ANNOTATION-EFFICIENT LANGUAGE MODEL ALIGNMENT VIA DIVERSE AND REPRESENTATIVE RESPONSE TEXTS

Anonymous authors
Paper under double-blind review

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

Large Language Models (LLMs) trained on massive datasets are capable of solving a variety of tasks in natural language understanding and generation (Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023; OpenAI et al., 2024). However, they have been shown to generate texts containing toxic, untruthful, biased, and harmful outputs (Bai et al., 2022; Lin et al., 2022; Touvron et al., 2023; Casper et al., 2023; Huang et al., 2024b; Guan et al., 2024). Language model alignment aims to address these issues by guiding 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).

The common strategies to align an LLM are Reinforcement learning from human feedback (RLHF) and Direct Preference Optimization (DPO) (Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023). RLHF and DPO use the human preference dataset to train a reward model or a language model directly. The performance of these algorithms is highly dependent on the choice of the preference dataset. However, building a human preference dataset requires human annotations, which are expensive to collect. Thus, the main bottleneck in building a preference dataset is the annotation cost.

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., 2024b; Chakraborty 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,

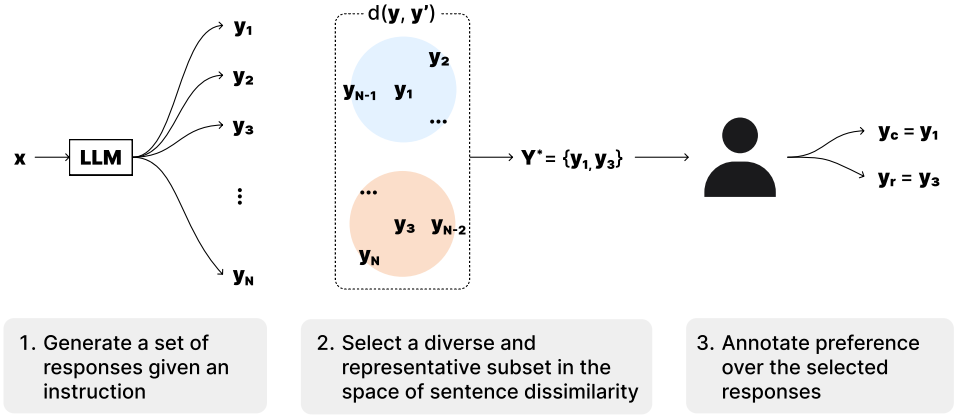


Figure 1: Annotation-Efficient Preference Optimization (AEPO) is a process for generating a preference dataset with diverse and representative responses with fewer annotations. See Section 3 for details. Here we set $k = 2$ and select two responses from the generated responses to annotate.

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.

The question is how to generate an effective preference dataset with a limited 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).
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)

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 required 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 (Table 1). 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, result-

Table 1: Comparison of annotation strategies for preference dataset.

Preference dataset	Human feedback	Scalable	Annotation-efficient
Human feedback	✓	✗	✗
RLAIF (Lee et al., 2024)	✗	✓	✓
West-of-N (Pace et al., 2024)	✓	✓	✗
AEPO (Proposed)	✓	✓	✓

ing 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. 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, Anthropic's hh-rlhf, and JCommonsensMorality 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. We conduct ablation studies and observe that AEPO consistently outperforms WoN with varying settings and hyperparameters (Appendix D). 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:

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

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 of the preference dataset $\mathcal{D}_p$. Existing approaches explore the use of high-performance models (e.g., GPT-4) to synthesize high-quality instructions, responses, and preference feedback (Ding et al., 2023; Honovich et al., 2023; Cui et al., 2023; Mukherjee et al., 2023; Xu et al., 2024a; Liu et al., 2024a).

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) suggested selecting examples to annotate from a pool of unlabeled data to improve the efficiency of in-context learning. Zhou et al. (2023) shows that fine-tuning a model with carefully selected and authored instructions can improve performance. Chen et al. (2024) points out that public instruction datasets contain many low-quality instances and proposes a method to filter out low-quality data, resulting in more efficient fine-tuning.

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 $\mathcal{D}_p$:

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

Algorithm 1 Annotation-Efficient Preference Optimization (AEPO)

Input: A set of pairs of an instruction and a set of 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: $\quad Y^* \leftarrow \arg\max_{Y \subseteq Y_{\text{cand}}, |Y|=k} f_{rep}(Y) + \lambda f_{div}(Y)$ (See Eq. 18)
4: $\quad y_c \leftarrow \arg\max_{y \in Y^*} R(x, y)$
5: $\quad y_r \leftarrow \arg\min_{y \in Y^*} R(x, y)$
6: $\quad \mathcal{D}_{AE} \leftarrow \mathcal{D}_{AE} \cup \{(x, y_c, y_r)\}$
7: **end for**
8: **return** $\mathcal{D}_{AE}$

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 Algorithm 1. We assume that a set of N responses is available for each instruction: $\mathcal{D} = \{(x, \{y_i\}_{i=1}^N)\}$. Instead of annotating the preference over all responses in $\{y_i\}_{i=1}^N$, AEPO subsamples k responses (e.g., $k = 2$) from the candidate set of samples according to the objective function (Eq. 18) that heuristically maximizes the information gain (line 3). We explain the objective function later. Then, it deploys the WoN strategy (Eq. 2) on the subsampled subset of responses Y^* instead of all N responses $\{y_i\}_{i=1}^N$. It annotates the preference over Y^* 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 informative 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.

The performance of the procedure is highly dependent on how we subsample a subset Y from the candidate set of responses $Y_{\text{cand}} := \{y_i\}_{i=1}^N$. We propose to maximize the information gain (IG) (Cover, 1999) as the criteria to select the subset Y. Let R^y be a random variable for the estimated probability distribution of y's reward value ($R(x, y)$) and R^Y be a set of random variables R^y for $y \in Y$. The information gain $\text{IG}(R^{Y_{\text{cand}}}; R^Y)$ measures the reduction in the entropy of the predicted values of $R^{Y_{\text{cand}}}$ when we observe the values of R^Y:

$$\text{IG}(R^{Y_{\text{cand}}}; R^Y) = \mathbf{H}[R^{Y_{\text{cand}}}] - \mathbf{H}[R^{Y_{\text{cand}}} \mid R^Y], \tag{3}$$

where $\mathbf{H}$ is the joint entropy. Our goal is to find an informative subset Y where $\text{IG}(R^{Y_{\text{cand}}}; R^Y)$ is maximized.

Information gain is one of the primary objectives used in active learning, where the goal is to selectively label the most informative unlabeled examples (Lewis & Gale, 1994; Engelson & Dagan, 1996; Guo & Greiner, 2007; Siddhant & Lipton, 2018; Nguyen et al., 2021; Huang et al., 2024a). We choose the subset Y to label the preference so that the information gain for $R^{Y_{\text{cand}}}$ is maximized, which we assume will lead to better alignment.

Since the information gain is not computable in a feasible time for LLMs, we instead make two assumptions to heuristically estimate the information gain. Let d be a cost function that represents the dissimilarity of the two response texts: $d : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, where $d(y, y') = 0$ if $y = y'$.

Heuristic 1 *The preference annotation over Y (R^Y) is more likely to be informative to R^y if it is closer to y. That is, if*

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

then,

$$\mathrm{IG}(R^y; R^Y) \geq \mathrm{IG}(R^y; R^{Y'}) \tag{5}$$

with high probability.

Figure 2 illustrates the intuition behind the heuristic. We assume that similar texts are more likely to have similar preferences. Thus, we assume that selecting a subset Y closer to y is more informative for estimating R^y than a more distant subset Y'.

From Eq. 4, we are motivated to choose a subset Y so that they are closer to $y \in Y_{\text{cand}}$:

$$f_{rep}(Y; y) := -\frac{1}{N} \sum_{y_i \in Y} d(y, y_i), \tag{6}$$

as a smaller f_{rep} leads to larger expected information gain for R^y (Eq. 5). Let $f_{rep}(Y)$ be the sum of $f_{rep}(y; Y)$ for $y \in Y_{\text{cand}}$:

$$f_{rep}(Y) := -\sum_{y \in Y_{\text{cand}}} f_{rep}(y; Y). \tag{7}$$

From the heuristic, the larger $f_{rep}(Y)$ is, the more likely it is that the information gain of $f_{rep}(Y)$ is greater.

