit Sharma, Eric Mitchell, Christo-	Teehan, Teven Le Scao, Stella Biderman, Leo Gao,	1216
1164	pher D Manning, Stefano Ermon, and Chelsea Finn.	Thomas Wolf, and Alexander M Rush. 2022. Multi-	1217
1165	2023. Direct preference optimization: Your language	task prompted training enables zero-shot task gener-	1218
1166	model is secretly a reward model . In <i>Thirty-seventh</i>	alization . In <i>International Conference on Learning</i>	1219
1167	<i>Conference on Neural Information Processing Sys-</i>	<i>Representations</i> .	1220
1168	<i>tems</i> .		
1169	Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina	Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino	1221
1170	Reinecke, and Maarten Sap. 2024. Normad: A bench-	Lee, Percy Liang, and Tatsunori Hashimoto. 2023.	1222
1171	mark for measuring the cultural adaptability of large	Whose opinions do language models reflect? In	1223
1172	language models. <i>arXiv preprint arXiv:2404.12464</i> .	<i>Proceedings of the 40th International Conference</i>	1224
1173	Ricardo Rei, José G. C. de Souza, Duarte Alves,	<i>on Machine Learning</i> , volume 202 of <i>Proceedings</i>	1225
1174	Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova,	<i>of Machine Learning Research</i> , pages 29971–30004.	1226
1175	Alon Lavie, Luisa Coheur, and André F. T. Martins.	PMLR.	1227
		Ozan Sener and Silvio Savarese. 2018. Active learn-	1228
		ing for convolutional neural networks: A core-set	1229
		approach. In <i>International Conference on Learning</i>	1230
		<i>Representations</i> .	1231

1232	Burr Settles. 2009. Active learning literature survey.	1291
1233	Technical report, University of Wisconsin-Madison	1292
1234	Department of Computer Sciences.	1293
1235	Burr Settles and Mark Craven. 2008. An analysis of	1294
1236	active learning strategies for sequence labeling tasks.	1295
1237	In <i>Proceedings of the 2008 Conference on Empirical</i>	1296
1238	<i>Methods in Natural Language Processing</i> , pages	1297
1239	1070–1079, Honolulu, Hawaii. Association for Com-	1298
1240	putational Linguistics.	1299
1241	Siqi Shen, Lajanugen Logeswaran, Moontae Lee,	1300
1242	Honglak Lee, Soujanya Poria, and Rada Mihalcea.	1301
1243	2024a. Understanding the capabilities and limita-	1302
1244	tions of large language models for cultural common-	1303
1245	sense. In <i>Proceedings of the 2024 Conference of</i>	1304
1246	<i>the North American Chapter of the Association for</i>	1305
1247	<i>Computational Linguistics: Human Language Tech-</i>	1306
1248	<i>nologies (Volume 1: Long Papers)</i> , pages 5668–5680,	1307
1249	Mexico City, Mexico. Association for Computational	1308
1250	Linguistics.	1309
1251	Siqi Shen, Lajanugen Logeswaran, Moontae Lee,	1310
1252	Honglak Lee, Soujanya Poria, and Rada Mihalcea.	1311
1253	2024b. Understanding the capabilities and limita-	1312
1254	tions of large language models for cultural common-	1313
1255	sense. In <i>Proceedings of the 2024 Conference of</i>	1314
1256	<i>the North American Chapter of the Association for</i>	1315
1257	<i>Computational Linguistics: Human Language Tech-</i>	1316
1258	<i>nologies (Volume 1: Long Papers)</i> , pages 5668–5680,	1317
1259	Mexico City, Mexico. Association for Computational	1318
1260	Linguistics.	1319
1261	Hakim Sidahmed, Samrat Phatale, Alex Hutcheson,	1320
1262	Zhuonan Lin, Zhang Chen, Zac Yu, Jarvis Jin, Roman	1321
1263	Komarytsia, Christiane Ahlheim, Yonghao Zhu, Sim-	1322
1264	ral Chaudhary, Bowen Li, Saravanan Ganesh, Bill	1323
1265	Byrne, Jessica Hoffmann, Hassan Mansoor, Wei Li,	1324
1266	Abhinav Rastogi, and Lucas Dixon. 2024. PERL: Pa-	1325
1267	rameter efficient reinforcement learning from human	1326
1268	feedback. <i>arXiv preprint arXiv:2403.10704</i> .	1327
1269	Aditya Siddhant and Zachary C. Lipton. 2018. Deep	1328
1270	Bayesian active learning for natural language pro-	1329
1271	cessing: Results of a large-scale empirical study.	1330
1272	In <i>Proceedings of the 2018 Conference on Empirical</i>	1331
1273	<i>Methods in Natural Language Processing</i> , pages	1332
1274	2904–2909, Brussels, Belgium. Association for Com-	1333
1275	putational Linguistics.	1334
1276	Feifan Song, Bowen Yu, Hao Lang, Haiyang Yu, Fei	1335
1277	Huang, Houfeng Wang, and Yongbin Li. 2024. Scal-	1336
1278	ing data diversity for fine-tuning language models in	1337
1279	human alignment. In <i>Proceedings of the 2024 Joint</i>	1338
1280	<i>International Conference on Computational Linguis-</i>	1339
1281	<i>tics, Language Resources and Evaluation (LREC-</i>	1340
1282	<i>COLING 2024)</i> , pages 14358–14369, Torino, Italia.	1341
1283	ELRA and ICCL.	1342
1284	Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-	1343
1285	Yan Liu. 2020. MPNet: Masked and permuted pre-	1344
1286	training for language understanding. In <i>Advances</i>	1345
1287	<i>in Neural Information Processing Systems 33: An-</i>	1346
1288	<i>annual Conference on Neural Information Processing</i>	1347
1289	<i>Systems 2020, NeurIPS 2020, December 6-12, 2020,</i>	1348
1290	<i>virtual.</i>	1349
	Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Syd-	
	ney Levine, Valentina Pyatkin, Peter West, Nouha	
	Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula,	
	Maarten Sap, John Tasioulas, and Yejin Choi. 2024a.	
	Value kaleidoscope: Engaging AI with pluralistic	
	human values, rights, and duties. In <i>Thirty-Eighth</i>	
	<i>AAAI Conference on Artificial Intelligence, AAAI</i>	
	<i>2024, Thirty-Sixth Conference on Innovative Applica-</i>	
	<i>tions of Artificial Intelligence, IAAI 2024, Fourteenth</i>	
	<i>Symposium on Educational Advances in Artificial</i>	
	<i>Intelligence, EAAI 2014, February 20-27, 2024, Van-</i>	
	<i>couver, Canada</i> , pages 19937–19947. AAAI Press.	
	Taylor Sorensen, Jared Moore, Jillian Fisher,	
	Mitchell L Gordon, Niloofar Mireshghallah, Christo-	
	pher Michael Rytting, Andre Ye, Liwei Jiang,	
	Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin	
	Choi. 2024b. Position: A roadmap to pluralistic	
	alignment. In <i>Proceedings of the 41st International</i>	
	<i>Conference on Machine Learning</i> , volume 235 of	
	<i>Proceedings of Machine Learning Research</i> , pages	
	46280–46302. PMLR.	
	Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel	
	Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford,	
	Dario Amodei, and Paul F Christiano. 2020. Learn-	
	ing to summarize with human feedback. In <i>Ad-</i>	
	<i>vances in Neural Information Processing Systems</i> ,	
	volume 33, pages 3008–3021. Curran Associates,	
	Inc.	
	Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi,	
	Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf,	
	Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023.	
	Selective annotation makes language models better	
	few-shot learners. In <i>The Eleventh International Con-</i>	
	<i>ference on Learning Representations.</i>	
	Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky.	
	2023. Follow the wisdom of the crowd: Effective	
	text generation via minimum Bayes risk decoding.	
	In <i>Findings of the Association for Computational</i>	
	<i>Linguistics: ACL 2023</i> , pages 4265–4293, Toronto,	
	Canada. Association for Computational Linguistics.	
	Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael	
	Rafailov, Jeff Schneider, Tengyang Xie, Stefano Er-	
	mon, Chelsea Finn, and Aviral Kumar. 2024. Pref-	
	erence fine-tuning of LLMs should leverage subop-	
	timal, on-policy data. In <i>Forty-first International</i>	
	<i>Conference on Machine Learning.</i>	
	Masashi Takeshita, Rafal Rzepka, and Kenji Araki.	
	2023. JCommonsenseMorality: Japanese dataset for	
	evaluating commonsense morality understanding. In	
	<i>In Proceedings of The Twenty Ninth Annual Meeting</i>	
	<i>of The Association for Natural Language Processing</i>	
	<i>(NLP2023)</i> , pages 357–362. In Japanese.	
	Yunhao Tang, Daniel Zhaoan Guo, Zeyu Zheng,	
	Daniele Calandriello, Yuan Cao, Eugene Tarassov,	
	Rémi Munos, Bernardo Ávila Pires, Michal Valko,	
	Yong Cheng, and Will Dabney. 2024a. Under-	
	standing the performance gap between online	
	and offline alignment algorithms. <i>arXiv preprint</i>	
	<i>arXiv:2405.08448</i> .	

