# Preference Ranking Optimization for Human Alignment

**Feifan Song**[1], **Bowen Yu**[2*], **Minghao Li**[2]
**Haiyang Yu**[2], **Fei Huang**[2], **Yongbin Li**[2], **Houfeng Wang**[1*]

[1]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2]Alibaba Group
songff@stu.pku.edu.cn, wanghf@pku.edu.cn
{yubowen.ybw, lmh397008, yifei.yhy, shuide.lyb}@alibaba-inc.com

## Abstract

Large language models (LLMs) often contain misleading content, emphasizing the need to align them with human values to ensure secure AI systems. Reinforcement learning from human feedback (RLHF) has been employed to achieve this alignment. However, it encompasses two main drawbacks: (1) RLHF exhibits complexity, instability, and sensitivity to hyperparameters in contrast to SFT. (2) Despite massive trial-and-error, multiple sampling is reduced to pairwise contrast, thus lacking contrasts from a macro perspective. In this paper, we propose Preference Ranking Optimization (PRO) as an efficient SFT algorithm to directly fine-tune LLMs for human alignment. PRO extends the pair-wise contrast to accommodate preference rankings of any length. By iteratively contrasting candidates, PRO instructs the LLM to prioritize the best response while progressively ranking the rest responses. In this manner, PRO effectively transforms human alignment into aligning the probability ranking of $n$ responses generated by LLM with the preference ranking of humans towards these responses. Experiments have shown that PRO outperforms baseline algorithms, achieving comparable results to ChatGPT and human responses through automatic-based, reward-based, GPT-4, and human evaluations.

## Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in meeting the diverse information needs of users (Brown et al. 2020; Chowdhery et al. 2022; Bubeck et al. 2023; Touvron et al. 2023; Li et al. 2023). Despite leveraging the extensive global knowledge and human behavior encoded within their trillion-token pretraining corpus, LLMs are unavoidably impacted by the existence of misleading, toxic, and detrimental content encompassed within it (Bai et al. 2022b; Ouyang et al. 2022). Consequently, reinforcement learning from human feedback (RLHF) is introduced to construct secure and manageable AI systems (Stiennon et al. 2020; Xue et al. 2023; Peng et al. 2023) by aligning linguistic space of LLMs to human values according to a set of candidates ranked by humans.

Nevertheless, RLHF remains more complex than supervised learning, prone to optimization instability, and sensitive to hyperparameters. These limitations arise mainly from
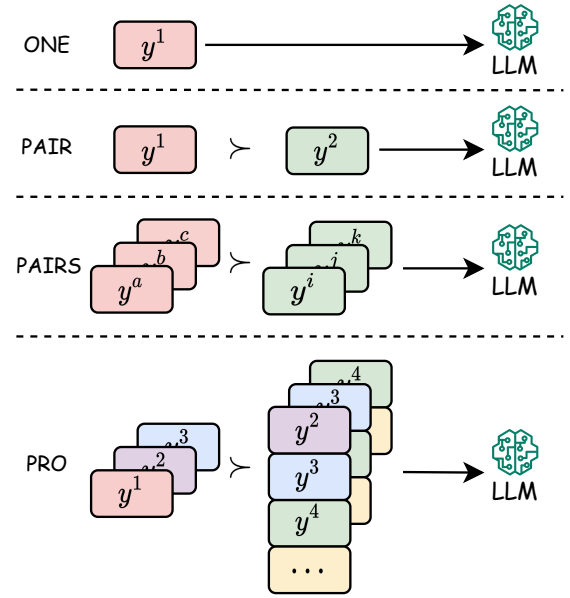
---

Figure 1: Comparison among different SFT paradigms. PRO is distinguished from others for multi-positional and multi-dimensional contrasts.

employing the agent, i.e. LLM to experience repetitive trial-and-error rather than directly aligning it to human preference. Hence, Supervised Fine-Tuning (SFT) is expected as the possibility of more direct optimization to replace RLHF.

SFT initially serves as only a warm-up process for RLHF, where the best candidates are selected to tune LLMs to intimate human-preferred data. Recent works have proposed more complex strategies to enhance SFT (Rafailov et al. 2023; Wu et al. 2023; Yuan et al. 2023; Dong et al. 2023). Despite some progress, there remains room for improvement: (1) The essence that powers RLHF to perform well is ignored. That is multiple sampling with scoring from broad linguistic space during training. Constrained by the given ranking length, most methods pay attention to pair-wise contrasts from semantic or scalar perspectives. (2) Even longer rankings are available, they tend to cut it into pairs, thus lacking distinction of candidates from a macro perspective.

In this work, we thoroughly investigate the effect of enlarging sampling from linguistic space on human alignment. Based on this scenario, we propose the **P**reference **R**anking **O**ptimization (PRO) as an efficient framework of direct policy optimization. Figure 1 shows how PRO stands out from different SFT-based formulations. To be specific, we rethink the essence of RLHF and extend pair-wise contrasts from the Bradley-Terry model (Bradley and Terry 1952) to encompass one-to-N contrasts within a ranking of arbitrary lengths. Then, given an input prompt $x$ and a set of responses ranked by humans, represented as $y^1, y^2, \cdots, y^n$, the proposed PRO algorithm begins by tuning the agent LLM to treat the best response $y^1$ as the positive and the remaining as negatives. This prioritization suggests that the LLM generates $y^1$ with a higher likelihood than those humans consider inferior. It then repetitively ignores the current top response and considers the next one as the positive against the rest, until there are no more responses in the ranking.