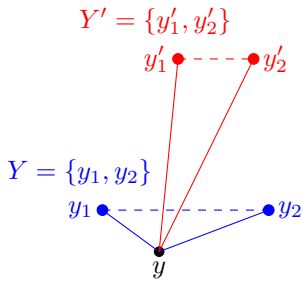


Figure 2: An illustrative example of response subsets for annotating preference. Our algorithm is based on the heuristic that the subset Y that is more diverse and closer to y is more likely to be informative than Y' to infer the value of y.

Remark 1 *Assume Heuristic 1. The preference over Y (R^Y) is more likely to be informative for estimating $R^{Y_{\text{cand}}}$ if it is closer to Y_{cand}. That is, If*

$$f_{rep}(Y) \geq f_{rep}(Y'), \tag{8}$$

then

$$\mathrm{IG}(R^{Y_{\text{cand}}}; R^Y) \geq \mathrm{IG}(R^{Y_{\text{cand}}}; R^{Y'}) \tag{9}$$

with high probability.

The remark is derived from the summation over $y \in Y_{\text{cand}}$ in Heuristic 1. As such, $f_{rep}(Y)$ is a reasonable objective to maximize the information gain (Eq. 3) under the given assumption.

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

$$f_{rep}(Y) = \sum_{y \in Y_{\text{cand}}} f_{rep}(y; Y) \tag{10}$$

$$= \sum_{y \in Y_{\text{cand}}} \left(-\frac{1}{N} \sum_{y \in Y} d(y, y') \right) \tag{11}$$

$$= -\sum_{y \in Y} \left(\frac{1}{N} \sum_{y' \in Y_{\text{cand}} \setminus \{y\}} d(y, y') \right) \tag{12}$$

where $\sum_{y' \in Y_{\text{cand}} \setminus \{y\}} d(y, y')$ can be interpreted as the average distance from y to all other samples. That is, it shows the closeness to the mean of the sample set. Thus, the objective is to select a subset Y that is closer to the center of the samples, making it more representative of the generated samples.

The second heuristic is about the effect of the diversity of a subset Y.

Heuristic 2 *The preference over Y (R^Y) is more likely to be informative for estimating $R^{Y_{\text{cand}}}$ if each pair of samples in Y is more distinct. That is, if*

$$\sum_{y_1 \in Y} \sum_{y_2 \in Y \setminus \{y_1\}} d(y_1, y_2) \geq \sum_{y_1 \in Y'} \sum_{y_2 \in Y' \setminus \{y_1\}} d(y_1, y_2), \tag{13}$$

then,

$$\mathrm{IG}(R^{Y_{\mathrm{cand}}}; R^{Y}) \geq \mathrm{IG}(R^{Y_{\mathrm{cand}}}; R^{Y'}) \tag{14}$$

with high probability.

An example of high and low diversity subsamples (Y and Y') is shown in Figure 2. If the selected samples are too similar (e.g., Y'), then it will be difficult to infer R^y when y is different from both of them. On the other hand, if the selected samples are distinct enough (e.g., Y), then we expect it to be easier to infer R^y.

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

$$f_{div}(Y) = \frac{1}{|Y|} \sum_{y_1 \in Y} \sum_{y_2 \in Y \setminus \{y_1\}} d(y_1, y_2). \tag{15}$$

The objective $f_{div}(Y)$ is equal to the value of Eq. 13, so maximizing it improves the information gain to $R^{Y_{\mathrm{cand}}}$.

An alternative view of f_{div} is that it serves as an upper bound on the difference in distance to a pair of samples in Y, under the assumption that d is a metric. Let y_1, y_2 be a pair of samples in Y with $R(x, y_1) > R(x, y_2)$. It is difficult to infer R^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 (Figure 2). Here, $d(y_1, y_2)$ is an upper bound of $|d(y, y_1) - d(y, y_2)|$ from the triangle inequality:

$$\forall y \; |d(y, y_1) - d(y, y_2)| \leq d(y_1, y_2). \tag{16}$$

Thus, $f_{div}(Y)$ serves as an upper bound on the sum of the difference in distance to a pair of subsampled texts y_1 and y_2:

Remark 2 *Assume Heuristic 2. Let d be a metric over $\mathcal{Y}$. f_{div} is an upper bound on the sum of the distance difference between the sample pairs in Y.:*

$$\frac{1}{|Y|} \sum_{y \in Y_{\mathrm{cand}}} \sum_{y_1 \in Y} \sum_{y_2 \in Y \setminus \{y_1\}} |d(y, y_1) - d(y, y_2)| \leq f_{div}(Y). \tag{17}$$

The proof is immediate from Eq. 16. Thus, it is ideal to have f_{div} large enough so that $|d(y, y_1) - d(y, y_2)|$ is not too small to infer $R^{Y_{\mathrm{cand}}}$. Although the cost functions used in NLP are often not metric (e.g., cosine distance), the remark serves as an intuitive explanation of the diversity objective f_{div}.

Based on the two heuristics, we propose to optimize the following objective to maximize the expected information gain from the subsample Y:

$$\begin{aligned} Y_k^* &:= \operatorname*{arg\,max}_{\substack{Y \subseteq Y_{\mathrm{cand}} \\ |Y| = k}} f_{rep}(Y) + \lambda f_{div}(Y) \\ &= \operatorname*{arg\,max}_{\substack{Y \subseteq Y_{\mathrm{cand}} \\ |Y| = k}} - \sum_{y \in Y} \left(\frac{1}{N} \sum_{y' \in Y_{\mathrm{cand}} \setminus \{y\}} d(y, y') \right) + \lambda \frac{1}{|Y|} \sum_{y_1 \in Y} \sum_{y_2 \in Y \setminus \{y_1\}} d(y_1, y_2), \end{aligned} \tag{18}$$

where λ is a hyperparameter to control the trade-off between the two objectives. We use the cosine distance of the embedding as the dissimilarity function:

$$d(y_1, y_2) = 1 - \cos(\mathrm{emb}(y_1), \mathrm{emb}(y_2)), \tag{19}$$

where $\cos$ is the cosine function and emb is the embedding function. We use the `all-mpnet-base-v2` sentence BERT model as the embedding model because it has been shown to be effective for a variety of sentence embedding tasks (Reimers & Gurevych, 2019; 2020; Song et al., 2020).

4 EXPERIMENTS

Setup. We evaluate the performance of AEPO on DPO using the AlpacaFarm (Dubois et al., 2023) and Anthropic's 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.

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. All the methods use the same set of responses to ensure a fair comparison. For AEPO, the number of subsampled responses 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. We evaluate random sampling and WoN strategy as baselines. We additionally evaluate a coreset-based subsampling strategy (Sener & Savarese, 2018) and a perplexity-based subsampling strategy for AlpacaFarm. 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. Note that we assume that the cost of annotating the preference rank for N responses is linear in N. This assumption favors WoN because it becomes increasingly difficult to annotate preference rank over a larger set of options (Ganzfried, 2017).

We use the OASST reward model (Köpf et al., 2023) to annotate the preference over the responses for the training data. Although it is ideal to use human annotations to evaluate the performance of the algorithms, human annotations are expensive and difficult to reproduce. To this end, we use existing open source reward models as preference annotators for the experiment.

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 Anthropic's 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.

We evaluate the quality of the trained models by sampling a response using nucleus sampling ($p = 0.7$). The model output is evaluated using Eurus-RM-7B (Eurus) (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).

Main Results. Figure 3 shows the Eurus score of the DPO 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 5 and 6 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 Eurus 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 dataset generated by AEPO with $k = 2$, $N \in \{2 \text{ (Random)}, 4, 8, 16, 32, 64, 128\}$, and $\lambda \in \{0, 0.3, 0.5, 1.0, 2.0\}$. To measure the semantic and

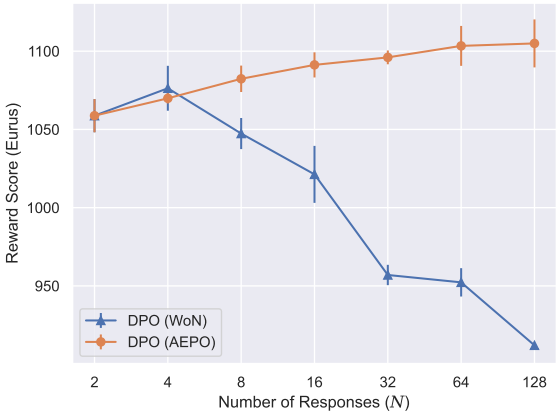


Figure 3: Evaluation of AEPO and West-of-N for DPO *with an annotation budget fixed to* 2 *times the number of instructions* on AlpacaFarm. The line represents the average reward score and the bar shows the standard deviation over three runs.