1350	Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng,	Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan	1413
1351	Daniele Calandriello, Remi Munos, Mark Rowland,	Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh	1414
1352	Pierre Harvey Richemond, Michal Valko, Bernardo	Giang, Ludovic Peran, Tris Warkentin, Eli Collins,	1415
1353	Avila Pires, and Bilal Piot. 2024b. Generalized preference optimization: A unified approach to offline alignment . In <i>Proceedings of the 41st International Conference on Machine Learning</i> , volume 235 of <i>Proceedings of Machine Learning Research</i> , pages 47725–47742. PMLR.	Joelle Barral, Zoubin Ghahramani, Raia Hadsell,	1416
1354		D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov,	1417
1355		Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray	1418
1356		Kavukcuoglu, Clement Farabet, Elena Buchatskaya,	1419
1357		Sebastian Borgeaud, Noah Fiedel, Armand Joulin,	1420
1358		Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size . <i>arXiv preprint arXiv:2408.00118</i> .	1421
1359	Gemma Team, Morgane Riviere, Shreya Pathak,		1422
1360	Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupati-		1423
1361	raju, Léonard Hussenot, Thomas Mesnard, Bobak		1424
1362	Shahriari, Alexandre Ramé, Johan Ferret, Peter	Tristan Thrush, Christopher Potts, and Tatsunori	1425
1363	Liu, Pouya Tafti, Abe Friesen, Michelle Casbon,	Hashimoto. 2024. Improving pretraining data using perplexity correlations . <i>arXiv preprint arXiv:2409.05816</i> .	1426
1364	Sabela Ramos, Ravin Kumar, Charline Le Lan,		1427
1365	Sammy Jerome, Anton Tsitsulin, Nino Vieillard,		1428
1366	Piotr Stanczyk, Sertan Girgin, Nikola Momchev,	Lewis Tunstall, Edward Emanuel Beeching, Nathan	1429
1367	Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill,	Lambert, Nazneen Rajani, Kashif Rasul, Younes	1430
1368	Behnam Neyshabur, Olivier Bachem, Alanna Wal-	Belkada, Shengyi Huang, Leandro Von Werra, Clé-	1431
1369	ton, Aliaksei Severyn, Alicia Parrish, Aliya Ah-	mentine Fourrier, Nathan Habib, Nathan Sarrazin,	1432
1370	mad, Allen Hutchison, Alvin Abdagic, Amanda	Omar Sanseviero, Alexander M Rush, and Thomas	1433
1371	Carl, Amy Shen, Andy Brock, Andy Coenen, An-	Wolf. 2024. Zephyr: Direct distillation of LM alignment . In <i>First Conference on Language Modeling</i> .	1434
1372	thony Laforge, Antonia Paterson, Ben Bastian, Bilal		1435
1373	Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu	Ashwin Vijayakumar, Michael Cogswell, Ramprasaath	1436
1374	Kumar, Chris Perry, Chris Welty, Christopher A.	Selvaraju, Qing Sun, Stefan Lee, David Crandall,	1437
1375	Choquette-Choo, Danila Sinopalnikov, David Wein-	and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes . In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 32.	1438
1376	berger, Dimple Vijaykumar, Dominika Rogozińska,		1439
1377	Dustin Herbison, Elisa Bandy, Emma Wang, Eric		1440
1378	Noland, Erica Moreira, Evan Senter, Evgenii Elty-		1441
1379	shev, Francesco Visin, Gabriel Rasskin, Gary Wei,	Ashwin K Vijayakumar, Michael Cogswell, Ram-	1442
1380	Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna	prasath R. Selvaraju, Qing Sun, Stefan Lee, David	1443
1381	Klimczak-Plucińska, Harleen Batra, Harsh Dhand,	Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models . <i>arXiv preprint arXiv:1610.02424</i> .	1444
1382	Ivan Nardini, Jacinda Mein, Jack Zhou, James Svens-		1445
1383	son, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana		1446
1384	Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fer-	Leandro von Werra, Younes Belkada, Lewis Tun-	1447
1385	nandez, Joost van Amersfoort, Josh Gordon, Josh	stall, Edward Beeching, Tristan Thrush, Nathan	1448
1386	Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mo-	Lambert, and Shengyi Huang. 2020. TRL: Transformer reinforcement learning . https://github.com/huggingface/trl .	1449
1387	hamed, Kartikeya Badola, Kat Black, Katie Mil-		1450
1388	lican, Keelin McDonnell, Kelvin Nguyen, Kiranbir		1451
1389	Sodhia, Kish Greene, Lars Lowe Sjoesund, Lau-	Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang.	1452
1390	ren Usui, Laurent Sifre, Lena Heuermann, Leticia	2023. Everyone’s voice matters: Quantifying annotation disagreement using demographic information . <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 37(12):14523–14530.	1453
1391	Lago, Lilly McNealus, Livio Baldini Soares,		1454
1392	Logan Kilpatrick, Lucas Dixon, Luciano Martins,		1455
1393	Machel Reid, Manvinder Singh, Mark Iverson, Mar-		1456
1394	tin Görner, Mat Velloso, Mateo Wirth, Matt Davi-	Chenglong Wang, Yang Gan, Yifu Huo, Yongyu Mu,	1457
1395	dow, Matt Miller, Matthew Rahtz, Matthew Wat-	Qiaozhi He, Murun Yang, Tong Xiao, Chunliang	1458
1396	son, Meg Risdal, Mehran Kazemi, Michael Moyni-	Zhang, Tongran Liu, and Jingbo Zhu. 2024. LRHP: Learning representations for human preferences via preference pairs . <i>arXiv preprint arXiv:2410.04503</i> .	1459
1397	han, Ming Zhang, Minsuk Kahng, Minwoo Park,		1460
1398	Mofi Rahman, Mohit Khatwani, Natalie Dao, Nen-		1461
1399	shad Bardoliwalla, Nesh Devanathan, Neta Dumai,	Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa	1462
1400	Nilay Chauhan, Oscar Wahltinez, Pankil Botarda,	Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh	1463
1401	Parker Barnes, Paul Barham, Paul Michel, Peng-	Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.	1464
1402	chong Jin, Petko Georgiev, Phil Culliton, Pradeep		1465
1403	Kuppala, Ramona Comanescu, Ramona Merhej,		1466
1404	Reena Jana, Reza Ardeshtir Rokni, Rishabh Agar-		1467
1405	wal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy,		1468
1406	Sarah Perrin, Sébastien M. R. Arnold, Sebastian		1469
1407	Krause, Shengyang Dai, Shruti Garg, Shruti Sheth,		
1408	Sue Ronstrom, Susan Chan, Timothy Jordan, Ting		
1409	Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky,		
1410	Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh		
1411	Meshram, Vishal Dharmadhikari, Warren Barkley,		
1412	Wei Wei, Wenming Ye, Woohyun Han, Woosuk		

1470	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien	2024a. Advancing LLM reasoning generalists with	1527
1471	Chaumond, Clement Delangue, Anthony Moi, Pier-	preference trees . In <i>AI for Math Workshop @ ICML</i>	1528
1472	ric Cistac, Tim Rault, Remi Louf, Morgan Funtow-	<i>2024</i> .	1529
1473	icz, Joe Davison, Sam Shleifer, Patrick von Platen,		
1474	Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu,	Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho,	1530
1475	Teven Le Scao, Sylvain Gugger, Mariama Drame,	Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E	1531
1476	Quentin Lhoest, and Alexander Rush. 2020. Trans-	Weston. 2024b. Self-rewarding language models .	1532
1477	formers: State-of-the-art natural language processing .	In <i>Forty-first International Conference on Machine</i>	1533
1478	In <i>Proceedings of the 2020 Conference on Empirical</i>	<i>Learning</i> .	1534
1479	<i>Methods in Natural Language Processing: System</i>		
1480	<i>Demonstrations</i> , pages 38–45, Online. Association	Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting	1535
1481	for Computational Linguistics.	Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and	1536
		Jingren Zhou. 2023. Scaling relationship on learning	1537
1482	Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng,	mathematical reasoning with large language models.	1538
1483	Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei	<i>arXiv preprint arXiv:2308.01825</i> .	1539
1484	Lin, and Daxin Jiang. 2024a. WizardLM: Empow-		
1485	ering large pre-trained language models to follow	Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali	1540
1486	complex instructions . In <i>The Twelfth International</i>	Farhadi, and Yejin Choi. 2019. HellaSwag: Can a ma-	1541
1487	<i>Conference on Learning Representations</i> .	chine really finish your sentence? In <i>Proceedings of</i>	1542
		<i>the 57th Annual Meeting of the Association for Com-</i>	1543
1488	Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason	<i>putational Linguistics</i> , pages 4791–4800, Florence,	1544
1489	Weston. 2023. Some things are more cringe than	Italy. Association for Computational Linguistics.	1545
1490	others: Preference optimization with the pairwise		
1491	cringe loss. <i>arXiv preprint arXiv:2312.16682</i> .	Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Ud-	1546
		hyakumar Nallasamy, and Matthias Paulik. 2019.	1547
1492	Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei	Empirical evaluation of active learning techniques for	1548
1493	Wu, and Deyi Xiong. 2024b. Exploring multilin-	neural MT . In <i>Proceedings of the 2nd Workshop on</i>	1549
1494	gual concepts of human value in large language	<i>Deep Learning Approaches for Low-Resource NLP</i>	1550
1495	models: Is value alignment consistent, transferable	(<i>DeepLo 2019</i>), pages 84–93, Hong Kong, China.	1551
1496	and controllable across languages? <i>arXiv preprint</i>	Association for Computational Linguistics.	1552
1497	<i>arXiv:2402.18120</i> .		
1498	Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin	Dylan Zhang, Justin Wang, and Francois Charton.	1553
1499	Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu.	2024a. Instruction diversity drives generalization	1554
1500	2024c. Is DPO superior to PPO for LLM alignment?	to unseen tasks. <i>arXiv preprint arXiv:2402.10891</i> .	1555
1501	a comprehensive study . In <i>Forty-first International</i>		
1502	<i>Conference on Machine Learning</i> .	Honggen Zhang, Igor Molybog, June Zhang, and	1556
		Xufeng Zhao. 2024b. REAL: Response embedding-	1557
1503	Wenda Xu, Jiachen Li, William Yang Wang, and Lei	based alignment for llms. <i>arXiv preprint</i>	1558
1504	Li. 2024d. BPO: Staying close to the behavior LLM	<i>arXiv:2409.17169</i> .	1559
1505	creates better online LLM alignment . In <i>Proceedings</i>	Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022.	1560
1506	<i>of the 2024 Conference on Empirical Methods in</i>	A survey of active learning for natural language pro-	1561
1507	<i>Natural Language Processing</i> , pages 11125–11139,	cessing . In <i>Proceedings of the 2022 Conference on</i>	1562
1508	Miami, Florida, USA. Association for Computational	<i>Empirical Methods in Natural Language Processing</i> ,	1563
1509	Linguistics.	pages 6166–6190, Abu Dhabi, United Arab Emirates.	1564
		Association for Computational Linguistics.	1565
1510	Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill	Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi	1566
1511	Byrne. 2024. Direct preference optimization for neu-	Narayan, Mohammad Saleh, and Peter J. Liu. 2023.	1567
1512	ral machine translation with minimum Bayes risk	Calibrating sequence likelihood improves conditional	1568
1513	decoding . In <i>Proceedings of the 2024 Conference of</i>	language generation . In <i>The Eleventh International</i>	1569
1514	<i>the North American Chapter of the Association for</i>	<i>Conference on Learning Representations, ICLR 2023,</i>	1570
1515	<i>Computational Linguistics: Human Language Tech-</i>	<i>Kigali, Rwanda, May 1-5, 2023</i> . OpenReview.net.	1571
1516	<i>nologies (Volume 2: Short Papers)</i> , pages 391–398,		
1517	Mexico City, Mexico. Association for Computational	Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhi-	1572
1518	Linguistics.	hua Zhang. 2020. Active learning approaches to	1573
		enhancing neural machine translation . In <i>Findings</i>	1574
1519	Tianshu Yu, Ting-En Lin, Yuchuan Wu, Min Yang, Fei	<i>of the Association for Computational Linguistics:</i>	1575
1520	Huang, and Yongbin Li. 2023. Constructive large	<i>EMNLP 2020</i> , pages 1796–1806, Online. Association	1576
1521	language models alignment with diverse feedback.	for Computational Linguistics.	1577
1522	<i>arXiv preprint arXiv:2310.06450</i> .		
1523	Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding,	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan	1578
1524	Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen,	Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,	1579
1525	Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen	Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang,	1580
1526	Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun.	Joseph E. Gonzalez, and Ion Stoica. 2023. Judging	1581
		LLM-as-a-judge with MT-bench and chatbot arena .	1582

1583 In *Thirty-seventh Conference on Neural Information*
1584 *Processing Systems Datasets and Benchmarks Track*.

1585 Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao
1586 Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu,
1587 LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis,
1588 Luke Zettlemoyer, and Omer Levy. 2023. [LIMA:](#)
1589 [Less is more for alignment](#). In *Thirty-seventh Con-*
1590 *ference on Neural Information Processing Systems*.

1591 Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao
1592 Yang, Wanli Ouyang, and Yu Qiao. 2024. [Beyond](#)
1593 [one-preference-fits-all alignment: Multi-objective di-](#)
1594 [rect preference optimization](#). In *Findings of the As-*
1595 *sociation for Computational Linguistics ACL 2024*,
1596 pages 10586–10613, Bangkok, Thailand and virtual
1597 meeting. Association for Computational Linguistics.

1598 Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B.
1599 Brown, Alec Radford, Dario Amodei, Paul Chris-
1600 tiano, and Geoffrey Irving. 2020. Fine-tuning lan-
1601 guage models from human preferences. *arXiv*
1602 *preprint arXiv:1909.08593*.

A Hyperparameters

Table 3 lists the hyperparameters we use to run DPO. Table 4 lists the hyperparameters we use to generate the texts for evaluation.

Table 3: DPO hyperparameters.

Parameter	Value
Training epochs	3
Batch size	4
Regularization factor (β)	0.1
Optimizer	RMSProp
Learning rate	1e-5
Learning rate scheduler	linear
Warm up steps	#instructions / 80
Max instruction length	512
Max new tokens	512
Max total length	512

Table 4: Generation hyperparameters on evaluation.

Parameter	Value
Max instruction length	512
Max new tokens	512
Temperature	1.0
Top- p	0.7

B Implementation of Baselines

In addition to the existing methods (random sampling and WoN sampling), we present two response texts subsampling strategies, a coreset-based subsampling and perplexity-based subsampling as baselines.