Apart from focusing on obtaining more candidates, we particularly deploy proxies of different levels to sample utterances with various qualities and degrees of human alignment. Inspired by RLHF, we also add a self-bootstrapping method to dynamically sample new candidates from the recipient LLM and label it with an additional reward model. The new candidate is added to the original set for training. All of these extended rankings are designed to check the impact of quantity, quality, and diversity of texts.

In general, PRO directly subjects the LLM to a $n$-length human preference ranking. By equipping LLMs with multi-positional and multi-dimensional contrasts among candidates, PRO fully utilizes given ranking sequences of any length. As $n$ approaches infinity, the recipient LLM is exposed to more and more samples with scores, and should continuously become perfectly aligned with human preferences. Especially when $n = 2$, PRO effectively optimizes the LLM using the pair-wise contrast.

We thoroughly evaluate PRO through numerous experiments, including automatic reward model scoring, GPT-4 evaluation, and human evaluation. The main observations are as follows: (1) With a 2-ranking, PRO surpasses the current competitive baselines by a large margin. Also, the high quality and diversity of candidates in preference rankings can be crucial for ultimate performance. (2) The longer the length is, the more prominent improvement PRO can acquire. Even by adding responses generated by ChatGPT to continuously increase the ranking length, PRO achieves reward scores similar to ChatGPT, but with just 7B parameters. (3) Heterogeneous candidates manage to bringing better improvement of PRO than relatively homogeneous ones.

## Preliminary

We commence by providing a brief review of RLHF. In order to train LLMs to generate responses that align with human preferences, RLHF consists of three stages:
(1) Supervised Fine-tuning (SFT) where labelers furnish the desired behavior's response with $t$ tokens, denoted as $y = y_{1,\cdots,t}$, for a given prompt, denoted as $x$. Subsequently, The policy LLM goes through direct fine-tuning (maximum likelihood) on this data, resulting in a model denoted as $\pi_{\text{SFT}}$.

(2) Reward Model (RM) Training where $\pi_{\text{SFT}}$ is utilized by prompts $x$ to generate pairs of responses, which are then denoted by human labelers as a more favored answer $y^1$ against the other one $y^2$, i.e. $y^1 \succ y^2 \mid x$. To predict these preferences, previous works employ the Bradley-Terry (BT) RM, which essentially constructs a pairwise contrast:

$$\mathcal{L}_{\text{RM}} = -\log \frac{\exp\left(r_\phi(x, y^1)\right)}{\exp\left(r_\phi(x, y^1)\right) + \exp\left(r_\phi(x, y^2)\right)} \quad (1)$$

(3) Reinforcement Learning (RL) stage where $\pi_{\text{SFT}}$ is further optimized in a trial-and-error process containing repetitive sampling from linguistic space and corresponding feedbacks simultaneously from the RM and reference policy.

Regrettably, RLHF is criticized for several drawbacks, including increased complexity compared to supervised learning, sensitivity to hyperparameters, and the requirement for additional training of reward models and value networks.

## Methodology

In this section, we first achieve a shift from the pair-wise contrast to PRO that leverages multi-positional one-to-N contrasts. With candidate rankings extending, PRO has access to more samples from linguistic space, and efficiently distinguishes the human-preferred feature of the positive samples from other negative samples. The whole process is completed in SFT settings, thus contributing to avoiding the numerous drawbacks associated with RLHF. Furthermore, we demonstrate the flexibility of our PRO in its ability to integrate with RM, thereby attaining advantages such as affordable preference ranking and differentiated contrast.

### From RLHF to PRO

We re-examine the objective of the Bradley-Terry RM (Equation 1), which helps LLMs understand $y^1 \succ y^2$ through score contrast. The RM is trained in supervised settings and is expected to capture human preference. For a given prompt $x$ and two responses $y^1$ and $y^2$, the RM should prefer $y^1$. To directly optimize the policy model, i.e. the LLM, we can similarly transfer the pair-wise contrast to it. In this way, the LLM is considered as both RM and policy network, denoted as $r_\pi$:

$$\mathcal{L} = -\log \frac{\exp\left(r_\pi(x, y^1)\right)}{\exp\left(r_\pi(x, y^1)\right) + \exp\left(r_\pi(x, y^2)\right)} \quad (2)$$

Naturally, if we expand the candidate set, i.e., increase sampling, $r_\pi$ is able to reach more shots, which replaces the trial-and-error experience. Considering there exist $n$ candidate responses $\{y^i\}$, the human-annotated order $y^{1,\cdots,n}$ is $y^1 \succ y^2 \succ \cdots \succ y^n$. We first define the partial order between $y^1$ and responses behind it as $y^{1,2:n} = y^1 \succ \{y^2, \cdots, y^n\}$. With reference to InfoNCE Loss (He et al. 2020), we derive Equation 2 to a multi-dimensional one-to-N contrast:

$$\mathcal{L} = -\log \frac{\exp\left(r_\pi(x, y^1)\right)}{\sum_{i=1}^n \exp\left(r_\pi(x, y^i)\right)} \quad (3)$$