Figure 4: The number of instructions (#Insts) and annotations (#Annots) used by the preference annotation strategies in Figures 5, 6, and 8.

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

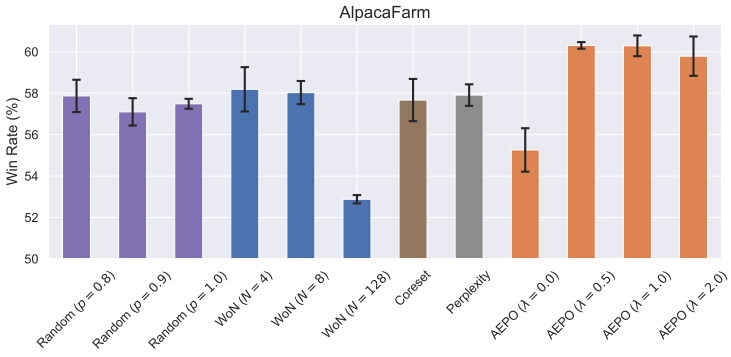


Figure 5: Evaluation of preference annotation strategies for DPO on AlpacaFarm using Mistral under the annotation budget fixed to 2 times the number of instructions. The win rate against the SFT model is evaluated. The bar represents the mean, and the error bar indicates the standard deviation of three runs.

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 $-f_{rep}(Y)/|Y_{\text{cand}}|$ which is the average similarity ($-d(y, y')$) of the selected texts Y to the whole sample set Y_{cand}. The quality of the responses is measured by the average reward score of the selected responses.

The result is shown in Figure 7a. 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 7b). This aligns with prior work reporting that diversity and representativeness objectives may also improve the quality of the output texts (Vijayakumar et al., 2016; 2018; Eikema & 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 Anthropic's datasets (Figures 15, 16, and 17 (Appendix H).

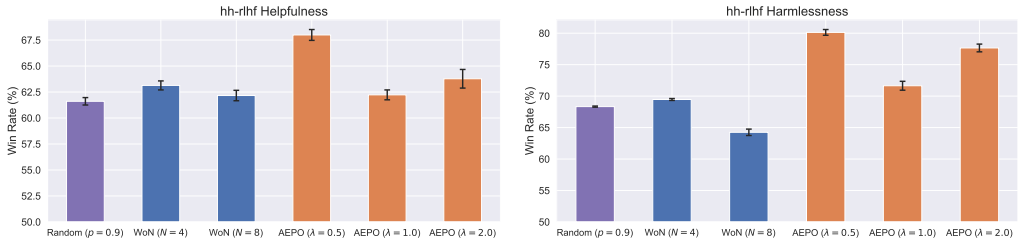
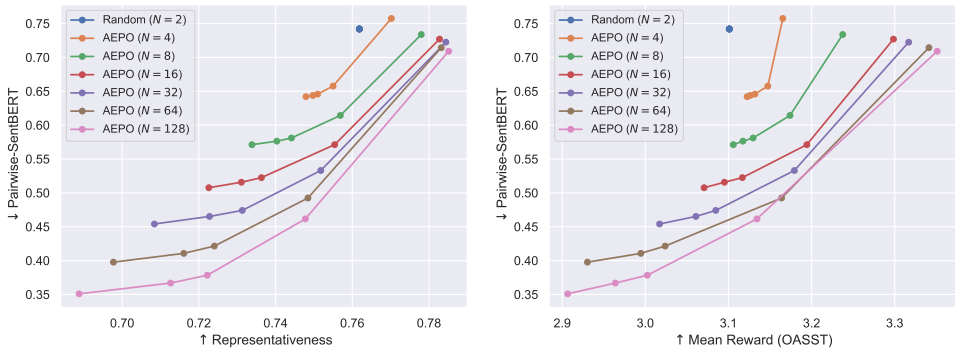


Figure 6: Evaluation of preference dataset annotation strategies for DPO on hh-rlhf's Helpfulness and Harmlessness dataset using Mistral under the annotation budget fixed to 2 times the number of instructions. The win rate against the SFT model is evaluated. The bar represents the mean, and the error bar indicates the standard deviation of three runs.



(a) Representativeness (b) Reward Score

Figure 7: Diversity (↓Sentence BERT), representativeness, and quality (↑mean reward) of the responses of the preference datasets $\mathcal{D}_{AE}$ generated by the subsampling process of 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 and diversity-quality trade-offs without requiring additional annotations.

Both diversity and representativeness of the preference dataset are important for preference learning. The question is what contributes to the improved performance of AEPO. We evaluate AEPO with $\lambda \in \{0.0, 0.5, 1.0, 2.0\}$ to investigate the importance of diversity and representativeness of responses on AlpacaFarm dataset. AEPO with moderate size of λ outperforms AEPO with higher or lower λ (Figure 5 and 10). The result shows 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).

AEPO is effective for learning Japanese commonsense morality with a limited annotation budget. To evaluate the proposed method in an application where the annotation budget is often limited, we conduct an experiment using the 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 (Hendrycks et al., 2021) 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., 2024a), it is difficult to collect a large number of annotations. Therefore, we consider the task of learning Japanese commonsense morality to be a good benchmark for evaluating AEPO in a realistic application where AEPO is needed.

We use 800 entries ($|\mathcal{D}| = 800$) from the train split for training and 500 entries from the test split for evaluation. We train a Japanese LLM (`calm2-7b-chat`) using the train set of the JCM dataset (Sugimoto, 2024). As a reward model, we evaluate the accuracy of the output with respect to the

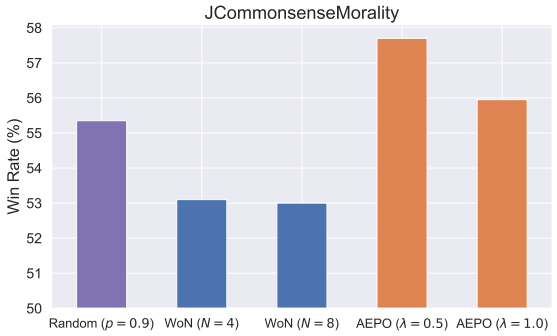


Figure 8: Evaluation of preference annotation strategies for DPO on the JCommonsenseMorality (JCM) dataset using `calm2-7b-chat` under a fixed annotation budget. The win rate against the SFT model is evaluated.

label provided in the dataset, as well as the overall quality. See Appendix G for the evaluation procedure. The results are summarized in Figure 8. Overall, AEPO outperforms the baselines within the same annotation budget constraint. The result on the JCM dataset suggests that AEPO is an effective strategy in one of the tasks where the available annotations are limited.

5 Related Work

Minimum Bayes risk decoding. Eq. 7 and 18 are largely inspired by Minimum Bayes Risk (MBR) decoding (Kumar & Byrne, 2002; 2004; Eikema & 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. 7. 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. 18 is strongly inspired by the objective function of Diverse MBR (DMBR) decoding (Jinnai et al., 2024). The novelty of our work is to introduce the objective function of DMBR as a strategy to subsample representative and diverse responses from many candidate responses so that the annotation budget can be used efficiently.

Active learning. Related work in active learning is described in Appendix C.

6 Conclusions

We propose Annotation-Efficient Preference Optimization (AEPO), an annotation-efficient dataset subsampling strategy for language model alignment. The subsampling strategy aims to maximize the information gain using two heuristics on how the preference information is propagated between samples. By focusing the annotation effort on the selected responses, AEPO achieves efficient preference optimization with a limited annotation budget. We evaluate the subsampling strategy and show that it successfully selects diverse and representative samples from the candidates (Figure 7). Experimental results show that AEPO outperforms the baselines on AlpacaFarm, Anthropic's hh-rlhf, and JCM datasets (Figures 5, 6, and 8). Our ablation study covers various settings, including GPT-4 evaluation, off-policy training, out-of-domain evaluation, and using different hyperparameters (Appendix D). The study shows that AEPO consistently outperforms the baselines in various settings. We believe that AEPO is a critical contribution to promoting preference optimization research by addressing the severe obstacle, the cost of creating better preference data.