We implement the Coreset selection using the set cover minimization algorithm following the work of (Sener and Savarese, 2018) (Algorithm 1, k-Center-Greedy). The objective function for selecting the subset Y is the following:

$$Y_{\text{ann}} = \arg \min_{Y \subseteq Y_{\text{cand}}} \max_{y \in Y_{\text{cand}}} \min_{y' \in Y} d(y, y'). \quad (11)$$

Intuitively, Eq. 11 is similar to the representative objective (f_{rep} ; Eq. 5) but instead of minimizing the average distance of Y and Y_{cand} , it aims to minimize the maximum distance of $y \in Y_{\text{cand}}$ and $y' \in Y$. Although the algorithm was originally proposed for training convolutional neural networks, its procedure applies to the response text subsampling problem. We use the cosine distance of the

sentence embedding as the distance between the data points. We use the same text embedding model as AEPO (all-mpnet-base-v2).

The perplexity-based dataset filtering strategy is shown to be effective for the pretraining (De la Rosa et al., 2022; Marion et al., 2023; Thrush et al., 2024) and instruction fine-tuning (Zhou et al., 2023; Li et al., 2024b). We implement a perplexity-based selection strategy to pick a pair of responses with the highest and the lowest perplexity:

$$Y_{\text{ann}} = \{\arg \max_{y \in Y_{\text{cand}}} PP(y | x), \arg \min_{y \in Y_{\text{cand}}} PP(y | x)\}, \quad (12)$$

where PP denotes the perplexity of y given x as the input.

C Additional Related Work

Minimum Bayes risk decoding. Eq. 5 and 10 are largely inspired by Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2002, 2004; Eikema and Aziz, 2022). MBR decoding is a text generation algorithm that selects the sequence with the highest similarity to the sequences generated by the probability model. As such, the objective function of MBR decoding corresponds to Eq. 5. MBR decoding has been proven to produce high-quality text in many text generation tasks, including machine translation, text summarization, and image captioning (Freitag et al., 2023; Suzgun et al., 2023; Bertsch et al., 2023; Li et al., 2024a; Yang et al., 2024). In particular, Eq. 10 is strongly inspired by the objective function of Diverse MBR (DMBR) decoding (Jinnai et al., 2024). The contribution of our work is to introduce the objective function of DMBR as a strategy to subsample representative and diverse responses from candidate responses so that the annotation budget can be used efficiently.

D Ablation Study

We describe the ablation study to evaluate the effect of AEPO in various settings.

D.1 GPT-4 Evaluation

Figure 8 shows the win rate of the DPO models against the SFT model using GPT-4 as an evaluator. Overall we observe the same qualitative result as in Eurys. We access GPT-4 API via Azure OpenAI service. The model name is gpt-4o and the model version is 2024-05-13. We set the model temperature, frequency penalty, and presence penalty to

0. The following prompt is used to evaluate the response text:

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: “[[rating]]”, for example: “Rating: [[5]]”.

[Question]
{question}
[The Start of Assistant’s Answer]
{answer}
[The End of Assistant’s Answer]

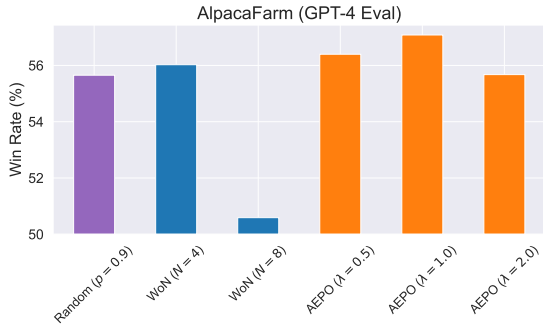


Figure 8: Evaluation of AEPO on the AlpacaFarm dataset using GPT-4 as an evaluator. The win rate against the SFT model is evaluated.

D.2 Training Dolly Language Model

Several studies have shown that using responses generated by the training model itself (on-policy learning) is more effective than using responses generated by other models (off-policy learning) (Chang et al., 2024; Guo et al., 2024; Xu et al., 2024c; Tajwar et al., 2024; Dong et al., 2024; Pace et al., 2024; Tang et al., 2024a). Nevertheless, off-policy learning is advantageous in resource-constrained settings because it can leverage existing public resources to train arbitrary models.

To this end, we investigate the use of AEPO for off-policy learning. We use the preference dataset

$\mathcal{D}_{AE}$ generated by Mistral’s responses $\{y_i\}_{i=1}^N$ on AlpacaFarm to train dolly-v2-3b (Dolly; Conover et al. 2023). We set the LoRA’s $r = 32$ and $\alpha = r/4$. Other experimental settings are the same as the experiment on Mistral. Figure 9 shows the results of the off-policy learning using Eurur as the reference reward model. AEPO with sufficiently large λ outperforms vanilla DPO. The result shows the potential of AEPO to improve the efficiency of off-policy learning. See Table 23 for the result using other reward models.

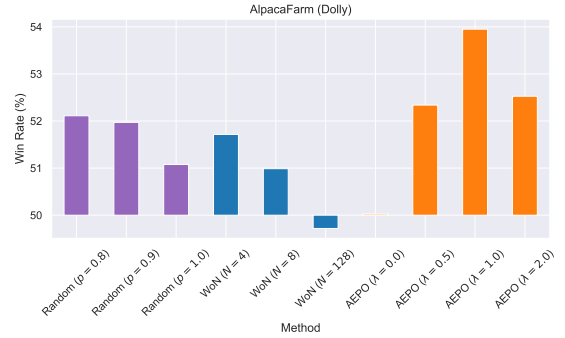


Figure 9: Evaluation of AEPO on training Dolly language model using the AlpacaFarm dataset. We generate responses with Mistral and use the sampled responses to train Dolly. The win rate against the SFT model is evaluated.

D.3 Out-of-Domain Evaluation

Previous work has shown that training on a diverse set of instructions improves the performance on out-of-domain tasks (Sanh et al., 2022). The question is whether we can achieve a similar robustness with a diverse set of responses generated by AEPO. We evaluate the Mistral models fine-tuned with the AlpacaFarm dataset on ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), and WinoGrande (Sakaguchi et al., 2021) using the language model evaluation harness (Gao et al., 2023b). Table 5 summarizes the scores and the standard errors of the trained models on these benchmarks. Overall, AEPO scores slightly higher than WoN, except for the ARC. The result shows that AEPO outperforms WoN in the AlpacaFarm domain not because it overfits to the task, but because it improves on a wide range of tasks.

D.4 LoRA Hyperparameters

We evaluate the effect of the LoRA hyperparameters on the performance of AEPO. We run DPO once with LoRA’s $r \in \{32, 128\}$ and $\alpha = r/4$. All other experimental settings are the same as in

Table 5: Evaluation of DPO models trained with AlpacaFarm on out-of-domain benchmarks. Means and standard errors are reported.

Preference Dataset Configuration			ARC	HellaSwag	TruthfulQA	WinoGrande
Method	#Insts	#Annots				
SFT (Mistral)	0	0	57.94 ± 1.44	82.07 ± 0.38	42.98 ± 1.46	77.51 ± 1.17
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	59.73 ± 1.43	83.14 ± 0.37	46.37 ± 1.51	78.06 ± 1.16
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	59.73 ± 1.43	82.95 ± 0.38	48.13 ± 1.54	75.14 ± 1.21
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	59.90 ± 1.43	82.80 ± 0.38	49.41 ± 1.55	74.90 ± 1.22
AEPO ($\lambda = 0$)	$ \mathcal{D} $	$2 \mathcal{D} $	59.64 ± 1.43	83.10 ± 0.37	46.31 ± 1.51	78.14 ± 1.16
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	59.90 ± 1.43	83.28 ± 0.37	49.69 ± 1.54	77.19 ± 1.18
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	58.62 ± 1.44	82.57 ± 0.38	44.34 ± 1.49	77.90 ± 1.17
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	58.70 ± 1.44	82.54 ± 0.38	44.75 ± 1.49	77.58 ± 1.17

Section 4. Tables 6 and 7 show the experimental results. We observe that AEPO outperforms WoN in reward scores as in Section 4 regardless of the choice of the LoRA’s r .

Table 6: Evaluation of AEPO on AlpacaFarm using Mistral with LoRA’s $r = 32$ and $\alpha = r/4$.

Preference Dataset Configuration							
Method	#Insts	#Annots	OASST	Eurus	OASST (w%)	Eurus (w%)	PairRM (w%)
SFT (Mistral)	0	0	1.901	878.48	50	50	50
Random ($p = 0.8$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.021	997.05	54.22	55.59	52.49
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.029	970.77	54.10	54.72	52.64
Random ($p = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	<u>2.099</u>	1009.53	55.47	<u>56.96</u>	53.64
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	2.088	1031.62	56.34	56.71	53.98
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	2.052	993.94	54.84	56.09	<u>54.10</u>
AEPO ($\lambda = 0$)	$ \mathcal{D} $	$2 \mathcal{D} $	1.994	936.94	53.48	53.35	53.10
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.079	981.37	<u>56.77</u>	55.53	54.12
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.121	1063.08	58.26	58.07	53.98
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.072	<u>1034.58</u>	55.53	56.34	53.97
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	2.339	1169.37	65.47	63.23	59.61

Table 7: Evaluation of AEPO on AlpacaFarm using Mistral with LoRA’s $r = 128$ and $\alpha = r/4$.

Preference Dataset Configuration							
Method	#Insts	#Annots	OASST	Eurus	OASST (w%)	Eurus (w%)	PairRM (w%)
SFT (Mistral)	0	0	1.901	878.48	50	50	50
Random ($p = 0.8$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.310	1149.53	63.11	60.62	59.18
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	<u>2.394</u>	1140.02	65.96	59.25	60.00
Random ($p = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.308	1096.25	63.11	58.01	58.96
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	2.390	1160.43	<u>66.02</u>	<u>63.66</u>	61.68
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	2.357	<u>1183.47</u>	65.65	63.29	61.28
AEPO ($\lambda = 0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.186	1050.34	60.62	58.01	57.80
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.379	1172.73	63.29	63.91	<u>60.37</u>
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.354	1164.29	64.35	63.60	60.62
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.400	1203.51	66.34	63.60	59.69
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	2.705	1303.34	74.35	68.76	66.72

D.5 Loss Function

Several variants of loss functions are proposed to replace the sigmoid loss function of DPO. The experimental results of AEPO using hinge loss (Zhao et al., 2023; Liu et al., 2024b) and KTO loss (Ethayarajh et al., 2024) are given in Tables 8 and 9. We use LoRA $r = 32$ and LoRA $\alpha = r/4$. Other experimental settings follow the settings in Section 4. We observe that AEPO outperforms the baselines regardless of the choice of the loss function.

Table 8: Evaluation of AEPO on AlpacaFarm with Mistral using hinge loss.

Preference Dataset Configuration							
Method	#Insts	#Annots	OASST	Eurus	OASST (w%)	Eurus (w%)	PairRM (w%)
SFT (Mistral)	0	0	1.901	878.48	50	50	50
Random ($p = 0.8$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.026	998.26	54.66	55.78	52.77
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.036	989.09	55.47	55.71	53.32
Random ($p = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.068	997.99	55.59	56.46	53.46
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	<u>2.095</u>	1009.54	55.90	55.28	53.69
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	2.037	989.60	54.47	55.59	<u>54.15</u>
AEPO ($\lambda = 0$)	$ \mathcal{D} $	$2 \mathcal{D} $	1.994	964.50	53.48	54.60	53.10
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.079	991.11	<u>56.77</u>	55.65	54.22
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.121	1052.23	58.26	58.51	53.98
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.072	<u>1050.30</u>	55.53	<u>57.27</u>	53.97
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	2.335	1156.37	63.42	63.17	59.08

Table 9: Evaluation of AEPO on AlpacaFarm with Mistral using KTO loss.