However, it does not fully leverage the ranking since it only characterizes $y^{1,2:n}$, disregarding the $n-2$ valuable rankings
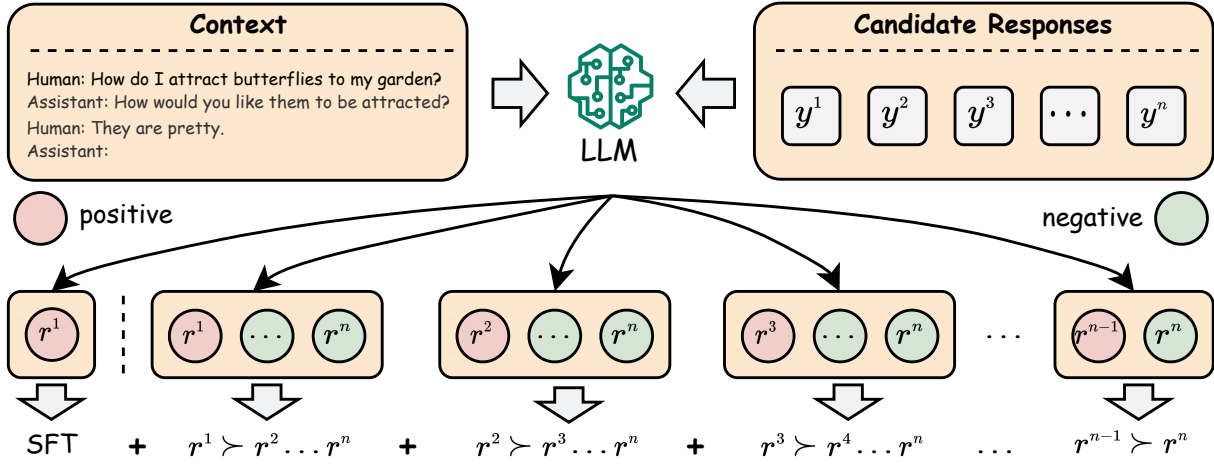
Figure 2: The pipeline of PRO for Human Feedback Alignment learning. Each candidate is concatenated with the prompt first, then processed by the LLM to estimate corresponding rewards, which are optimized by Equation 5.

such as $y^{2,3:n}$ and $y^{n-1,n}$. Instead, recursive contrasts can be imposed to exploit multi-positional patterns, which start with the first response, treat the remaining responses as negatives, drop the current response, and move to the next. This procedure repeats until there are no candidates left. Consequently, the further extension to Equation 3 is as follows:

$$\mathcal{L} = -\log \prod_{k=1}^{n-1} \frac{\exp\left(r_\pi(x, y^k)\right)}{\sum_{i=k}^{n} \exp\left(r_\pi(x, y^i)\right)} \quad (4)$$

Equation 4 provides a comprehensive alignment with human feedback. In addition to adhering to human preferences, it is also desirable for the model to generate fluent replies. Therefore, the original supervised loss that requires the model to fit the responses considered the best by humans can also be incorporated. We conclude the above as the Preference Ranking Optimization (PRO) algorithm, as is demonstrated in Figure 2. Instead of optimizing the agent to approximate the RM, PRO enables the agent LLM to be directly trained by the following objective:

$$\mathcal{L}_{\text{PRO}}(y^{1,\cdots,n} \mid x) = \mathcal{L} + \beta \mathcal{L}_{\text{SFT}} \quad (5)$$

where $\mathcal{L}_{\text{SFT}}$ is the NLL loss of the top 1 candidate and $\beta$ is the hyper-parameter to maintain the balance between text quality and human preference. The policy agent (also RM) $r_\pi$ is parameterized as $r_{\pi_{\text{PRO}}}$ in PRO:

$$r_{\pi_{\text{PRO}}}(x, y^k) = \frac{1}{|y^k|} \sum_{t=1}^{|y^k|} \log P(y_t^k | x, y_{<t}^k) \quad (6)$$

PRO and RLHF share a similar objective, that is, understanding human preferences through more exposure to labeled samples. The difference is that RLHF relies on trial-and-error experience and pair-wise contrasts, whereas PRO learns by assembling multiple samples into long rankings, which can be more efficient.

Surprisingly, Equation 4 has a similar formulation with Plackett-Luce (PL) model (Plackett 1975; Luce 2012), a classic algorithm for ranking aggregation. Believing it is not a coincidence, we assume that PL model and PRO accomplish similar targets. PL model aims to acquire a global ranking of fixed candidates by combining multiple rankings, whose parameters correspond to these candidates, while PRO aims to learn general human preference, but the involved rankings contain different $n$ candidates from each other. With modeling language space, the parameters of agent LLM should theoretically correspond to infinite candidates for each ranking (i.e. $n = \infty$). Although $n$ is limited in practice, the larger it is, the more perfect the LLM should be. We accordingly implement experiments towards $n$ in the next section.

### Grafting RLHF onto PRO

While PRO can be directly valid on the human-annotated preference ranking sequence without the need for introducing concepts like the reward model in RLHF, grafting RLHF onto PRO can bring more flexibility to PRO. We outline three possible upgrades as follows:

**Affordable Preference Ranking.** PRO is highly flexible, relying solely on ranked preference sequences. The source of the sequence is unrestricted, allowing for various possibilities. One approach involves requesting annotators to imagine multiple responses of different quality. Alternatively, a more efficient method entails utilizing different existing LLMs, such as ChatGPT and Alpaca, to generate multiple responses. These responses can be ranked using an additional reward model $r_\phi$, similar to RLHF.