REFERENCES

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling "culture" in LLMs: A survey. *arXiv preprint arXiv:2403.15412*, 2024.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in LLMs. *arXiv preprint arXiv:2402.14740*, 2024.

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 9954–9972, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.592. URL https://aclanthology.org/2024.findings-acl.592.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

Michael Bloodgood and Chris Callison-Burch. Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 854–864, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1088.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. Instruction mining: When data mining meets large language model finetuning. *arXiv preprint arXiv:2307.06290*, 2023a.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti (eds.), *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pp. 53–67, Dubrovnik, Croatia, May 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.7. URL https://aclanthology.org/2023.c3nlp-1.7.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Bedi, and Mengdi Wang. Maxmin-RLHF: Towards equitable alignment of large language models with diverse human preferences. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024. URL https://openreview.net/forum?id=NCQp4KpT8R.

Jonathan D. Chang, Wenhao Zhan, Owen Oertell, Kianté Brantley, Dipendra Misra, Jason D. Lee, and Wen Sun. Dataset reset policy optimization for RLHF. *arXiv preprint arXiv:2404.08495*, 2024.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a better Alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FdVXgSJhvz.

Ruitao Chen and Liwei Wang. The power of active multi-task learning in reinforcement learning from human feedback. *arXiv preprint arXiv:2405.11226*, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018.

David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15:201–221, 1994.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the world's first truly open instruction-tuned LLM, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Javier De la Rosa, Eduardo G Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Maria Grandury. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *arXiv preprint arXiv:2207.06814*, 2022.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.183. URL https://aclanthology.org/2023.emnlp-main.183.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. RLHF workflow: From reward modeling to online RLHF. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a13aYUU9eU.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 30039–30069. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5fc47800ee5b30b8777fdd30abcaaf3b-Paper-Conference.pdf.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled AlpacaEval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2024.

Matthias Eck, Stephan Vogel, and Alex Waibel. Low cost portability for statistical machine translation based on n-gram frequency and TF-IDF. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA, October 24-25 2005. URL https://aclanthology.org/2005.iwslt-1.7.

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

Sean P. Engelson and Ido Dagan. Minimizing manual annotation cost in supervised training from corpora. In *34th Annual Meeting of the Association for Computational Linguistics*, pp. 319–326, Santa Cruz, California, USA, June 1996. Association for Computational Linguistics. doi: 10.3115/981863.981905. URL https://aclanthology.org/P96-1042.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9198–9209, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.617. URL https://aclanthology.org/2023.findings-emnlp.617.

Sam Ganzfried. Optimal number of choices in rating contexts. In Tatiana V. Guy, Miroslav Kárný, David Rios-Insua, and David H. Wolpert (eds.), *Proceedings of the NIPS 2016 Workshop on Imperfect Decision Makers*, volume 58 of *Proceedings of Machine Learning Research*, pp. 61–74. PMLR, 09 Dec 2017. URL https://proceedings.mlr.press/v58/ganzfried17a.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/gao23h.html.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023b. URL https://zenodo.org/records/10256836.

Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Hongxia Ma, Li Zhang, Hao Yang, and Tong Xiao. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. *arXiv preprint arXiv:2402.18191*, 2024.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Travis Greene, Galit Shmueli, and Soumya Ray. Taking the person seriously: Ethically aware is research in the era of reinforcement learning-based personalization. *Journal of the Association for Information Systems*, 24(6):1527–1561, 2023.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, June 2024.

Arnav Gudibande, Eric Wallace, Charlie Victor Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Kz3yckpCN5.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct language model alignment from online AI feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Yuhong Guo and Russ Greiner. Optimistic active learning using mutual information. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pp. 823–829, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14409–14428, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.806. URL https://aclanthology.org/2023.acl-long.806.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Ido Dagan. Selective annotation via data allocation: These data should be triaged to experts for annotation rather than the model. *arXiv preprint arXiv:2405.12081*, 2024a.

Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7):175, 2024b.

Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. NEFTune: Noisy embeddings improve instruction finetuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=0bMmZ3fkCk.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.792. URL https://aclanthology.org/2023.acl-long.792.

Yuu Jinnai. Does cross-cultural alignment change the commonsense morality of language models? In Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Laura Cabello, Yong Cao, Ife Adebara, and Li Zhou (eds.), *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pp. 48–64, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.c3nlp-1.5. URL https://aclanthology.org/2024.c3nlp-1.5.

Yuu Jinnai, Ukyo Honda, Tetsuro Morimura, and Peinan Zhang. Generating diverse and high-quality texts by minimum Bayes risk decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 8494–8525, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.503. URL https://aclanthology.org/2024.findings-acl.503.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *arXiv preprint arXiv:2303.05453*, 2023.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. OpenAssistant conversations - democratizing large language model alignment. In *Thirty-seventh Conference on Neural Information Processing*

Systems Datasets and Benchmarks Track, 2023. URL https://openreview.net/forum?id=VSJotgbPHF.

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

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

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Forty-first International Conference on Machine Learning*, 2024.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In Bruce W. Croft and C. J. van Rijsbergen (eds.), *SIGIR '94*, pp. 3–12, London, 1994. Springer London. ISBN 978-1-4471-2099-5.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024a.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.421. URL https://aclanthology.org/2024.naacl-long.421.

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

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*, 2024a. URL https://openreview.net/forum?id=OJaWBhh61C.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=xbjSwwrQOe.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024c. URL https://openreview.net/forum?id=BTKAeLqLMw.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pszewhybU9.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining LLMs at scale. *arXiv preprint arXiv:2309.04564*, 2023.

Andrew McCallum and Kamal Nigam. Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pp. 350–358, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1558605568.

Tetsuro Morimura, Mitsuki Sakamoto, Yuu Jinnai, Kenshi Abe, and Kaito Air. Filtered direct preference optimization. *arXiv preprint arXiv:2404.13846*, 2024.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv preprint arXiv:2306.02707*, 2023.

Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16366–16393, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.862. URL https://aclanthology.org/2024.acl-long.862.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. An information-theoretic framework for unifying active learning problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9126–9134, 2021.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv

Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. West-of-N: Synthetic preference generation for improved reward modeling. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2024. URL https://openreview.net/forum?id=7kNwZhMefs.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. Valuenet: A new dataset for human value driven dialogue system. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 11183–11191. AAAI Press, 2022. doi: 10.1609/AAAI.V36I10.21368. URL https://doi.org/10.1609/aaai.v36i10.21368.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*, 2024.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

(EMNLP-IJCNLP), pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. URL https://arxiv.org/abs/2004.09813.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9Vrb9D0WI4.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 1070–1079, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL https://aclanthology.org/D08-1112.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. Understanding the capabilities and limitations of large language models for cultural commonsense. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5668–5680, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.316. URL https://aclanthology.org/2024.naacl-long.316.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. Understanding the capabilities and limitations of large language models for cultural commonsense. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5668–5680, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.316. URL https://aclanthology.org/2024.naacl-long.316.

Hakim Sidahmed, Samrat Phatale, Alex Hutcheson, Zhuonan Lin, Zhang Chen, Zac Yu, Jarvis Jin, Roman Komarytsia, Christiane Ahlheim, Yonghao Zhu, Simral Chaudhary, Bowen Li, Saravanan Ganesh, Bill Byrne, Jessica Hoffmann, Hassan Mansoor, Wei Li, Abhinav Rastogi, and Lucas Dixon. PERL: Parameter efficient reinforcement learning from human feedback. *arXiv preprint arXiv:2403.10704*, 2024.

Aditya Siddhant and Zachary C. Lipton. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2904–2909, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1318. URL https://aclanthology.org/D18-1318.

Feifan Song, Bowen Yu, Hao Lang, Haiyang Yu, Fei Huang, Houfeng Wang, and Yongbin Li. Scaling data diversity for fine-tuning language models in human alignment. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14358–14369, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1251.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.

Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 19937–19947. AAAI Press, 2024a. doi: 10.1609/AAAI.V38I18.29970. URL https://doi.org/10.1609/aaai.v38i18.29970.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024b.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

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. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=qY1hlv7gwg.