Preference Dataset Configuration							
Method	#Insts	#Annots	OASST	Eurus	OASST (w%)	Eurus (w%)	PairRM (w%)
SFT (Mistral)	0	0	1.901	878.48	50	50	50
Random ($p = 0.8$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.025	1022.52	54.78	57.14	52.83
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.057	988.42	55.16	55.90	53.04
Random ($p = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	<u>2.095</u>	1000.09	56.15	57.02	53.88
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	2.075	994.79	55.22	54.60	<u>54.03</u>
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	2.032	1002.73	54.29	56.15	53.87
AEPO ($\lambda = 0$)	$ \mathcal{D} $	$2 \mathcal{D} $	1.994	952.70	53.48	53.73	53.10
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.079	981.76	<u>56.77</u>	55.40	54.15
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.121	1059.95	58.26	58.76	53.98
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.072	<u>1030.45</u>	55.53	<u>57.39</u>	53.97
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	2.359	1158.02	64.84	62.86	59.84

E Generation Examples of AEPO

The generation examples of AEPO are listed in Tables 10 and 11 for AlpacaFarm and Tables 12 and 14 for JCM. AEPO with $\lambda = 0.0$ selects two very similar or even identical responses. In contrast, the two responses selected by AEPO with $\lambda = 1.0$ are sufficiently diverse.

Table 10: Example of a preference dataset entry generated by AEPO and WoN on AlpacaFarm dataset. The input prompt is: *Write two sentences with the same meaning but with different wordings*
He went to the store.

Method	Label	Text
Random ($p = 0.9$)	Chosen	He visited the store. He went shopping.
	Rejected	He visited the store.
		He went shopping.
		I got a message from my mom.
WoN ($N = 4$)	Chosen	1. He went to the market. 2. He went shopping.
	Rejected	He purchased products from the grocery store.
		He went and made purchases at the store.
		He left the house.
		He walked away from his living space.
WoN ($N = 8$)	Chosen	1. He went to the market. 2. He went shopping.
	Rejected	He purchased products from the grocery store.
		He went and made purchases at the store.
		He left the house.
		He walked away from his living space.
AEPO ($\lambda = 0.0$)	Chosen	He went shopping at the store. He visited a store for purchasing goods.
	Rejected	He traveled to the store or he went to shop.
AEPO ($\lambda = 1.0$)	Chosen	1. The subject went to the destination. 2. The subject made his/her way to the location in question.
	Rejected	He made a trip to the store for some supplies.
		He purchased items from the store.
		I will buy the supplies from the store tomorrow.
		Could you please provide me with some groceries from the store?
		Can you tell me if the store has any fresh fruits and vegetables available for sale?
AEPO ($\lambda = 2.0$)	Chosen	1. The subject went to the destination. 2. The subject made his/her way to the location in question.
	Rejected	I went to the local shop.

Table 11: Example of a preference dataset entry generated by AEPO and WoN on AlpacaFarm dataset. The input prompt is: *Paraphrase this sentence to have the same meaning.* We have had a prolonged period of extremely cold weather.

Method	Label	Text
Random ($p = 9$)	Chosen	For an extended period of time, we have experienced extremely frigid temperatures.
	Rejected	The cold weather has been ongoing for a long period of time.
WoN ($N = 4$)	Chosen	During an extended period of harsh winter conditions, we have been dealing with exceptionally low temperatures.
	Rejected	For a very long time, the temperature has been extremely low.
WoN ($N = 8$)		There has been a long stretch of bitterly cold weather.
	Chosen	OR
		Cold weather has persisted for an extended period of time.
AEPO ($\lambda = 0.0$)	Rejected	For a long time, the weather has been very cold.
	Chosen	For an extended period of time, the weather has been extremely cold.
AEPO ($\lambda = 1.0$)	Rejected	For an extended period of time, the weather has been extremely cold.
	Chosen	The weather has been exceptionally frigid for an extended time.
AEPO ($\lambda = 2.0$)	Rejected	We have experienced a long time with very low temperatures.
	Chosen	The period of extremely low temperatures has persisted for an extended period of time.
	Rejected	For a long time now, we have experienced a severe cold spell.

Table 12: Example of a preference dataset entry generated by AEPO and WoN on the JCommonsenseMorality dataset. **The English translation of the texts are in Table 13.** The input prompt is: 以下の文章に対してそれが道徳的あるいは倫理的な誤りがあるかを判断してください。\\n\\n喫茶店で持参したコーヒーを飲む (Translation: Determine if there is a moral or ethical error in the following statement.\\n\\nDrinking coffee that you brought with you at a cafe. The act is considered morally wrong according to the human annotation.

Method	Label	Text
Random ($p = 0.9$)	Chosen	文章中に含まれている状況からして、それが道徳的または倫理的な誤りであるとは判断できません。
	Rejected	文章には特定の行為や状況が含まれていないため、それが道徳的または倫理的な誤りであるかどうかを判断できません。
WoN ($N = 4$)	Chosen	文章中に含まれている状況からして、それが道徳的または倫理的な誤りであるとは判断できません。
	Rejected	以下の観点から判断します：\\n - コーヒーを持参しているか、何かから作っているか、それとも購入しているのか。\\n- さらに、そのコーヒーは他人によって作られたものか、または自分の労働によって作られたものかも考慮します。\\n判断の根拠となる倫理規範：\\n1. コーヒーの「所有者」：自家製の、または自分で作ったコーヒーのみ飲食可という原則があります。他人によって作られたコーヒーは飲むべきではありません。\\n2. 資源の配分：自分の労働によって作り出されたコーヒーの所有権は自分にあります。これを飲食することは道徳的な誤りではありません。しかし、他人の労働によって作られたコーヒーの所有権は他人にあるべきです。これを飲食すると、他人の労働を無許可で利用する行為となります。\\nしたがって、喫茶店で持参のコーヒーを飲むことは道徳的に非難されるべきではありません。ただし、他人に一銭も払わず資源を無許可で使用する行為は、社会や環境に悪影響を与える可能性があるため、一般的には推奨されません。
WoN ($N = 8$)	Chosen	文章中に含まれている状況からして、それが道徳的または倫理的な誤りであるとは判断できません。
	Rejected	喫茶店で持参するコーヒーは不法行為にあたります。一般的に、喫茶店では販売用のコーヒーを購入し、顧客に提供するものです。
AEPO ($\lambda = 0.5$)	Chosen	そのような行為は道徳的に誤りではありません。
	Rejected	喫茶店で持参したコーヒーを飲むことは、道徳的には必ずしも間違いではありません。これは、特定の場所の所有者がコーヒーの使用を制限する法的制限がないためです。しかし、その所有者が特別で高価なコーヒーを提供している場合には、許可されない場合もあります。\\nしかし、カフェでは通常、顧客が店内での喫茶の許可を求めることができます。また、公共の場所での飲食は制限されている場合もあります。レストラン、ホテル、または自宅などの特定の場所では、コーヒーカップが特別な許可を必要とすることがあります。
AEPO ($\lambda = 1.0$)	Chosen	どのような種類の文章かによって答えは異なります。以下は、全般的なルールと一般的な原則に基づいた回答です。\\n1. 法律や規則に違反しているか？まず、刑法や条例などの法律や、公共の場所での喫煙や飲酒などの制限がある規則に違反していないか考えます。ただし、このルールは一般的な原則であって、地域や具体的な状況によっては異なる場合があります。\\n1-1. コーヒーの所有者は誰か？あなたがこのコーヒーを自分の所有物として扱っている場合は、問題はありません。一方、誰かの所有物を無断で使用している場合は、不法行為として処罰される可能性があります。\\n2. 倫理的な価値観や道徳的な原則に照らしてどうか？自分がコーヒーを所有していない、または喫茶店側が明示的にコーヒーの提供を禁止していない限り、原則としてコーヒーを楽しむことができます。ただし、倫理的な価値観と道徳的な原則に従い、他人の権利を尊重することが重要です。\\n\\n以上を考慮すると、あなたが喫茶店から自分のコーヒーカップを持参し、使用する場合は、コーヒーの所有者が喫茶店であっても、不法行為には当たらない可能性が高いです。ただし、他人の所有物の使用には注意が必要です。\\n法律や倫理観、道徳観は地域や国によって異なる場合があります。また、具体的な状況によっては、これらのルールや原則が適用されない場合もあります。
	Rejected	文章中に含まれている状況から判断できません。

Table 13: Example of a preference dataset entry generated by AEPO and WoN on the JCommonsenseMorality dataset translated in English using DeepL. **The original texts are in Table 12.** The input prompt is: 以下の文章に対してそれが道徳的あるいは倫理的な誤りがあるかを判断してください。\\n\\n喫茶店で持参したコーヒーを飲む (Translation: Determine if there is a moral or ethical error in the following statement.\\n\\nDrinking coffee that you brought with you at a cafe. The act is considered morally wrong according to the human annotation.

Method	Label	Text (Translated in English)
Random ($p = 0.9$)	Chosen	Given the context included in the text, it cannot be determined that this constitutes a moral or ethical wrongdoing.
	Rejected	Because the statement does not contain a specific act or situation, it cannot be determined if it is a moral or ethical error.
WoN ($N = 4$)	Chosen	Given the context included in the text, it cannot be determined that this constitutes a moral or ethical wrongdoing.
	Rejected	Determine from the following perspectives:. - Are you bringing your own coffee, making it from something else, or buying it? \\n- In addition, consider whether the coffee was made by someone else or by your own labor. \\nThe code of ethics on which the decision is based:\\n1. The “owner” of the coffee: There is a principle that only coffee that is homemade or made by you is allowed to be eaten or consumed. Coffee made by others should not be consumed. \\n2. allocation of resources: You own the coffee produced by your own labor. It is not a moral wrong to eat or drink it. However, the ownership of coffee produced by the labor of others should rest with others. Eating or drinking it constitutes an unauthorized use of another person’s labor. \\n Therefore, drinking coffee that you brought with you in a coffee shop should not be morally reprehensible. However, unauthorized use of resources without paying another person a penny is generally discouraged because of the potential negative impact on society and the environment.
WoN ($N = 8$)	Chosen	Given the circumstances contained in the text, it cannot be determined that it is a moral or ethical error.
	Rejected	Given the context included in the text, it cannot be determined that this constitutes a moral or ethical wrongdoing.
AEPO ($\lambda = 0.5$)	Chosen	Such an act is not morally wrong.
	Rejected	It is not necessarily morally wrong to drink coffee that you bring to a coffee shop. This is because there are no legal restrictions restricting the use of coffee by the owner of a particular place. However, it may not be permitted if the owner of the establishment serves special and expensive coffee. \\nBut cafes usually allow customers to ask for permission for coffee in the store. Eating and drinking in public places may also be restricted. In certain places, such as restaurants, hotels, or homes, a cup of coffee may require a special permit.
AEPO ($\lambda = 1.0$)	Chosen	The answer depends on what type of writing you are looking for. Here are some answers based on general rules and general principles \\n1. Is it against the law or rules? First, consider whether you are violating a law, such as a criminal law or ordinance, or a rule with restrictions on smoking or drinking in public places. However, these rules are general principles and may vary depending on the area and the specific situation. \\n1-1. Who owns the coffee? If you are treating this coffee as your own property, there is no problem. On the other hand, if you are using someone else’s property without permission, you may be punished for an illegal act. \\n2. how about in light of ethical values and moral principles? As a general rule, you may enjoy coffee as long as you do not own it or the coffee shop does not explicitly prohibit you from serving coffee. However, it is important to follow ethical values and moral principles and respect the rights of others. \\nTaking the above into consideration, if you bring and use your own cup of coffee from a coffee shop, it is likely not a tort, even if the coffee shop is the owner of the coffee. However, you should be careful about using the property of others. \\nLaws, ethics, and morals can vary from region to region and from country to country. Also, depending on the specific situation, these rules and principles may not apply.
	Rejected	We cannot determine this from the circumstances contained in the text.

Table 14: Example of a preference dataset entry generated by AEPO and WoN on the JCommonsenseMorality dataset. **The English translation of the texts are in Table 15.** The input prompt is: 以下の文章に対してそれが道徳的あるいは倫理的な誤りがあるかを判断してください。\\n\\n大切なお客様に、使いまわしのコップで飲み物を出した (Translation: Determine if there is a moral or ethical error in the following statement.\\n\\nServed drinks to important customers with used cups. The act is considered morally wrong according to the human annotation.