**Differentiated Contrast.** The formulation as shown in Equation 4, treats all responses $y^i \prec y^k$ as negative examples of $y^k$ and applies the same penalty to them. However, this approach may not be reasonable, especially when the preference scores of different $y^i$ are similar. For instance,

when the preference of $y^{k+1}$ is only slightly worse than $y^k$, while $y^n$ is significantly worse than $y^k$, the model should differentiate and apply different penalty strengths, slightly penalizing $y^{k+1}$ and heavily penalizing $y^n$ compared to $y^k$. To address this, we propose using the score $r_\phi(x, y^i)$ from a reward model $r_\phi$ to indicate the numerical preference of $y^i$, and modify Equation 4 as follows:

$$\mathcal{L}_{\text{PRO}} = -\sum_{k=1}^{n-1} \log \frac{\exp\left(\frac{r_{\pi_{\text{PRO}}}(x,y^k)}{\mathcal{T}_k^k}\right)}{\sum_{i=k}^{n} \exp\left(\frac{r_{\pi_{\text{PRO}}}(x,y^i)}{\mathcal{T}_k^i}\right)} \quad (7)$$

where

$$\mathcal{T}_k^{i>k} = \frac{1}{r_\phi(x,y^k) - r_\phi(x,y^i)} \quad (8)$$

$$\mathcal{T}_k^k = \min_{i>k} \mathcal{T}_k^i \quad (9)$$

When the difference between $r_\phi(x, y^k)$ and $r_\phi(x, y^i)$ increases, the preference gap between $y^k$ and $y^i$ becomes more evident. Consequently, the temperature $\mathcal{T}_k^i$ decreases, amplifying the penalty of positive example $y^k$ towards $y^i$, while it decreases when the difference is smaller. $\mathcal{T}_k^k$ is defined as the minimum temperature among all the negative examples to maintain a balance between the numerator and denominator. Our experiments (§) reveal that the dynamic temperature design significantly increases performance when optimizing $\mathcal{L}_{\text{PRO}}$ alone while excluding $\mathcal{L}_{\text{SFT}}$. It also provides some performance gains when jointly optimizing $\mathcal{L}_{\text{PRO}}$ and $\mathcal{L}_{\text{SFT}}$.

## Experiments

### Data Prepration

We choose **HH-RLHF** Bai et al. (2022a) as the experimental dataset. It has 4 sub-sets, namely Harmless$_{\text{base}}$, Helpful$_{\text{base}}$, Helpful$_{\text{online}}$ and Helpful$_{\text{rejection}}$. Each sample contains two different conversations rated by human annotators and is grouped into train/test splits.

To further evaluate the performance of different methods on longer rankings, we augment each sample with new candidates from diverse LLMs to expand the range of ranked responses. These augmented datasets are denoted as *HH-RLHF*$_{\text{LLM},i}$, where LLM represents the language models used (e.g. Alpaca, ChatGPT), and $i$ is the length of the rankings. The unmodified dataset is referred to as *HH-RLHF*raw.

**Disclaimer:** Data augmentation and inference from Curie/ChatGPT, as well as the following GPT-4 evaluation, are completed where the related services are available.

### Evaluation Metrics

We present the findings using various evaluation methods: automatic, model-based, and human-based metrics. Specifically, we utilize BLEU (Papineni et al. 2002) to assess the text quality and RMs to measure the level of human preference gained. To avoid unfairness, we select two different RMs for training and evaluation, which we denote as RM$_{\text{train}}$ and RM$_{\text{eval}}$, respectively. These metrics allow us to automatically evaluate numerous models. Human evaluation is the gold standard for assessing human preferences (Zhou et al. 2023). An annotator judge is presented with a question

and two responses and tasked with determining the better option or declaring a tie. Furthermore, recent studies have shown that GPT-4 (OpenAI 2023) effectively evaluates the responses of chat assistants and aligns with human preferences (Zheng et al. 2023; Wang et al. 2023). Consequently, we involve GPT-4 to select the best from the two options. To mitigate positional bias (Zheng et al. 2023; Wang et al. 2023), we evaluate each candidate in both positions during two separate runs, and the final score is computed as the average of the two runs.

### Implementation Details

We choose the popular LLaMA-7B (Touvron et al. 2023) as the backbone model, and implement PRO using Transformers (Wolf et al. 2020) and Accelerate (Gugger et al. 2022).

We assign $\beta$, the weight SFT loss, to $0.05*(l-1)^2$ where $l$ is the ranking length. The sequence length, epoch, and learning rate are set to 512, 2, and 5e-6, respectively, while the maximum number of new tokens generated during inference is 128, and the total batch size is 112.

Moreover, the expanded candidate rankings in each augmented dataset need to be re-sorted. However, the numerous manual sortings are time-consuming and costly. Therefore, we employ RM$_{\text{train}}$ to score and rearrange all candidate rankings during the pre-processing stage of training. In addition, values from RM$_{\text{eval}}$ will be normalized through Sigmoid function in case it occasionally provides extreme values that excessively influence the overall performance. RM$_{\text{train}}$ and RM$_{\text{eval}}$ are all implemented using open-source checkpoints. More particulars can be found in our code[1].

### Main Experiment

We compare PRO with several LLMs (zero-shot), as well as baselines for fine-tuning LLaMA-7B (Touvron et al. 2023). Table 1 contains the results. It can be found that each fine-tuned LLaMA-7B gets a notable improvement on BLEU and Reward against the initial version without any specific alignment with human preference. Also, even without fine-tuning on *HH-RLHF*, LLMs can still show certain performance, while ChatGLM (Du et al. 2022) and ChatGPT with RLHF training beat LLaMA-7B, Curie (Brown et al. 2020), and Alpaca-7B (Taori et al. 2023). All of these prove the significance of Human Alignment.

Next, we compare PRO with strong baselines on the same dataset using LLaMA-7B, including SFT, RLHF, CoH (Liu, Sferrazza, and Abbeel 2023), DPO (Rafailov et al. 2023) and RRHF (Yuan et al. 2023). In general, PRO can outperform all baselines, or show competitive performance in terms of Reward score while maintaining considerable BLEU scores. Specifically, even in the basic *HH-RLHF*$_{\text{raw}}$ containing just rankings of 2 candidates, PRO achieves a 6.52 improvement of Reward over SFT, and 2.6 over the well-performed DPO. CoH (Liu, Sferrazza, and Abbeel 2023) gets higher BLEU scores but falls short of PRO in Reward, which is mediocre.