Kaito Sugimoto. Exploring Open Large Language Models for the Japanese Language: A Practical Guide. *Jxiv preprint*, 2024. doi: 10.51094/jxiv.682.

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

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=bWNPx6t0sF.

Masashi Takeshita, Rafal Rzpeka, and Kenji Araki. JCommonsenseMorality: Japanese dataset for evaluating commonsense morality understanding. In *In Proceedings of The Twenty Nineth Annual Meeting of The Association for Natural Language Processing (NLP2023)*, pp. 357–362, 2023. URL https://www.anlp.jp/proceedings/annual_meeting/2023/pdf_dir/D2-1.pdf. in Japanese.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, and Will Dabney. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024a.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Remi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Avila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47725–47742. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/tang24b.html.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Tristan Thrush, Christopher Potts, and Tatsunori Hashimoto. Improving pretraining data using perplexity correlations. *arXiv preprint arXiv:2409.05816*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher,

Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aKkAwZB6JV.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 10 2016. URL http://arxiv.org/abs/1610.02424.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. TRL: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530, Jun. 2023. doi: 10.1609/aaai.v37i12.26698. URL https://ojs.aaai.org/index.php/AAAI/article/view/26698.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=CfXh93NDgH.

Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*, 2023.

Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. Exploring multilingual concepts of human value in large language models: Is value alignment consistent, transferable and controllable across languages? *arXiv preprint arXiv:2402.18120*, 2024b.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is DPO superior to PPO for LLM alignment? a comprehensive study. In *Forty-first International Conference on Machine Learning*, 2024c. URL https://openreview.net/forum?id=6XH8R7YrSk.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 391–398, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.34. URL https://aclanthology.org/2024.naacl-short.34.

Tianshu Yu, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. Constructive large language models alignment with diverse feedback. *arXiv preprint arXiv:2310.06450*, 2023.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. Advancing LLM reasoning generalists with preference trees. In *AI for Math Workshop @ ICML 2024*, 2024a. URL https://openreview.net/forum?id=2Y1iiCqM5y.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=0NphYCmgua.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Udhyakumar Nallasamy, and Matthias Paulik. Empirical evaluation of active learning techniques for neural MT. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 84–93, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6110. URL https://aclanthology.org/D19-6110.

Dylan Zhang, Justin Wang, and Francois Charton. Instruction diversity drives generalization to unseen tasks. *arXiv preprint arXiv:2402.10891*, 2024a.

Honggen Zhang, Igor Molybog, June Zhang, and Xufeng Zhao. REAL: Response embedding-based alignment for llms. *arXiv preprint arXiv:2409.17169*, 2024b.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. A survey of active learning for natural language processing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6166–6190, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.414. URL https://aclanthology.org/2022.emnlp-main.414.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. Active learning approaches to enhancing neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1796–1806, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.162. URL https://aclanthology.org/2020.findings-emnlp.162.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=KBMOKmX2he.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 10586–10613, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.630. URL https://aclanthology.org/2024.findings-acl.630.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2020.

A HYPERPARAMETERS

Table 2 lists the hyperparameters we use to run DPO. Table 3 lists the hyperparameters we use to generate the texts for evaluation.

Table 2: 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 3: 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 & Savarese (2018) (Algorithm 1, k-Center-Greedy). The objective function for selecting the subset Y is the following:

$$Y^* = \underset{Y \subseteq Y_{\text{cand}}}{\arg\min} \max_{y \in Y_{\text{cand}}} \min_{y' \in Y} d(y, y'). \tag{20}$$

Intuitively, Eq. 20 is similar to the representative objective (f_{rep}; Eq. 7) 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:

$$Y^* = \{\underset{y \in Y_{\text{cand}}}{\arg\max}\, PP(y \mid x), \underset{y \in Y_{\text{cand}}}{\arg\min}\, PP(y \mid x)\}, \tag{21}$$

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

C Additional 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). There are roughly two active learning strategies used in NLP (Zhang et al., 2022). One uses the informativeness of the data instances, such as uncertainty and disagreement of the models (Lewis & Gale, 1994; Engelson & Dagan, 1996; Siddhant & Lipton, 2018; Huang et al., 2024a). 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 & Nigam, 1998; Settles & Craven, 2008; Zhao et al., 2020; Chen & 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 & 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. The novelty of our study lies in applying these ideas to the language model alignment problem to reduce the annotation cost.

D Ablation Study

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

D.1 GPT-4 Evaluation

Figure 9 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 Eurus. 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]

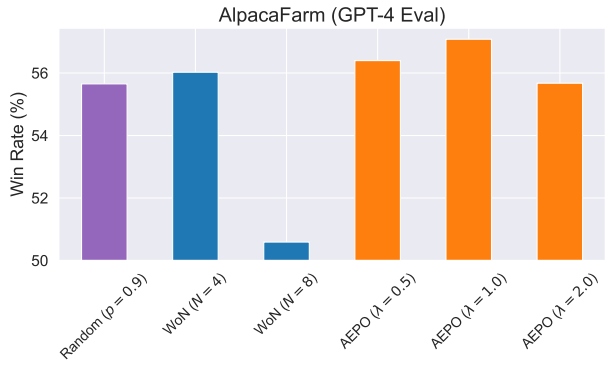


Figure 9: Evaluation of AEPO on the AlpacaFarm dataset using GPT-4 as an evaluator. The win rate against the SFT model is evaluated.

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 10 shows the results of the off-policy learning using Eurus 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 18for the result using other reward models.

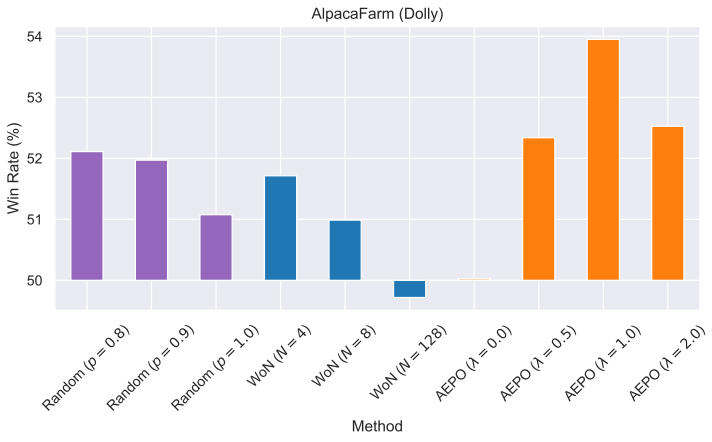


Figure 10: Evaluation of AEPO on training Dolly language model using the AlpacaFarm dataset. We generate responses with Mistral and use the sampled responses to train Dolly. The win rate against the SFT model is evaluated.

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 4 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.

Table 4: Evaluation of DPO models trained with AlpacaFarm on out-of-domain benchmarks. Means and standard errors are reported.

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

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 Section 4. Tables 5 and 6 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 5: 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 ($p = 0.8$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.021	997.05	54.22	55.59	52.49
Random ($p = 0.9$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.029	970.77	54.10	54.72	52.64
Random ($p = 1.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	<u>2.099</u>	1009.53	55.47	<u>56.96</u>	53.64
WoN ($N = 4$)	$\lvert\mathcal{D}\rvert/2$	$2\lvert\mathcal{D}\rvert$	2.088	1031.62	56.34	56.71	53.98
WoN ($N = 8$)	$\lvert\mathcal{D}\rvert/4$	$2\lvert\mathcal{D}\rvert$	2.052	993.94	54.84	56.09	<u>54.10</u>
AEPO ($\lambda = 0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	1.994	936.94	53.48	53.35	53.10
AEPO ($\lambda = 0.5$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.079	981.37	<u>56.77</u>	55.53	**54.12**
AEPO ($\lambda = 1.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	**2.121**	**1063.08**	**58.26**	**58.07**	53.98
AEPO ($\lambda = 2.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.072	<u>1034.58</u>	55.53	56.34	53.97
WoN ($N = 128$)	$\lvert\mathcal{D}\rvert$	$128\lvert\mathcal{D}\rvert$	2.339	1169.37	65.47	63.23	59.61