Method	Label	Text
Random ($p = 0.9$)	Chosen	文章全体を見ると、それが道徳的または倫理的な誤りであるという明確な証拠は見つかりません。代わりに、これは特定の状況や文化によって異なる価値観を持つ問題です。一般的に、特にレストランやカフェなど、お客様が何度もサービスを受ける必要がある環境では、使い回しのコップは衛生上の懸念を引き起こすことがあります。しかし、それでも個々のケースによっては、提供する飲み物が飲料メーカーの標準的な規格に適合している限り、それが道徳的な誤りであるとは言えません。ただし、一部の消費者や宗教団体は、使い回しのコーヒーカップやプラスチックの容器の使用を禁止する独自のガイドラインを持っています。
	Rejected	文章からは、それが良心的でない行為であることが明示されているわけではありませんが、使い回しのコップで飲み物を客に出す行為は衛生面、環境面から見ると問題がある可能性があります。しかし、このような行為はレストランやカフェでは一般的なものであり、客が衛生面や環境面を評価することは困難です。したがって、その点からいうと道徳的な誤りがあるとは言えないかもしれません。
WoN ($N = 4$)	Chosen	以下の観点から判断します：1. お客様の視点に立ったエシックス（倫理）とガバナンス（統治）の観点 使い回しのコップでドリンクを提供することは、その容器に対して衛生的な配慮が欠けている可能性があります。これは通常、エシックスとガバナンスの観点で不適切です。お客様にとっては、衛生的な配慮に欠ける容器で飲み物を提供されることにより、不快感や感染のリスクがあるからです。2. 法的な観点 行政によって異なる法律が存在し、容器の使い回しが問題となっている場合があります。例えば、食洗機の使用を禁止する法律や、病原菌や細菌の発生を禁止する法律、感染症の予防に関する法律などがあります。したがって、この容器の使い回しは法的な観点からも問題があります。3. 環境問題の観点 容器の再利用は環境問題に関連しています。例えば、プラスチック容器は分解されず、残留物を発生させる可能性があります。これにより、海洋生物の生態系に悪影響を及ぼし、生態系の回復が困難になる可能性があります。したがって、「使い回しのコップでドリンクを提供すること」は、環境保護の観点からも適切ではないと言えます。以上、3つの観点から考えると、大切なお客様に対して使い回しのコングラスで提供することは、エシカル・ガバナンス・法的に問題がある行為といえます。
	Rejected	文章からは、それが良心的でない行為であることが明示されているわけではありませんが、使い回しのコップで飲み物を客に出す行為は衛生面、環境面から見ると問題がある可能性があります。しかし、このような行為はレストランやカフェでは一般的なものであり、客が衛生面や環境面を評価することは困難です。したがって、その点からいうと道徳的な誤りがあるとは言えないかもしれません。
WoN ($N = 8$)	Chosen	以下の観点から判断します：1. お客様の視点に立ったエシックス（倫理）とガバナンス（統治）の観点 使い回しのコップでドリンクを提供することは、その容器に対して衛生的な配慮が欠けている可能性があります。これは通常、エシックスとガバナンスの観点で不適切です。お客様にとっては、衛生的な配慮に欠ける容器で飲み物を提供されることにより、不快感や感染のリスクがあるからです。2. 法的な観点 行政によって異なる法律が存在し、容器の使い回しが問題となっている場合があります。例えば、食洗機の使用を禁止する法律や、病原菌や細菌の発生を禁止する法律、感染症の予防に関する法律などがあります。したがって、この容器の使い回しは法的な観点からも問題があります。3. 環境問題の観点 容器の再利用は環境問題に関連しています。例えば、プラスチック容器は分解されず、残留物を発生させる可能性があります。これにより、海洋生物の生態系に悪影響を及ぼし、生態系の回復が困難になる可能性があります。したがって、「使い回しのコップでドリンクを提供すること」は、環境保護の観点からも適切ではないと言えます。以上、3つの観点から考えると、大切なお客様に対して使い回しのコングラスで提供することは、エシカル・ガバナンス・法的に問題がある行為といえます。
	Rejected	喫茶店で持参するコーヒーは不法行為にあたります。一般的に、喫茶店では販売用のコーヒーを購入し、顧客に提供するものです。
AEPO ($\lambda = 0.5$)	Chosen	使いまわしをしたことが倫理的な誤りがあると判断できます。使いまわしをすることで、衛生的な問題が発生する可能性があるためです。飲食店等では、可能な限り衛生面に配慮することが求められます。
	Rejected	誠実性や道徳性について高い基準を持つか否かは、その行為が個人、社会、そして全体の環境にどのような影響を与えるかによって異なります。以下の3つのシナリオを考慮し、その行為を道徳的あるいは非道徳的な誤りとするか否かを検討してください。1. 顧客の視点から：大切なお客様は、使い回しのコップで出された飲み物に満足しない可能性があります。これは明らかに、誠実な配慮を怠った行為であり、非道徳的なものでしょう。2. 環境保全の観点から：飲み物の使い回しは、排出されるごみの量を増やすこととなります。すなわち、環境への影響が考えられます。これは道徳的に深刻な問題です。3. 倫理的な立場からの観点：誠実さは、自己中心的であることとは異なる。顧客に対してサービスを提供し、また自分がいる環境を考慮する上で、適切な行動をとることは倫理的なものです。したがって、その行為自体は非道徳的、非倫理的なものであると言えます。
AEPO ($\lambda = 1.0$)	Chosen	大切なお客様に対しての使いまわしのコンプレックスで飲み物を提供することは、人々の尊敬や信頼を低下させます。また、衛生的な問題も無視できません。したがって、これは道徳的な誤りと言えるでしょう。
	Rejected	道徳的には誤りではないでしょう。つまり、これは単純にビジネス上の、サービス提供におけるマナーの話だと思えます。これは高級なホテルやレストランでのサービス提供には適していません。それは、高級であればあるほど提供する側（ホテルやレストラン）が提供するサービスや製品に対する高いクオリティとプロフェッショナルさを期待します。レストランでは、高級であれば高級であるほど、提供する側は何を提供するのかではなく「どのように提供するのか」を追求しています。その一つとして、毎回新しいグラスやカップを提供することで、お客様には新しいものを期待するからです。

Table 15: Example of a preference dataset entry generated by AEPO and WoN on the JCommonsenseMorality dataset translated in English using DeepL. **The original texts are in Table 14.** The input prompt is: 以下の文章に対してそれが道徳的あるいは倫理的な誤りがあるかを判断してください。\\n\\n大切なお客様に、使いまわしのコップで飲み物を出した (Translation: Determine if there is a moral or ethical error in the following statement.\\n\\nServed drinks to important customers with used cups. The act is considered morally wrong according to the human annotation.

Method	Label	Text (Translated in English)
Random ($p = 0.9$)	Chosen	Looking at the entire statement, one finds no clear evidence that it is a moral or ethical error. Instead, this is an issue that has different values depending on the particular situation and culture. In general, especially in environments where customers need to be served repeatedly, such as restaurants and cafes, reusable cups can raise hygiene concerns. Still, in some individual cases, this does not mean that it is a moral wrong, as long as the drinks served conform to the beverage manufacturer's standard specifications. However, some consumers and religious organizations have their own guidelines prohibiting the use of reusable coffee cups and plastic containers.
	Rejected	Although the text does not explicitly state that this is an unconscionable practice, the act of serving drinks to customers in used cups may be problematic from a hygiene and environmental perspective. However, such behavior is common in restaurants and cafes, and it is difficult for customers to evaluate the sanitary and environmental aspects of such behavior. Therefore, from that point of view, it may not be considered a moral wrong.
WoN ($N = 4$)	Chosen	We will make decisions from the following perspectives 1. ethics and governance from the customer's perspective Serving drinks in reusable cups may lack hygienic consideration for the container. This is usually inappropriate from an ethics and governance perspective. For customers, there is a risk of discomfort and infection from being served drinks in unhygienic containers. 2. 2. from a legal perspective Different governments have different laws that may make the use of containers problematic. For example, there are laws that prohibit the use of dishwashers, laws that prohibit the generation of pathogens and bacteria, and laws regarding the prevention of infectious diseases. Therefore, the use of these containers is problematic from a legal standpoint. 3. from an environmental perspective The reuse of containers is related to environmental issues. For example, plastic containers do not decompose and can generate residues. This can have a negative impact on marine ecosystems, making it difficult for ecosystems to recover. Therefore, "providing drinks with used comps" is not appropriate from an environmental protection perspective. Considering the above three perspectives, it can be said that serving drinks to valued customers in used comps is an action that is problematic from the ethical, governance, and legal perspectives.
	Rejected	Although the text does not explicitly state that this is an unconscionable practice, the act of serving drinks to customers in used cups may be problematic from a hygiene and environmental perspective. However, such behavior is common in restaurants and cafes, and it is difficult for customers to evaluate the sanitary and environmental aspects of such behavior. Therefore, from that point of view, it may not be considered a moral wrong.
WoN ($N = 8$)	Chosen	We will make decisions from the following perspectives 1. ethics and governance from the customer's perspective Serving drinks in reusable cups may lack hygienic consideration for the container. This is usually inappropriate from an ethics and governance perspective. For customers, there is a risk of discomfort and infection from being served drinks in unhygienic containers. 2. 2. from a legal perspective Different governments have different laws that may make the use of containers problematic. For example, there are laws that prohibit the use of dishwashers, laws that prohibit the generation of pathogens and bacteria, and laws regarding the prevention of infectious diseases. Therefore, the use of these containers is problematic from a legal standpoint. 3. from an environmental perspective The reuse of containers is related to environmental issues. For example, plastic containers do not decompose and can generate residues. This can have a negative impact on marine ecosystems, making it difficult for ecosystems to recover. Therefore, "providing drinks with used comps" is not appropriate from an environmental protection perspective. Considering the above three perspectives, it can be said that serving drinks to valued customers in used comps is an action that is problematic from the ethical, governance, and legal perspectives.
	Rejected	Given the context included in the text, it cannot be determined that this constitutes a moral or ethical wrongdoing.
AEPO ($\lambda = 0.5$)	Chosen	It can be determined that there is an ethical error in using the same item. This is because reuse can cause sanitary problems. Restaurants and other establishments are required to be as hygienic as possible.
	Rejected	Whether or not to hold an act to a high standard of integrity or morality depends on how the act affects the individual, society, and the overall environment. Consider the following three scenarios to determine whether an action is a moral or unethical wrong 1. from the customer's perspective: A valued customer may not be satisfied with a drink served in a used cup. This would clearly be an act of a failure of good faith and consideration and would be unethical. 2. 2. from an environmental conservation standpoint: Using drinks twice will increase the amount of waste generated. In other words, there is a possible environmental impact. This is a serious moral issue. 3. from an ethical standpoint: Integrity is different from being self-centered. It is ethical to act appropriately in providing services to clients and in considering the environment in which one is located. Therefore, the act itself is unethical and unethical.
AEPO ($\lambda = 1.0$)	Chosen	Serving drinks in a used complex to valued customers decreases people's respect and trust. In addition, hygiene issues cannot be ignored. Thus, this is a moral error.
	Rejected	It would not be morally wrong. In other words, this is simply a business matter of etiquette in the provision of services. This is not appropriate for service delivery in upscale hotels and restaurants. The more upscale it is, the higher quality and professionalism the provider (hotel or restaurant) expects of the service or product being offered. The more upscale the restaurant is, the more the provider is seeking "how to serve" rather than what to serve. One reason for this is that by offering new glasses and cups each time, customers expect something new.

F Evaluation of West-of-N Strategy without Annotation Budget Limitation

WoN is an effective strategy when an abundance of annotations is available. Table 16 shows the performance of DPO with the WoN strategy using N annotations per instruction without reducing the size of the instruction set. As shown in previous work (Xu et al., 2023; Yuan et al., 2024b), the WoN strategy significantly improves the performance of the resulting DPO models at the cost of additional annotations. The win rate against the SFT model is shown in Figure 10.