PRO exhibits a distinct advantage in terms of Harmlessness compared to Helpfulness. We attribute this to the fact that achieving Harmlessness is comparatively easier

---

[1]github.com/AlibabaResearch/DAMO-ConvAI/tree/main/PRO

| Training Set | Method | Harmless$_{base}$ | | Helpful$_{base}$ | | Helpful$_{online}$ | | Helpful$_{rejection}$ | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BLEU | Reward | BLEU | Reward | BLEU | Reward | BLEU | Reward | BLEU | Reward |
| Zero-shot | LLaMA | 10.82 | 51.16 | 12.78 | 31.71 | 15.02 | 38.91 | 14.60 | 34.85 | 13.13 | 38.94 |
| | Curie | 14.23 | 50.71 | 17.33 | 45.51 | 17.11 | 51.36 | 18.99 | 48.68 | 16.99 | 48.71 |
| | Alpaca | 15.07 | 53.03 | 19.68 | 49.80 | 18.77 | 55.74 | 22.21 | 53.72 | 19.12 | 52.72 |
| | ChatGLM | 15.39 | 63.30 | 20.16 | 59.14 | 30.99 | 61.10 | 25.41 | 61.45 | 21.99 | 61.27 |
| | ChatGPT | 15.51 | 71.44 | 21.38 | 65.94 | 29.81 | 67.94 | 26.52 | 68.39 | 22.56 | 68.48 |
| *HH-RLHF*$_{raw}$ | SFT | 15.07 | 55.96 | 20.40 | 41.36 | 29.36 | 54.08 | 25.54 | 47.08 | 21.80 | 48.83 |
| | RLHF | 14.54 | 55.05 | 19.86 | 42.16 | 28.04 | 53.40 | 25.11 | 47.73 | 21.19 | 48.93 |
| | CoH | 13.34 | 45.47 | 23.17 | 39.03 | 33.84 | 52.63 | 29.79 | 46.57 | 24.06 | 45.00 |
| | DPO | 16.29 | 54.43 | 21.37 | 50.13 | 27.73 | 54.09 | 26.91 | 53.04 | 22.62 | 52.75 |
| | RRHF | 13.49 | 53.98 | 18.76 | 48.23 | 30.68 | 56.44 | 24.95 | 52.51 | 20.91 | 52.25 |
| | **PRO** | 12.05 | 62.96 | 20.83 | 48.51 | 28.75 | 59.02 | 27.17 | 53.28 | 21.54 | 55.35 |
| *HH-RLHF*$_{Alpaca,3}$ | BoN | 16.75 | 59.24 | 22.81 | 54.04 | 29.89 | 61.00 | 27.76 | 58.04 | 23.7 | 57.66 |
| | RLHF | 16.33 | 56.61 | 23.12 | 54.85 | 30.54 | 60.97 | 27.94 | 58.4 | 23.82 | 57.28 |
| | CoH | 13.71 | 47.36 | 22.45 | 42.34 | 33.17 | 53.19 | 28.76 | 48.61 | 23.54 | 47.15 |
| | DPO | 16.37 | 63.93 | 21.82 | 55.86 | 27.84 | 58.49 | 27.53 | 58.60 | 22.98 | 59.27 |
| | RRHF | 12.79 | 54.18 | 19.21 | 53.23 | 31.53 | 59.04 | 25.14 | 56.76 | 21.02 | 55.39 |
| | **PRO** | 14.41 | 62.60 | 22.47 | 54.38 | 25.61 | 60.90 | 26.82 | 58.26 | 22.11 | 58.72 |
| *HH-RLHF*$_{ChatGPT,3}$ | BoN | 15.05 | 67.85 | 20.77 | 60.43 | 31.27 | 64.36 | 26.47 | 63.14 | 22.45 | 63.83 |
| | RLHF | 13.63 | 61.97 | 20.12 | 55.29 | 28.89 | 59.78 | 24.65 | 58.26 | 20.99 | 58.65 |
| | CoH | 13.44 | 56.87 | 21.89 | 51.52 | 34.04 | 59.51 | 28.24 | 56.35 | 23.26 | 55.58 |
| | DPO | 15.63 | 67.81 | 21.00 | 61.86 | 29.01 | 61.90 | 26.39 | 63.81 | 22.35 | 64.10 |
| | RRHF | 13.02 | 64.63 | 18.95 | 61.38 | 31.37 | 63.26 | 24.75 | 63.28 | 20.86 | 63.12 |
| | **PRO** | 15.53 | 73.08 | 22.30 | 64.78 | 29.35 | 66.66 | 27.49 | 66.95 | 23.07 | 67.97 |

Table 1: Main Results. PRO consistently acquires more reward than all fine-tuned baselines, while is close to or even exceeding ChatGLM and ChatGPT.
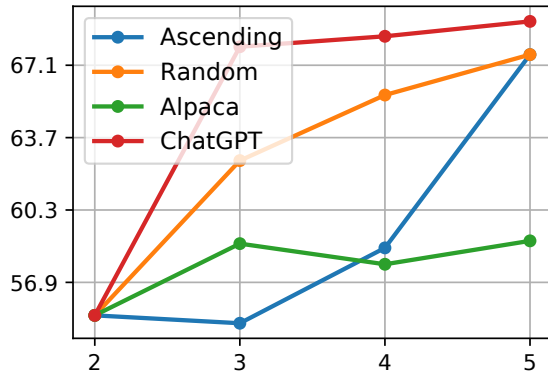


Figure 3: Results of experiments on different lengths.

for PRO as it primarily involves significant features such as adapting expression styles and maintaining politeness in most conversations. On the other hand, **Helpfulness** typically demands specific suggestions, which pose a greater challenge for language models due to their limited world knowledge, thus increasing the difficulty in this aspect.