Table 6: 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 ($p = 0.8$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.310	1149.53	63.11	60.62	59.18
Random ($p = 0.9$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	<u>2.394</u>	1140.02	65.96	59.25	60.00
Random ($p = 1.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.308	1096.25	63.11	58.01	58.96
WoN ($N = 4$)	$\lvert\mathcal{D}\rvert/2$	$2\lvert\mathcal{D}\rvert$	2.390	1160.43	<u>66.02</u>	<u>63.66</u>	61.68
WoN ($N = 8$)	$\lvert\mathcal{D}\rvert/4$	$2\lvert\mathcal{D}\rvert$	2.357	<u>1183.47</u>	65.65	63.29	61.28
AEPO ($\lambda = 0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.186	1050.34	60.62	58.01	57.80
AEPO ($\lambda = 0.5$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.379	1172.73	63.29	**63.91**	<u>60.37</u>
AEPO ($\lambda = 1.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.354	1164.29	64.35	63.60	**60.62**
AEPO ($\lambda = 2.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	**2.400**	**1203.51**	**66.34**	63.60	59.69
WoN ($N = 128$)	$\lvert\mathcal{D}\rvert$	$128\lvert\mathcal{D}\rvert$	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 7 and 8. 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 7: 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 8: 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

We show examples of the dataset entries generated by AEPO and WoN in Tables 9 and 10. 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 9: 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\n\nHe 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 10: 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.\n\nWe 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 ($N = 8$)	Chosen	There has been a long stretch of bitterly cold weather. OR Cold weather has persisted for an extended period of time.
	Rejected	For a long time, the weather has been very cold.
AEPO ($\lambda = 0.0$)	Chosen	For an extended period of time, the weather has been extremely cold.
	Rejected	For an extended period of time, the weather has been extremely cold.
AEPO ($\lambda = 1.0$)	Chosen	The weather has been exceptionally frigid for an extended time.
	Rejected	We have experienced a long time with very low temperatures.
AEPO ($\lambda = 2.0$)	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.

F Evaluation of West-of-N Strategy

WoN is an effective strategy when an abundance of annotations is available. Table 11 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 11.

Table 11: 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 ($p = 0.9$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.174	1058.78	59.71	57.10	55.54
WoN ($N = 4$)	$\lvert\mathcal{D}\rvert$	$4\lvert\mathcal{D}\rvert$	2.315	1105.60	64.35	61.37	59.26
WoN ($N = 8$)	$\lvert\mathcal{D}\rvert$	$8\lvert\mathcal{D}\rvert$	2.422	1225.22	66.09	67.20	62.73
WoN ($N = 16$)	$\lvert\mathcal{D}\rvert$	$16\lvert\mathcal{D}\rvert$	2.454	1237.81	68.14	64.66	63.42
WoN ($N = 32$)	$\lvert\mathcal{D}\rvert$	$32\lvert\mathcal{D}\rvert$	2.529	1246.53	70.56	66.09	64.51
WoN ($N = 64$)	$\lvert\mathcal{D}\rvert$	$64\lvert\mathcal{D}\rvert$	2.470	1248.19	68.07	66.58	63.72
WoN ($N = 128$)	$\lvert\mathcal{D}\rvert$	$128\lvert\mathcal{D}\rvert$	2.522	1262.77	70.12	67.31	64.17

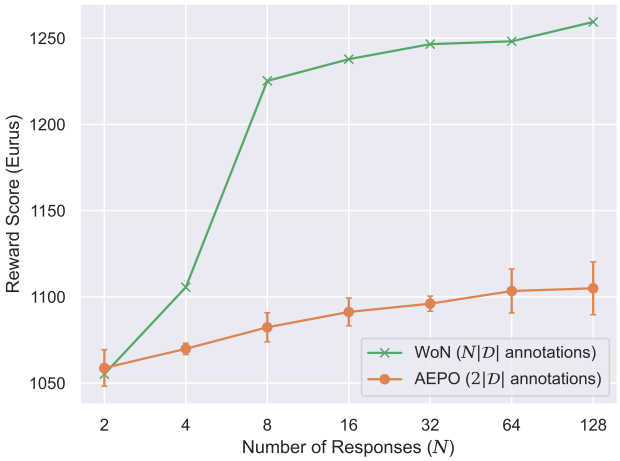


Figure 11: Evaluation of the DPO using the WoN strategy without limiting the annotation budget on AlpacaFarm using Mistral. The Eurus reward model is used for evaluation. With more annotations, WoN outperforms AEPO.

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 12 and 13 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 14, 15, 16, and 17. Table 18 is the result of training Dolly.

Interestingly, we observed that AEPO outperforms WoN with 64 times more annotations in Anthropic's datasets (Tables 14, 15, 16, and 17). 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 12: 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 ($p = 0.8$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.155 ± 0.010	1088.71 ± 17.90
Random ($p = 0.9$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.174 ± 0.009	1058.78 ± 10.60
Random ($p = 1.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.168 ± 0.007	1044.35 ± 0.98
WoN ($N = 4$)	$\lvert\mathcal{D}\rvert/2$	$2\lvert\mathcal{D}\rvert$	2.217 ± 0.012	1076.31 ± 14.35
WoN ($N = 8$)	$\lvert\mathcal{D}\rvert/4$	$2\lvert\mathcal{D}\rvert$	2.197 ± 0.005	1047.37 ± 9.94
WoN ($N = 128$)	$\lvert\mathcal{D}\rvert/64$	$2\lvert\mathcal{D}\rvert$	1.926 ± 0.005	912.03 ± 1.25
Coreset	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.107 ± 0.011	1037.100 ± 11.31
Perplexity	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.187 ± 0.008	1051.52 ± 15.54
AEPO ($\lambda = 0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.063 ± 0.009	999.03 ± 1.43
AEPO ($\lambda = 0.5$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	**2.230** ± 0.011	$\underline{1094.20} \pm 13.70$
AEPO ($\lambda = 1.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	$\underline{2.222} \pm 0.009$	**1104.97** ± 15.33
AEPO ($\lambda = 2.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	2.219 ± 0.010	1085.78 ± 9.72
WoN ($N = 128$)	$\lvert\mathcal{D}\rvert$	$128\lvert\mathcal{D}\rvert$	2.522 ± 0.008	1262.77 ± 5.62

Table 13: 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 ($p = 0.8$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	59.86 ± 1.44	57.87 ± 0.78	56.20 ± 0.31
Random ($p = 0.9$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	59.71 ± 0.52	57.10 ± 0.66	55.54 ± 0.62
Random ($p = 1.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	59.32 ± 0.85	57.49 ± 0.24	56.17 ± 0.74
WoN ($N = 4$)	$\lvert\mathcal{D}\rvert/2$	$2\lvert\mathcal{D}\rvert$	60.34 ± 1.09	58.19 ± 1.07	56.61 ± 0.24
WoN ($N = 8$)	$\lvert\mathcal{D}\rvert/4$	$2\lvert\mathcal{D}\rvert$	$\underline{60.64} \pm 0.61$	58.03 ± 0.56	56.00 ± 0.62
WoN ($N = 128$)	$\lvert\mathcal{D}\rvert/64$	$2\lvert\mathcal{D}\rvert$	51.55 ± 0.53	52.88 ± 0.20	50.16 ± 0.16
Coreset	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	56.71 ± 0.93	57.67 ± 0.52	56.57 ± 0.20
Perplexity	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	60.05 ± 0.52	57.91 ± 1.05	54.23 ± 0.56
AEPO ($\lambda = 0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	56.83 ± 0.49	55.26 ± 1.05	54.92 ± 0.16
AEPO ($\lambda = 0.5$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	59.23 ± 0.91	**60.31** ± 0.16	56.42 ± 0.31
AEPO ($\lambda = 1.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	**62.40** ± 0.22	$\underline{60.29} \pm 0.50$	$\underline{56.97} \pm 0.24$
AEPO ($\lambda = 2.0$)	$\lvert\mathcal{D}\rvert$	$2\lvert\mathcal{D}\rvert$	59.71 ± 0.45	59.79 ± 0.95	**57.36** ± 0.38
WoN ($N = 128$)	$\lvert\mathcal{D}\rvert$	$128\lvert\mathcal{D}\rvert$	70.12 ± 0.56	67.31 ± 0.25	64.17 ± 0.66