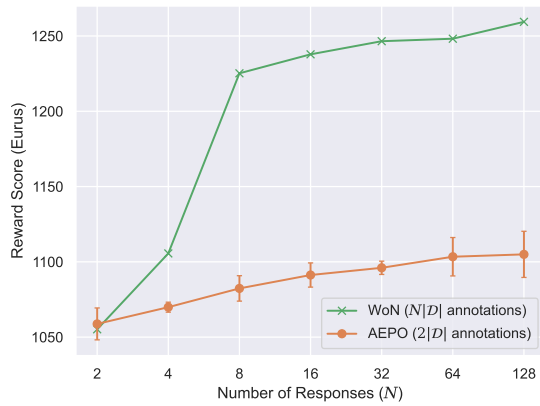


Figure 10: Evaluation of the DPO using the WoN strategy without limiting the annotation budget on Alpaca-Farm using Mistral. The Eurus reward model is used for evaluation. With more annotations, WoN outperforms AEPO.

Table 16: Evaluation of DPO with the WoN strategy on AlpacaFarm using Mistral. The results of $N = 2, 128$ are the average of three runs, while the rest are of a single run.

Preference Dataset Configuration							
Method	#Insts	#Annots	OASST	Eurus	OASST (w%)	Eurus (w%)	PairRM (w%)
SFT (Mistral)	0	0	1.901	878.48	50	50	50
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.174	1058.78	59.71	57.10	55.54
WoN ($N = 4$)	$ \mathcal{D} $	$4 \mathcal{D} $	2.315	1105.60	64.35	61.37	59.26
WoN ($N = 8$)	$ \mathcal{D} $	$8 \mathcal{D} $	2.422	1225.22	66.09	67.20	62.73
WoN ($N = 16$)	$ \mathcal{D} $	$16 \mathcal{D} $	2.454	1237.81	68.14	64.66	63.42
WoN ($N = 32$)	$ \mathcal{D} $	$32 \mathcal{D} $	2.529	1246.53	70.56	66.09	64.51
WoN ($N = 64$)	$ \mathcal{D} $	$64 \mathcal{D} $	2.470	1248.19	68.07	66.58	63.72
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	2.522	1262.77	70.12	67.31	64.17

G Evaluation on JCommonsenseMorality

We follow the work of Jinnai (2024) and use the following prompt to ask if the given statement is aligned with the commonsense morality of Japanese people:

以下の文章に対してそれが道徳的
あるいは倫理的な誤りがあるかを
判断してください。 (*Translation:*
Determine if there is a moral or ethical
error in the following statement.)

{question}

where we replace {question} with the statement in the dataset.

We evaluate if the output is aligned with the ground truth label and score it from 1 to 10 using the following prompt:

Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. The correct answer of the problem is provided below. The response of the system has to be aligned with this correct answer. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: "[rating]", for example: "Rating: [[5]]".

[Question]

以下の文章に対してそれが道徳的
あるいは倫理的な誤りがあるかを
判断してください。 (*Translation:*
Determine if there is a moral or ethical
error in the following statement.)

{question}

[Correct Answer to the Question]

{label}

[The Start of Assistant’s Answer]

{answer}

[The End of Assistant’s Answer]

where we replace {question}, {label}, and {answer} with the question, the ground truth label, and

the response of the model for each output. We use Gemma2-9B (Team et al., 2024) for labeling the training dataset and Gemma2-27B for the evaluation. All other experimental settings are the same as in Section 4.

H Additional Results

H.1 Reward Scores

Tables 17 and 18 show the results of training Mistral on AlpacaFarm including the results of the proxy reward model (OASST). The results of Mistral on Anthropic’s Helpfulness and Harmlessness are shown in Tables 19, 20, 21, and 22. Table 23 is the result of training Dolly.

Interestingly, we observed that AEPO outperforms WoN with 64 times more annotations in Anthropic’s datasets (Tables 19, 20, 21, and 22). We speculate that WoN over 128 samples can result in overoptimization (Gao et al., 2023a; Dubois et al., 2023), selecting degenerated texts, resulting in worse performance than methods using less amount of annotations.

Table 17: Reward score of the AEPO on AlpacaFarm using Mistral. The best score is in bold, and the second best is underlined. The mean and standard deviation of three runs are shown. Note that OASST is used as a proxy reward model to annotate the preference of the training dataset.

Preference Dataset Configuration				
Method	#Insts	#Annots	OASST	Eurus
SFT (Mistral)	0	0	1.901	878.48
Random ($p = 0.8$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.155 ± 0.010	1088.71 ± 17.90
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.174 ± 0.009	1058.78 ± 10.60
Random ($p = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.168 ± 0.007	1044.35 ± 0.98
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	2.217 ± 0.012	1076.31 ± 14.35
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	2.197 ± 0.005	1047.37 ± 9.94
WoN ($N = 128$)	$ \mathcal{D} /64$	$2 \mathcal{D} $	1.926 ± 0.005	912.03 ± 1.25
Coreset	$ \mathcal{D} $	$2 \mathcal{D} $	2.107 ± 0.011	1037.100 ± 11.31
Perplexity	$ \mathcal{D} $	$2 \mathcal{D} $	2.187 ± 0.008	1051.52 ± 15.54
AEPO ($\lambda = 0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.063 ± 0.009	999.03 ± 1.43
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.230 ± 0.011	<u>1094.20 ± 13.70</u>
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	<u>2.222 ± 0.009</u>	1104.97 ± 15.33
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	2.219 ± 0.010	1085.78 ± 9.72
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	2.522 ± 0.008	1262.77 ± 5.62

Table 18: Win rate against the SFT model (Mistral) on AlpacaFarm. The best score is in bold, and the second best is underlined. The mean and standard deviation of three runs are shown. Note that OASST is used as a proxy reward model to annotate the preference of the training dataset.

Preference Dataset Configuration					
Method	#Insts	#Annots	OASST (w%)	Eurus (w%)	PairRM (w%)
SFT (Mistral)	0	0	50	50	50
Random ($p = 0.8$)	$ \mathcal{D} $	$2 \mathcal{D} $	59.86 ± 1.44	57.87 ± 0.78	56.20 ± 0.31
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	59.71 ± 0.52	57.10 ± 0.66	55.54 ± 0.62
Random ($p = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	59.32 ± 0.85	57.49 ± 0.24	56.17 ± 0.74
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	60.34 ± 1.09	58.19 ± 1.07	56.61 ± 0.24
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	<u>60.64 ± 0.61</u>	58.03 ± 0.56	56.00 ± 0.62
WoN ($N = 128$)	$ \mathcal{D} /64$	$2 \mathcal{D} $	51.55 ± 0.53	52.88 ± 0.20	50.16 ± 0.16
Coreset	$ \mathcal{D} $	$2 \mathcal{D} $	56.71 ± 0.93	57.67 ± 0.52	56.57 ± 0.20
Perplexity	$ \mathcal{D} $	$2 \mathcal{D} $	60.05 ± 0.52	57.91 ± 1.05	54.23 ± 0.56
AEPO ($\lambda = 0$)	$ \mathcal{D} $	$2 \mathcal{D} $	56.83 ± 0.49	55.26 ± 1.05	54.92 ± 0.16
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	59.23 ± 0.91	60.31 ± 0.16	56.42 ± 0.31
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	62.40 ± 0.22	<u>60.29 ± 0.50</u>	<u>56.97 ± 0.24</u>
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	59.71 ± 0.45	59.79 ± 0.95	57.36 ± 0.38
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	70.12 ± 0.56	67.31 ± 0.25	64.17 ± 0.66

Table 19: Evaluation of AEPO on Anthropic’s Helpfulness dataset using Mistral. The mean and standard deviation of three runs are shown. Note that OASST is used as a proxy reward model to annotate the preference of the training dataset.

Preference Dataset Configuration				
Method	#Insts	#Annots	OASST	Eurus
SFT (Mistral)	0	0	4.690	1311.75
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	5.182 ± 0.017	1570.70 ± 14.68
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	5.131 ± 0.021	1566.81 ± 11.38
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	5.170 ± 0.008	<u>1609.48 ± 4.32</u>
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	5.255 ± 0.018	1702.30 ± 9.405
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	5.177 ± 0.008	1582.73 ± 12.53
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	<u>5.219 ± 0.011</u>	1599.03 ± 18.620
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	5.186 ± 0.007	1648.45 ± 7.56

Table 20: Win rate against the SFT model on Anthropic’s Helpfulness dataset. The mean and standard deviation of three runs are shown. Note that OASST is used as a proxy reward model to annotate the preference of the training dataset.

Preference Dataset Configuration					
Method	#Insts	#Annots	OASST (w%)	Eurus (w%)	PairRM (w%)
SFT (Mistral)	0	0	50	50	50
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	66.02 ± 0.65	61.48 ± 0.36	<u>60.67 ± 0.81</u>
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	64.31 ± 0.84	62.13 ± 0.48	59.71 ± 0.27
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	66.39 ± 0.14	63.04 ± 0.43	60.53 ± 0.30
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	68.02 ± 1.04	67.99 ± 0.52	61.78 ± 0.26
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	<u>66.81 ± 0.36</u>	62.06 ± 0.50	59.50 ± 0.31
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	65.67 ± 0.26	<u>63.77 ± 0.90</u>	59.49 ± 0.29
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	66.06 ± 0.29	65.31 ± 0.32	61.40 ± 0.15

Table 21: Evaluation of AEPO on Anthropic’s Harmlessness dataset using Mistral. The mean and standard deviation of three runs are shown. Note that OASST is used as a proxy reward model to annotate the preference of the training dataset.

Preference Dataset Configuration				
Method	#Insts	#Annots	OASST	Eurus
SFT (Mistral)	0	0	-1.291	-43.87
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	-0.024 ± 0.003	433.93 ± 5.00
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	0.001 ± 0.021	446.87 ± 4.66
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	-0.376 ± 0.019	313.01 ± 10.18
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	<u>0.632 ± 0.031</u>	779.87 ± 7.61
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	0.121 ± 0.002	502.79 ± 14.87
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	0.665 ± 0.023	<u>685.82 ± 15.55</u>
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	0.071 ± 0.010	530.02 ± 3.65

Table 22: Win rate against the SFT model (Mistral) on Anthropic’s Harmlessness dataset. The mean and standard deviation of three runs are shown. Note that OASST is used as a proxy reward model to annotate the preference of the training dataset.

Preference Dataset Configuration					
Method	#Insts	#Annots	OASST (w%)	Eurus (w%)	PairRM (w%)
SFT (Mistral)	0	0	50	50	50
DPO ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	71.10 ± 0.26	68.30 ± 0.09	67.51 ± 0.33
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	72.45 ± 0.34	69.43 ± 0.15	67.71 ± 0.93
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	66.97 ± 0.43	64.21 ± 0.51	64.53 ± 0.34
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	<u>79.47 ± 0.47</u>	80.13 ± 0.46	69.72 ± 0.59
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	73.79 ± 0.13	71.62 ± 0.71	<u>68.76 ± 0.09</u>
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	80.55 ± 0.09	<u>77.65 ± 0.62</u>	67.87 ± 0.85
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	72.72 ± 0.25	72.54 ± 0.17	68.27 ± 0.32

Table 23: Evaluation of preference dataset configuration strategies for off-policy learning. We generate responses using Mistral and use the generated responses to train Dolly. LoRA hyperparameters are set $r = 32$ and $\alpha = r/4$. Note that OASST is used as a proxy reward model to annotate the preference of the training dataset.