When expanding the ranking sequence using ChatGPT and sorting it with RM$_{train}$, the performance of each method also increases. On the expanded sequences, we observe that BoN (selecting the response with the highest reward model score for SFT) becomes a competitive baseline. This finding aligns with Rafailov et al. 2023, who observed that RLHF is

less tuning-efficient than BoN. The effectiveness of RRHF becomes less prominent because it relies on pairwise contrasts between candidates from given rankings. It fails to capture global differences corresponding to human preference in the long rankings, which can be achieved through Equation 4. Overall, in the expanded ranking, PRO remains the most competitive method, and the more powerful the LLM used for ranking augmentation, the more pronounced the improvement of PRO. This surprising characteristic fills us with anticipation for PRO's future development.

## Effect of Expanding Preference Ranking Sequence

In Table 1, we have observed that expanding the ranking sequence of *HH-RLHF*$_{raw}$ from length 2 to 3 using LLMs improves the performance of all models. This leads us to wonder how the effect would change if we further expand the preference ranking sequence. Specifically, we simulate 4 expansion strategies, each introducing 3 additional responses to extend the preference sequence to length 5, followed by reranking using a reward model.

**Alpaca:** Using Alpaca-7B, we generate 3 responses, adding 1, 2, and 3 responses, respectively, to form ranking sequences of lengths 3, 4, and 5.

**ChatGPT:** Using ChatGPT, we generate three responses, adding 1, 2, and 3 responses, respectively, to form ranking sequences of lengths 3, 4, and 5.

**Ascending:** We utilize three LLMs, namely Curie, Alpaca-7B, and ChatGPT. Based on the zero-shot results in Table 1, the quality of their responses can be ranked as ChatGPT ≻

Alpaca-7B $\succ$ Curie. In this setting, we add the responses in ascending order of quality, i.e. Curie's response in rankings of length 3, Curie and Alpaca-7B's responses in rankings of length 4 while Curie, Alpaca-7B, and ChatGPT's responses in rankings of length 5.

**Random:** The order of response additions is unrelated to response quality and is done randomly.

Figure 3 presents the impact of various expansion strategies on the effectiveness of PRO after expanding sequences of different lengths. Our observations are as follows:

**Longer ranking, better results:** Overall, longer ranking sequences generally lead to improved performance for most strategies, as longer rankings embrace more candidates. It implies that more sampling from linguistic space with feedback labels effectively helps LLMs align with human preference. This is an exciting finding because with a well-performed RM, which is relatively easy to obtain, expanding rankings can be a straightforward task compared to brainstorming for new prompts.

**Better added responses, better results:** If a single model is used to generate additional responses, supplementing one response is sufficient when the quality is average, such as with Alpaca, adding more responses provides limited improvement. However, when the quality of responses is high, as with ChatGPT, adding more responses leads to consistent performance gains. This could potentially offer new insights for the design of future Human Alignment algorithms.

**More diversified added responses, better results:** Incorporating lower-quality responses may better improve the LLM compared to using only high-quality responses. Interestingly, when the sequence length is 4, Ascending (blue line) = Curie + Alpaca surpasses the performance of Alpaca (green line) = Alpaca + Alpaca, even though Curie's response quality is not as good as Alpaca's. We attribute it to the fact that diverse responses, even if they are negative examples, help the language model become more aware of behaviors that should be avoided, thereby enhancing overall performance. Lastly, by combining Curie, Alpaca, and ChatGPT, we achieve a performance close to using three ChatGPT responses, demonstrating the truth in the saying, "Two heads are better than one."

## Human and GPT-4 Evaluation

Compared with reward models which may have a distortion in capturing human preferences, human annotation is considered the most accurate evaluation method, and recently, GPT-4-as-a-judge has emerged as a scalable approach for rapidly assessing human preference.

To verify whether PRO truly captures human preferences, we provide comprehensive evaluations conducted by both GPT-4 and humans. We hereby investigate the performance of **PRO vs. Golden**, i.e. the 1st candidate provided by the datasets. In detail, we aim to determine whether PRO trained on *HH-RLHF*<sub>raw</sub> can achieve or surpass human-preferred responses provided by the raw dataset, which contains ranking sequences of length 2 that do not fully exploit PRO's capabilities. On the other hand, this comparison serves as evidence to some extent for the validity of the reward model we use in evaluation.

| Evaluator | Sub-set | Win | Tie | Lose |
|---|---|---|---|---|
| GPT-4 | Harmless$_{\text{base}}$ | 60.00 | 5.00 | 35.00 |
| | Helpful$_{\text{base}}$ | 77.50 | 0.00 | 22.50 |
| | Helpful$_{\text{online}}$ | 27.50 | 12.50 | 60.00 |
| | Helpful$_{\text{rejection}}$ | 55.00 | 0.00 | 45.00 |
| | Average | **55.00** | 4.37 | 40.63 |
| Human | Harmless$_{\text{base}}$ | 20.00 | 55.00 | 25.00 |
| | Helpful$_{\text{base}}$ | 20.00 | 60.00 | 20.00 |
| | Helpful$_{\text{online}}$ | 20.00 | 50.00 | 30.00 |
| | Helpful$_{\text{rejection}}$ | 30.00 | 60.00 | 10.00 |
| | Average | **22.50** | 56.25 | 21.25 |

Table 2: Results of GPT-4 and Human Evaluation.