Table 14: 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</u> ± 4.32
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</u> ± 0.011	1599.03 ± 18.620
WoN ($N = 128$)	$\|\mathcal{D}\|$	$128\|\mathcal{D}\|$	5.186 ± 0.007	1648.45 ± 7.56

Table 15: 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</u> ± 0.81
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</u> ± 0.36	62.06 ± 0.50	59.50 ± 0.31
AEPO ($\lambda = 2.0$)	$\|\mathcal{D}\|$	$2\|\mathcal{D}\|$	65.67 ± 0.26	<u>63.77</u> ± 0.90	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 16: 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</u> ± 0.031	**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</u> ± 15.55
WoN ($N = 128$)	$\|\mathcal{D}\|$	$128\|\mathcal{D}\|$	0.071 ± 0.010	530.02 ± 3.65

Table 17: 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 ($\lambda = 0.5$)	$\|\mathcal{D}\|$	$2\|\mathcal{D}\|$	$\underline{79.47} \pm 0.47$	$\mathbf{80.13} \pm 0.46$	$\mathbf{69.72} \pm 0.59$
AEPO ($\lambda = 1.0$)	$\|\mathcal{D}\|$	$2\|\mathcal{D}\|$	73.79 ± 0.13	71.62 ± 0.71	$\underline{68.76} \pm 0.09$
AEPO ($\lambda = 2.0$)	$\|\mathcal{D}\|$	$2\|\mathcal{D}\|$	$\mathbf{80.55} \pm 0.09$	$\underline{77.65} \pm 0.62$	67.87 ± 0.85
WoN ($N = 128$)	$\|\mathcal{D}\|$	$128\|\mathcal{D}\|$	72.72 ± 0.25	72.54 ± 0.17	68.27 ± 0.32

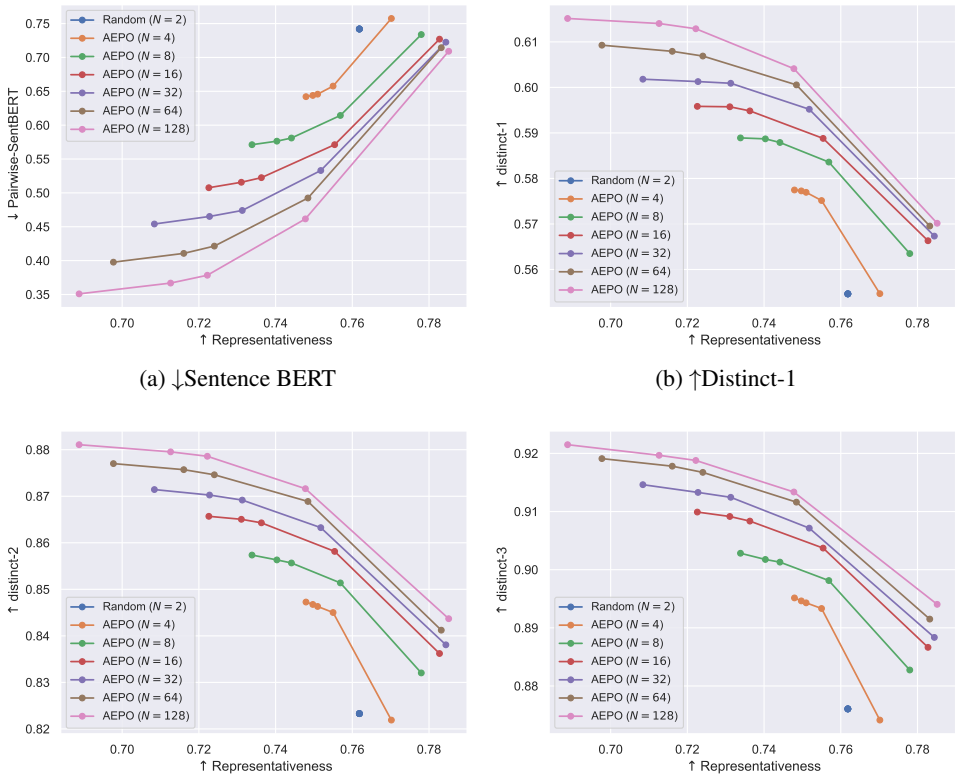
Table 18: Evaluation of preference dataset configuration strategies for off-policy learning. We generate responses using Mistral and use the generated responses to train Dolly. LoRA hyperparameters are set $r = 32$ and $\alpha = r/4$. Note that OASST is used as a proxy reward model to annotate the preference 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 12, 13, and 14 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 15, 16, and 17 show the diversity (distinct-n) and quality (mean reward) tradeoff. AEPO successfully improves the diverse-quality tradeoff with a larger number of response texts.



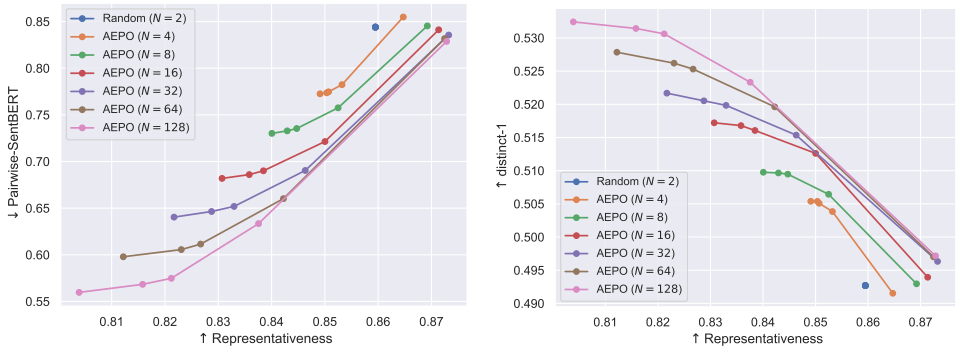
(a) ↓Sentence BERT

(b) ↑Distinct-1

(c) ↑Distinct-2

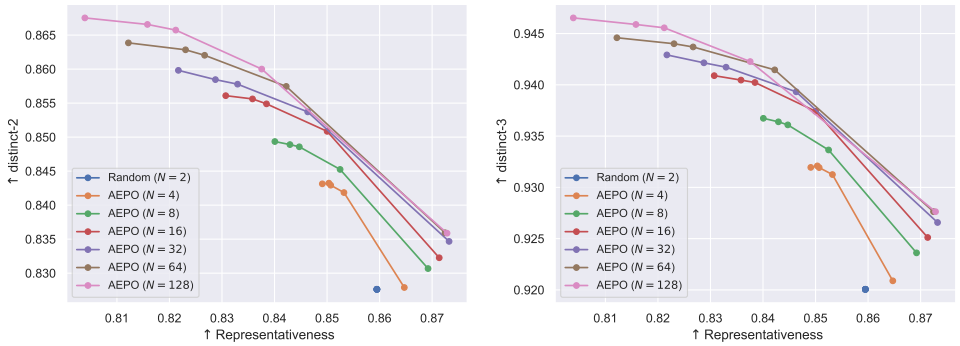
(d) ↑Distinct-3

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. AEPO successfully generates datasets with better diversity-representativeness tradeoffs.



(a) ↓Sentence BERT

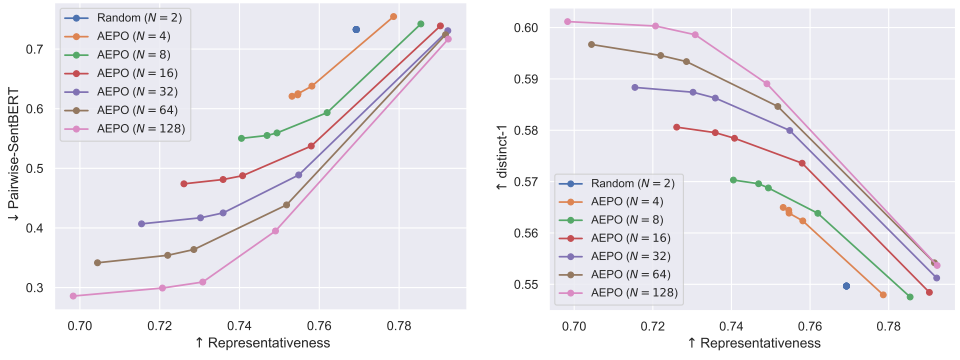
(b) ↑Distinct-1



(c) ↑Distinct-2

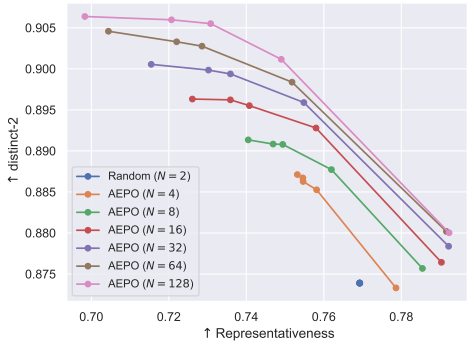
(d) ↑Distinct-3

Figure 13: 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.