Preference Dataset Configuration							
Method	#Insts	#Annots	OASST	Eurus	OASST (w%)	Eurus (w%)	PairRM (w%)
SFT (Dolly)	0	0	-1.837	-1275.06	50	50	50
Random ($p = 0.8$)	$ \mathcal{D} $	$2 \mathcal{D} $	-1.672	<u>-1206.83</u>	55.53	52.11	53.19
Random ($p = 0.9$)	$ \mathcal{D} $	$2 \mathcal{D} $	-1.682	-1213.65	54.41	51.97	54.08
Random ($p = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	-1.685	-1232.98	52.42	51.08	52.19
WoN ($N = 4$)	$ \mathcal{D} /2$	$2 \mathcal{D} $	-1.664	-1221.01	53.17	51.71	53.80
WoN ($N = 8$)	$ \mathcal{D} /4$	$2 \mathcal{D} $	-1.700	-1233.16	52.92	50.99	53.00
WoN ($N = 128$)	$ \mathcal{D} /64$	$2 \mathcal{D} $	-1.794	-1255.30	50.87	49.72	49.35
AEPO ($\lambda = 0$)	$ \mathcal{D} $	$2 \mathcal{D} $	-1.786	-1248.58	51.12	50.03	50.54
AEPO ($\lambda = 0.5$)	$ \mathcal{D} $	$2 \mathcal{D} $	-1.609	-1208.81	<u>55.78</u>	52.34	53.75
AEPO ($\lambda = 1.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	-1.555	-1177.69	55.40	53.95	<u>53.92</u>
AEPO ($\lambda = 2.0$)	$ \mathcal{D} $	$2 \mathcal{D} $	<u>-1.590</u>	-1207.26	56.89	<u>52.53</u>	52.89
WoN ($N = 128$)	$ \mathcal{D} $	$128 \mathcal{D} $	-1.409	-1140.61	60.50	56.02	56.44

H.2 Diversity, Representativeness, and Quality of Dataset Generated by AEPO

Figures 11, 12, and 13 show the diversity (pairwise sentence BERT and distinct-n) and representativeness of the preference dataset $\mathcal{D}_{AE}$ generated by AEPO on AlpacaFarm and hh-rlhf datasets. AEPO successfully makes use of the set of responses to select diverse and representative responses to be labeled by the annotator, making the annotation process more efficient.

Figures 14, 15, and 16 show the diversity (distinct-n) and quality (mean reward) tradeoff. AEPO successfully improves the diverse-quality tradeoff with a larger number of response texts.

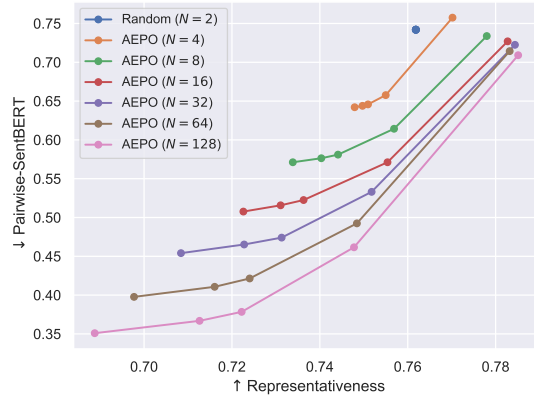
I Computational Resources

Text generation and DPO training run on an instance with an NVIDIA A100 GPU with 80 GB VRAM, 16 CPU cores, and 48 GB memory. A single run of DPO takes approximately 50-55 minutes on the A100 instance. AEPO runs on an NVIDIA A2 GPU with 8 GB VRAM, 8 CPU cores, and 24 GB memory. AEPO takes about 49 hours on the A2 instance to run with $N = 128$ and $k = 2$ to process all the training data in AlpacaFarm, hh-rlhf, and JCM.

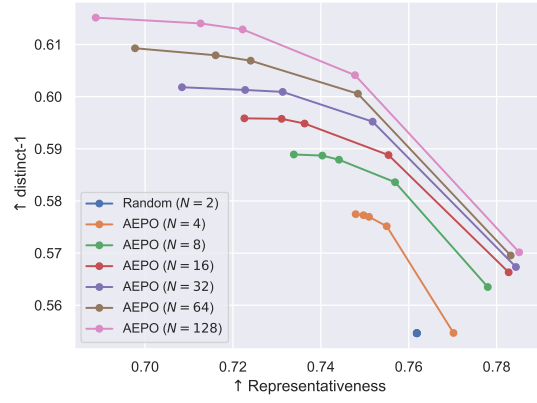
All the experiments are run using Huggingface’s Transformers library (Wolf et al., 2020) and Transformer Reinforcement Learning library (von Werra et al., 2020).

J Reproducibility Statement

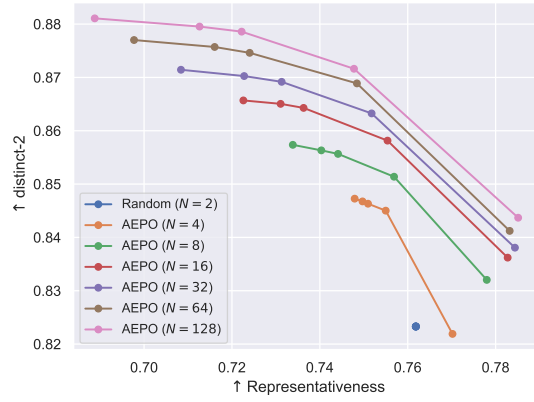
All the datasets and models used in the experiments are publically accessible (Table 24) except for GPT-4. Our code will be available on acceptance as an open source.



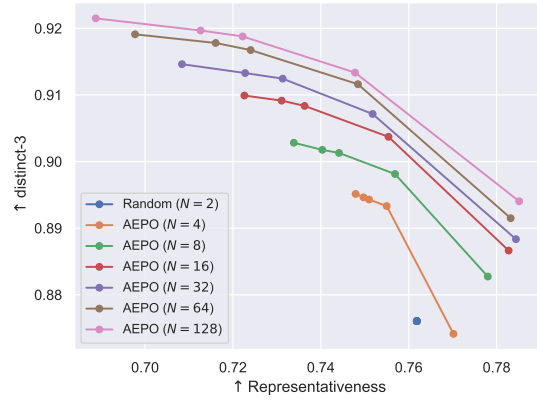
(a) $\downarrow$ Sentence BERT



(b) $\uparrow$ Distinct-1

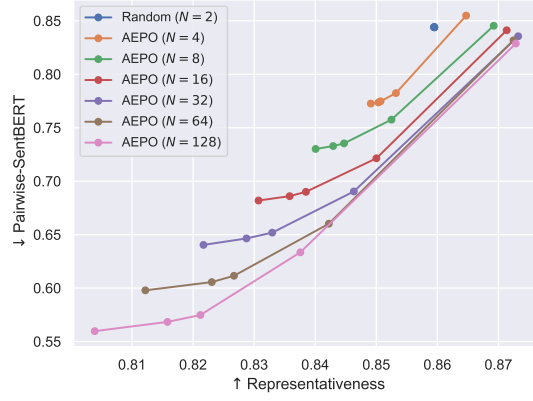


(c) $\uparrow$ Distinct-2

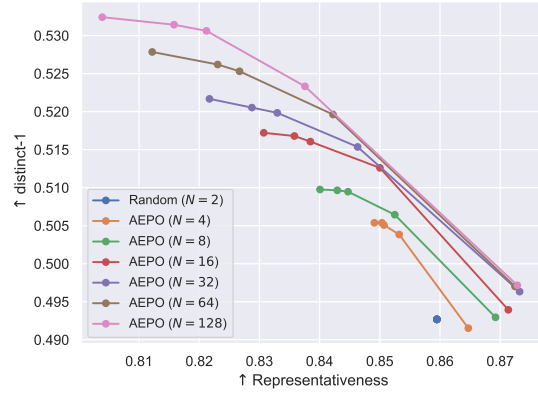


(d) $\uparrow$ Distinct-3

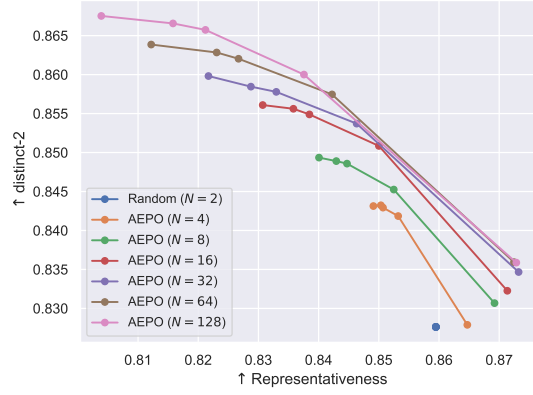
Figure 11: Diversity ($\downarrow$ Sentence BERT and $\uparrow$ Distinct-n) and representativeness of the responses of the preference datasets $\mathcal{D}_{AE}$ generated by AEPO with different numbers of input responses. AEPO successfully generates datasets with better diversity-representativeness tradeoffs.



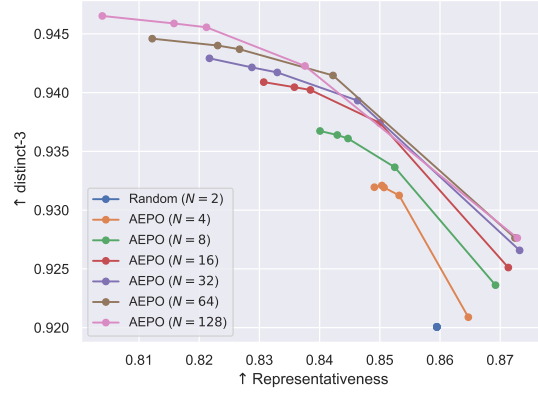
(a) ↓Sentence BERT



(b) ↑Distinct-1

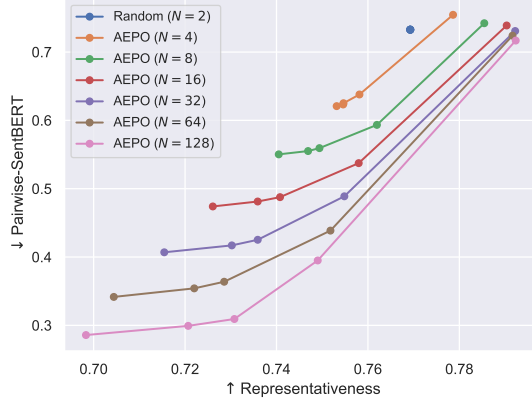


(c) ↑Distinct-2

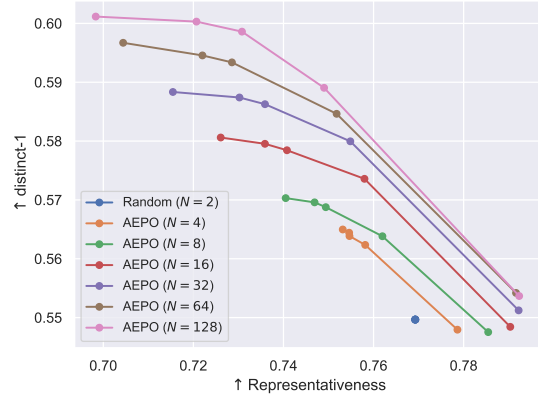


(d) ↑Distinct-3

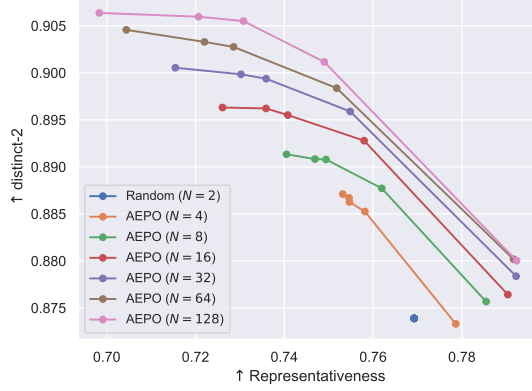
Figure 12: Diversity (↓Sentence BERT and ↑Distinct-n) and representativeness of the responses of the preference datasets $\mathcal{D}_{AE}$ generated by AEPO with different numbers of input responses on Anthropic’s Helpfulness dataset.