For GPT-4 evaluation, we first sample contexts in test sets. We assemble two corresponding responses from PRO and its counterpart into a modified version of the prompt template from Zheng et al. (2023) for GPT-4 scoring. We also refer to Wang et al. (2023) to provide two candidates in binary directions respectively, to eliminate unfairness triggered by order. For Human evaluation, we employ 3 annotators to estimate the same samples with GPT-4 evaluation, and directly distinguish one from another between two shuffled responses.

Table 2 gives the detailed results, where both GPT-4 and humans globally support PRO more, thus highlighting the strengths of PRO. This suggests that PRO is able to effectively capture the preferences of humans as reflected in the annotated data. Furthermore, our evaluation using the reward model yielded consistent results, with both humans and GPT-4 significantly favoring PRO. This not only reaffirms the effectiveness of PRO but also demonstrates that our reward model can reasonably evaluate human preferences.

## Ablation Study

In this part, we investigate the effectiveness of each part in PRO. Table 3 presents results.

**SFT Loss** To avoid the model solely catering to the reward model at the expense of text quality, we introduce $\mathcal{L}_{\text{SFT}}$. Therefore, removing $\mathcal{L}_{\text{SFT}}$ lowers BLEU scores.

**PRO Loss** Table 1 also demonstrates the influence of $\mathcal{L}_{\text{PRO}}$, as excluding it in PRO essentially equals to SFT (BoN) that gets lower Reward.

**Adequate Ranking** To fully leverage the ranking $y^{1,\cdots,n}$, we employ $n-1$ loss functions to model $y^{1,2:n}$, $y^{2,3:n}$, ..., $y^{n-1,n}$. Our objective is to adequately model all ranking orders and enable LLM to better differentiate between samples of different preferences. To validate this idea, we deactivate $\mathcal{L}_{\text{PRO}}$ except for its first term, $\mathcal{L}_{\text{PRO}}^1$. Experimental results demonstrate a decrease in both BLEU and Reward scores, thus confirming the effectiveness of Equation 4.

**Temperature** $\mathcal{T}$ slightly enhances overall performance. However, we observe a significant drop in performance when both $\mathcal{L}_{\text{SFT}}$ and $\mathcal{T}$ are removed simultaneously, whereas removing either one individually did not have such a noticeable impact. We believe this is because temperature helps

| Traning set | Method | Harmless$_{\text{base}}$ | | Helpful$_{\text{base}}$ | | Helpful$_{\text{online}}$ | | Helpful$_{\text{rejection}}$ | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BLEU | Reward | BLEU | Reward | BLEU | Reward | BLEU | Reward | BLEU | Reward |
| *HH-RLHF*$_{\text{raw}}$ | **PRO** | 12.05 | 62.96 | 20.83 | 48.51 | 28.75 | 59.02 | 27.17 | 53.28 | 21.54 | 55.35 |
| | $-\mathcal{L}_{\text{SFT}}$ | 6.94 | 67.20 | 10.37 | 46.60 | 11.17 | 49.33 | 11.32 | 48.84 | 9.85 | 53.25 |
| | $-\mathcal{T}$ | 12.04 | 62.91 | 20.63 | 47.92 | 28.73 | 58.52 | 26.94 | 53.08 | 21.41 | 55.04 |
| | $-\mathcal{L}_{\text{SFT}} - \mathcal{T}$ | 0.88 | 52.81 | 6.74 | 42.97 | 6.37 | 42.84 | 6.85 | 44.71 | 5.14 | 46.17 |
| *HH-RLHF*$_{\text{Alpaca,3}}$ | **PRO** | 14.41 | 62.6 | 22.47 | 54.38 | 25.61 | 60.90 | 26.82 | 58.26 | 22.11 | 58.72 |
| | $-\mathcal{L}_{\text{PRO}}^{k>1}$ | 13.38 | 62.88 | 21.50 | 53.48 | 24.56 | 60.32 | 25.81 | 57.15 | 21.10 | 58.11 |
| | $-\mathcal{L}_{\text{SFT}}$ | 9.06 | 65.78 | 18.77 | 54.18 | 23.90 | 62.26 | 23.33 | 58.29 | 18.29 | 59.71 |
| | $-\mathcal{T}$ | 13.71 | 63.40 | 21.70 | 53.77 | 24.84 | 60.36 | 26.01 | 57.34 | 21.34 | 58.40 |
| | $-\mathcal{L}_{\text{SFT}} - \mathcal{T}$ | 0.52 | 55.90 | 2.13 | 23.41 | 3.56 | 23.44 | 2.66 | 23.82 | 2.05 | 32.33 |
| *HH-RLHF*$_{\text{ChatGPT,3}}$ | **PRO** | 15.53 | 73.08 | 22.30 | 64.78 | 29.35 | 66.66 | 27.49 | 66.95 | 23.07 | 67.97 |
| | $-\mathcal{L}_{\text{PRO}}^{k>1}$ | 15.20 | 72.64 | 21.94 | 64.44 | 29.17 | 66.97 | 27.29 | 66.80 | 22.80 | 67.75 |
| | $-\mathcal{L}_{\text{SFT}}$ | 13.81 | 73.18 | 21.28 | 64.20 | 27.90 | 67.15 | 26.57 | 66.76 | 21.84 | 67.84 |
| | $-\mathcal{T}$ | 15.77 | 72.99 | 22.13 | 65.34 | 29.03 | 67.48 | 27.28 | 67.54 | 22.98 | 68.40 |
| | $-\mathcal{L}_{\text{SFT}} - \mathcal{T}$ | 5.93 | 69.61 | 5.22 | 33.92 | 9.33 | 31.81 | 6.11 | 33.52 | 6.25 | 43.16 |

Table 3: Ablation results. We investigate the effectiveness of $\mathcal{L}_{\text{PRO}}$, $\mathcal{L}_{\text{SFT}}$ and the dynamic temperature $\mathcal{T}$.

the LLM understand that some negative examples are neutral (with reward scores similar to positive examples), and thus should not be overly penalized to avoid confusion during LLM training. The inclusion of $\mathcal{L}_{\text{SFT}}$ plays a similar role by increasing the weight of the best response.