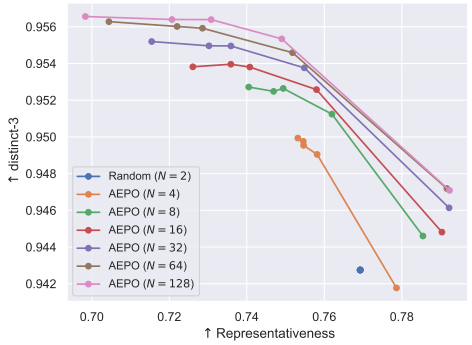


(a) ↓Sentence BERT

(b) ↑Distinct-1

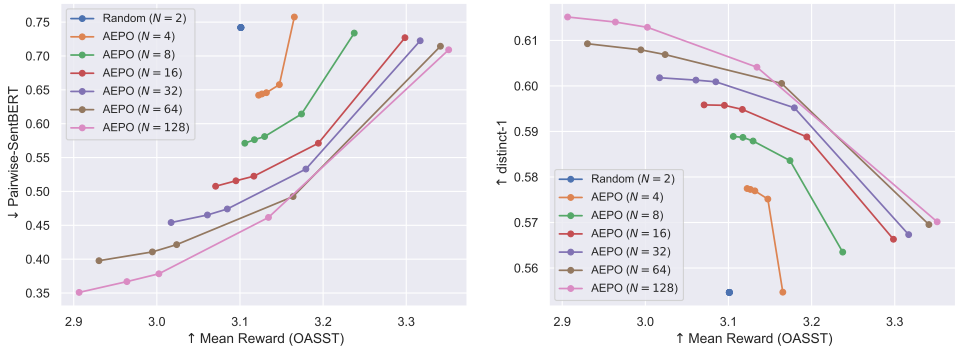


(c) ↑Distinct-2



(d) ↑Distinct-3

Figure 14: 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 Harmlessness dataset.



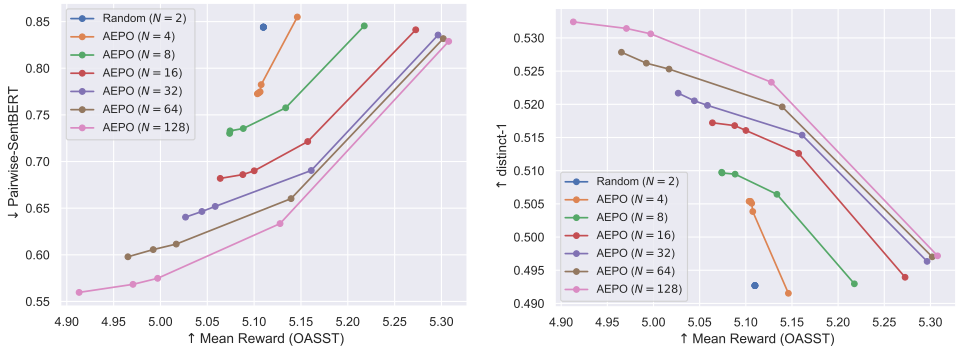
(a) ↓Sentence BERT

(b) ↑Distinct-1

(c) ↑Distinct-2

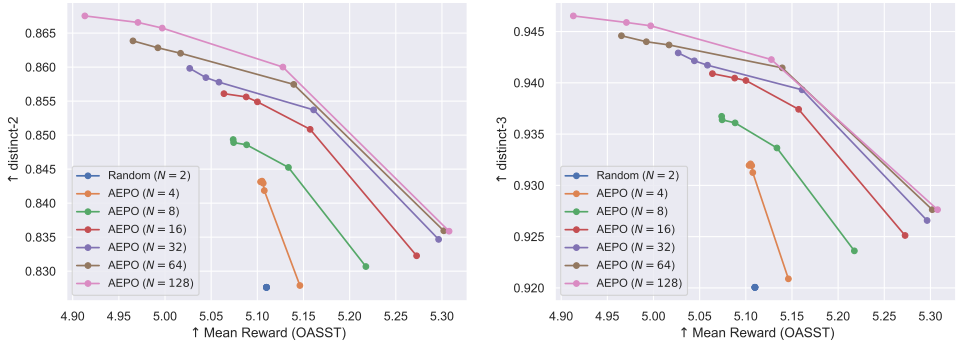
(d) ↑Distinct-3

Figure 15: Diversity (↓Sentence BERT and ↑Distinct-n) and quality (↑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) ↓Sentence BERT

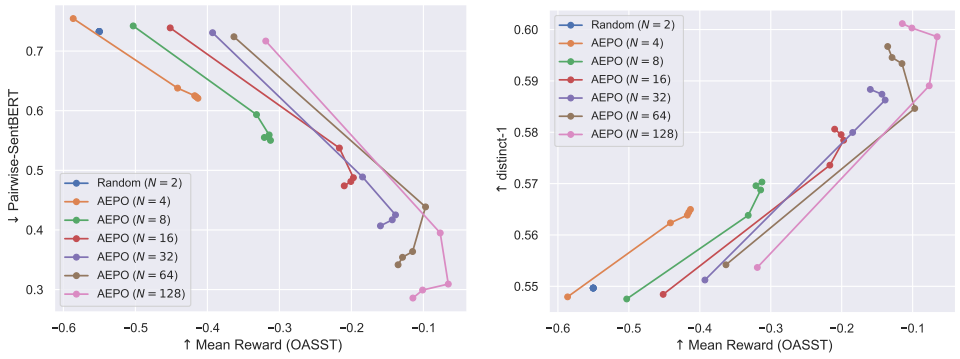
(b) ↑Distinct-1



(c) ↑Distinct-2

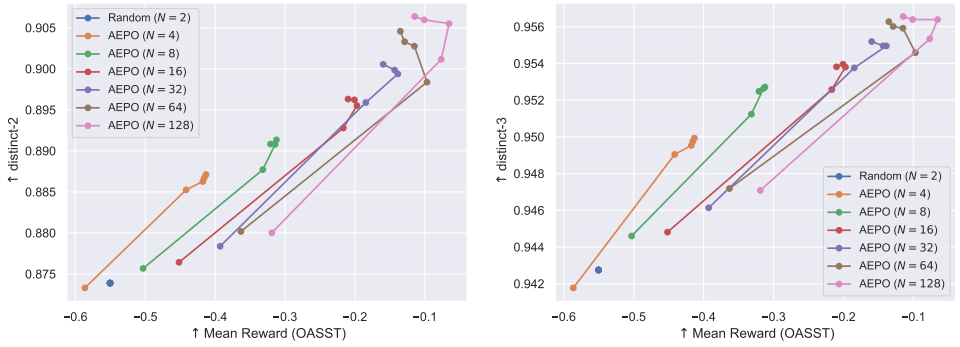
(d) ↑Distinct-3

Figure 16: Diversity (↓Sentence BERT and ↑Distinct-n) and quality (↑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) ↓Sentence BERT

(b) ↑Distinct-1



(c) ↑Distinct-2

(d) ↑Distinct-3

Figure 17: Diversity (↓Sentence BERT and ↑Distinct-n) and quality (↑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.

I LIMITATIONS

Although our method is motivated by the situation where the annotation is needed to align the language model, the majority of our experiments (AlpacaFarm and Anthropic's 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 use an LLM to automatically evaluate the agreement of the response text with the human annotation. Manual human annotation would be desirable for 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.

J 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).

K REPRODUCIBILITY STATEMENT

All the datasets and models used in the experiments are publically accessible (Table 19) except for GPT-4. Our code will be available on acceptance as an open source.

L IMPACT STATEMENT

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.

Table 19: 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