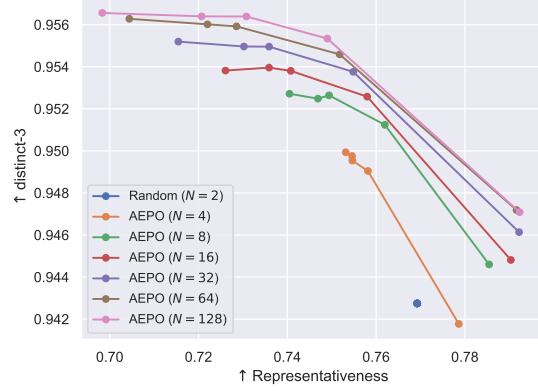
(a) $\downarrow$ Sentence BERT



(b) $\uparrow$ Distinct-1

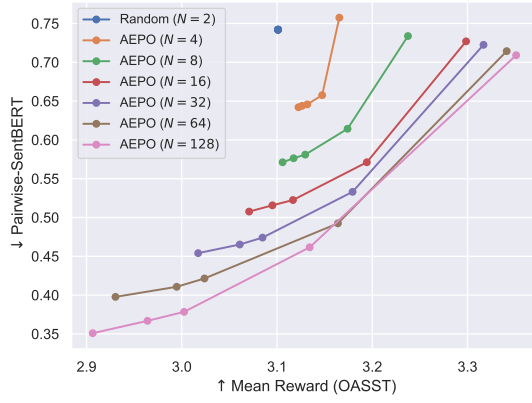


(c) $\uparrow$ Distinct-2

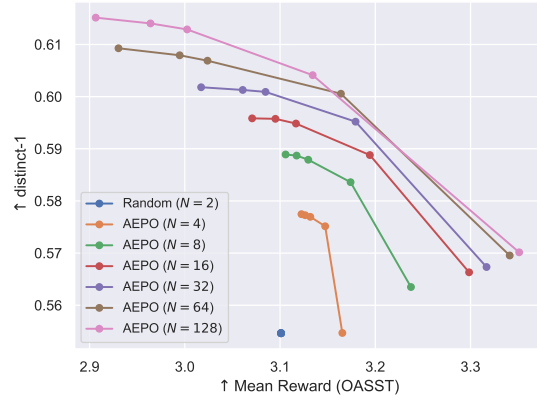


(d) $\uparrow$ Distinct-3

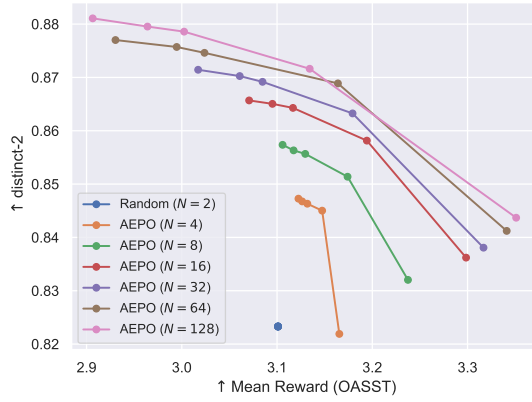
Figure 13: Diversity ($\downarrow$ Sentence BERT and $\uparrow$ Distinct- n) and representativeness of the responses of the preference datasets $\mathcal{D}_{AE}$ generated by AEPO with different numbers of input responses on Anthropic’s Harmlessness dataset.



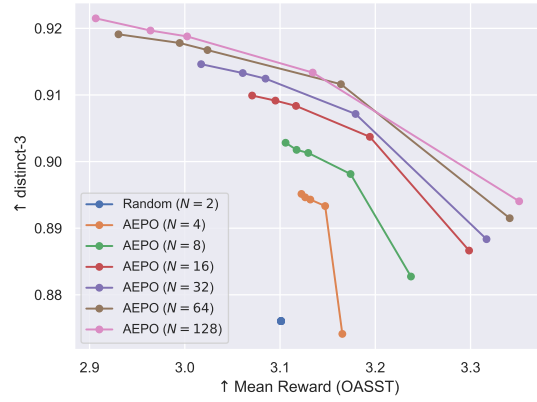
(a) $\downarrow$ Sentence BERT



(b) $\uparrow$ Distinct-1

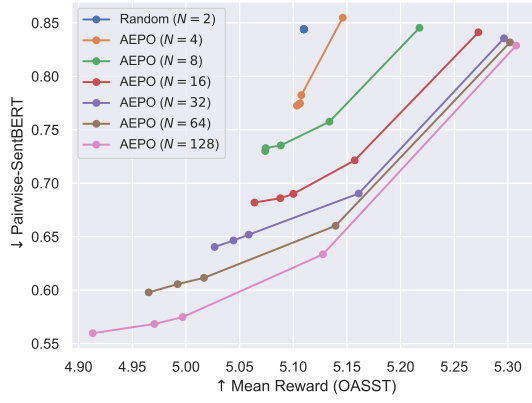


(c) $\uparrow$ Distinct-2

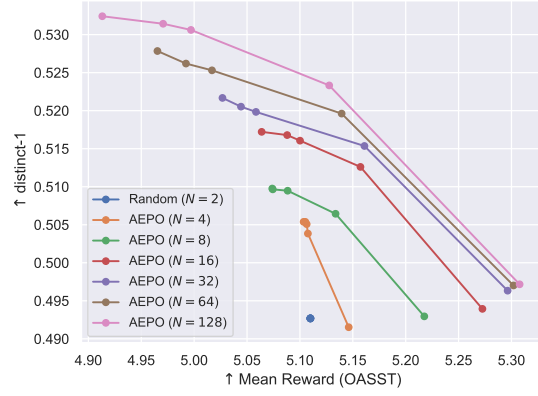


(d) $\uparrow$ Distinct-3

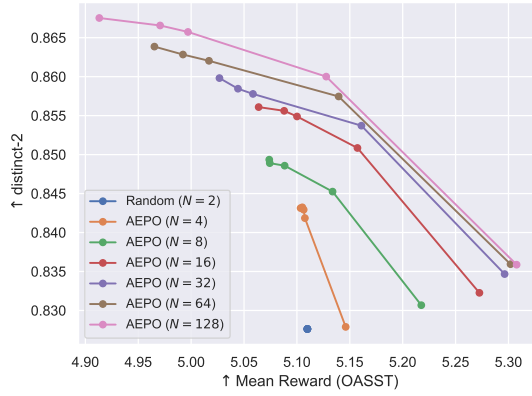
Figure 14: Diversity ($\downarrow$ Sentence BERT and $\uparrow$ Distinct-n) and quality ($\uparrow$ mean reward) of the responses of the preference datasets $\mathcal{D}_{AE}$ generated by AEPO with different numbers of input responses. AEPO successfully generates datasets with better diversity-quality tradeoffs.



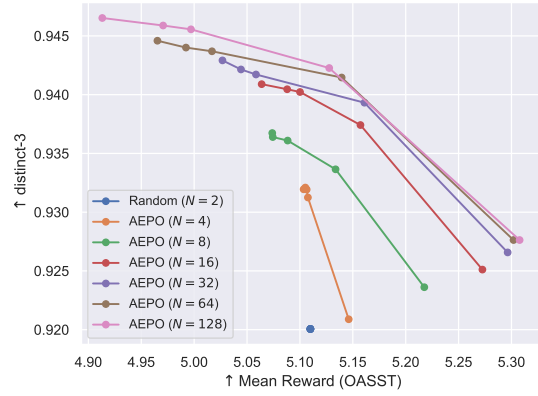
(a) $\downarrow$ Sentence BERT



(b) $\uparrow$ Distinct-1

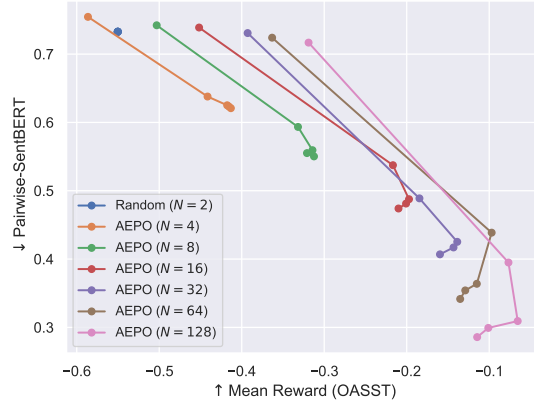


(c) $\uparrow$ Distinct-2

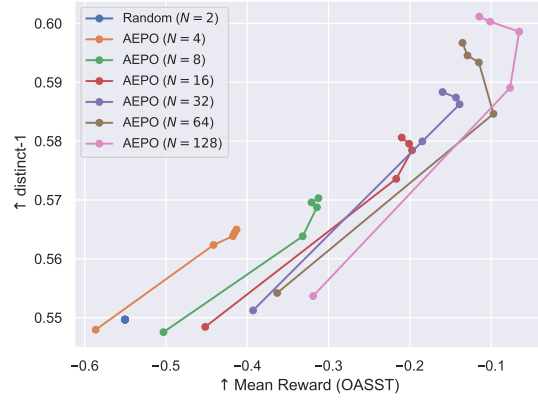


(d) $\uparrow$ Distinct-3

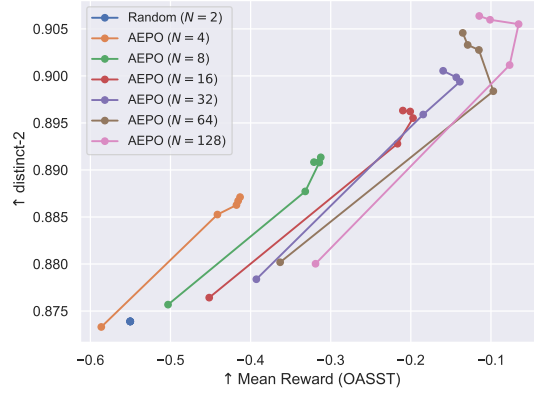
Figure 15: Diversity ($\downarrow$ Sentence BERT and $\uparrow$ Distinct-n) and quality ($\uparrow$ mean reward) of the responses of the preference datasets $\mathcal{D}_{AE}$ generated by AEPO with different numbers of input responses on Anthropic’s Helpfulness dataset.



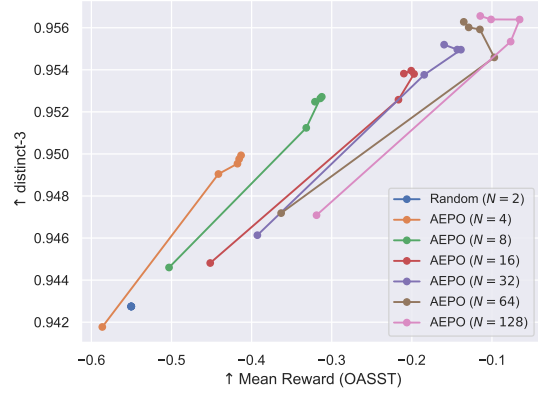
(a) $\downarrow$ Sentence BERT



(b) $\uparrow$ Distinct-1



(c) $\uparrow$ Distinct-2



(d) $\uparrow$ Distinct-3

Figure 16: Diversity ($\downarrow$ Sentence BERT and $\uparrow$ Distinct-n) and quality ($\uparrow$ mean reward) of the responses of the preference datasets $\mathcal{D}_{AE}$ generated by AEPO with different numbers of input responses on Anthropic’s Harmlessness dataset.

Table 24: 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
JCommonsenseMorality	(Takeshita et al., 2023) https://github.com/Language-Media-Lab/commonsense-moral-ja
mistral-7b-sft-beta (Mistral)	(Jiang et al., 2023a; Tunstall et al., 2024) https://huggingface.co/HuggingFaceH4/mistral-7b-sft-beta
dolly-v2-3b (Dolly)	(Conover et al., 2023) https://huggingface.co/databricks/dolly-v2-3b
calm2-7b-chat (CALM2)	https://huggingface.co/cyberagent/calm2-7b-chat
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
Gemma2-9B	(Team et al., 2024) https://huggingface.co/google/gemma-2-9b-it
Gemma2-27B	(Team et al., 2024) https://huggingface.co/google/gemma-2-27b-it
MPNet	(Song et al., 2020) https://huggingface.co/sentence-transformers/all-mpnet-base-v2