## Related Work

### Reinforcement Learning from Human Feedback

Fine-tuning LLMs to align with human preferences has emerged as a critical research problem. It can be formulated as given a context and corresponding suffixes ranked or scored by human annotators without more detailed labels, the agent is required to learn human preference and provide human-like results. Reinforcement Learning (RL) can be the most straightforward way to reach this goal, for the agent just scarce supervision signal from reward models as human proxies, then is modified through numerous trials under RL framework, namely **R**einforcement **L**earning from **H**uman **F**eedback (RLHF). Many explorations have been done on this path (Christiano et al. 2017; MacGlashan et al. 2017; Warnell et al. 2018; Ziegler et al. 2019; Stiennon et al. 2020; Nakano et al. 2021; Lee, Smith, and Abbeel 2021; Lei et al. 2022; Snell et al. 2022; Bai et al. 2022a; Ouyang et al. 2022; Zhu, Jiao, and Jordan 2023; Zhu et al. 2023). Stiennon et al. (2020) and Nakano et al. (2021) investigate the RLHF method for text summarization and question answering, respectively. Bai et al. (2022a) apply RLHF to enable LLMs to be harmless and helpful while releasing a new dialogue dataset with human feedback. Zhu, Jiao, and Jordan (2023) provide the bound of reward learning if formulated as the Bradely-Terry model and Plackett-Luce model. Known as a masterpiece, Ouyang et al. (2022) propose InstructGPT which first goes through SFT, then is continually modified with PPO algorithm (Schulman et al. 2017). This process is cyclic, during which the performance of the trained agent spirals upwards. The famous ChatGPT inherits it.

### SFT for Human Preference Alignment

Despite appealing advantages, RLHF has obvious limitations regarding training efficiency and complexity, consequently driving researchers to focus on SFT without these challenges. Liu, Sferrazza, and Abbeel (2023) combine desirable and undesirable suffixes in a template prompted by opposite keywords, fully depending on a highly semantic understanding of LLMs. Dong et al. (2023) rely on RMs to select data sampled from the tuned model itself, which in turn are utilized to extend the process of fine-tuning. Yuan et al. (2023) compose multiple pairwise contrasts between suffixes in the given ranking, which forms a new algorithm from the perspective of training objectives. Rafailov et al. (2023) similarly transform LLMs as a BT model to measure chosen and rejected candidates by human annotators. PRO chooses the path of modifying the SFT objective, but is further promoted from RLHF and inherits its straightforwardness. It transforms RL's indirect optimization into a direct one and extends pair-wise contrasts to multi-positional and multi-dimensional contrasts.

## Conclusion

In this paper, we derive from pair-wise contrasts of reward models in RLHF that human alignment can be modeled as aligning the probability ranking of $n$ responses generated by the LLM and the preference ranking of these responses by humans. Based on this derivation, we propose PRO algorithms. PRO inherits the advantages of RLHF, and further captures fine-grained distinction corresponding to human preference from multiple one-to-N contrasts. We conduct extensive experiments to verify the excellence of PRO against baselines and investigate the impact of multifaceted factors. Overall, the findings presented in this paper demonstrate the significance of PRO in effectively and efficiently aligning LLMs to human preference. This work can serve as a stepping stone for further quantifiable explorations.

## Ethics Statement

## Acknowledgments

## References

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022b. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

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

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv:2303.12712.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. arXiv:2204.02311.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Dong, H.; Xiong, W.; Goyal, D.; Pan, R.; Diao, S.; Zhang, J.; Shum, K.; and Zhang, T. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. arXiv:2304.06767.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335. Dublin, Ireland: Association for Computational Linguistics.

Gugger, S.; Debut, L.; Wolf, T.; Schmid, P.; Mueller, Z.; and Mangrulkar, S. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lee, K.; Smith, L.; and Abbeel, P. 2021. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. arXiv:2106.05091.

Lei, W.; Zhang, Y.; Song, F.; Liang, H.; Mao, J.; Lv, J.; Yang, Z.; and Chua, T.-S. 2022. Interacting with Non-Cooperative User: A New Paradigm for Proactive Dialogue Policy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, 212–222. New York, NY, USA: Association for Computing Machinery. ISBN 9781450387323.

Li, M.; Song, F.; Yu, B.; Yu, H.; Li, Z.; Huang, F.; and Li, Y. 2023. Api-bank: A benchmark for tool-augmented llms. arXiv:2304.08244.

Liu, H.; Sferrazza, C.; and Abbeel, P. 2023. Chain of Hindsight aligns Language Models with Feedback. arXiv:2302.02676.

Luce, R. D. 2012. *Individual choice behavior: A theoretical analysis*. Courier Corporation.

MacGlashan, J.; Ho, M. K.; Loftin, R.; Peng, B.; Wang, G.; Roberts, D. L.; Taylor, M. E.; and Littman, M. L. 2017. Interactive Learning from Policy-Dependent Human Feedback. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 2285–2294. PMLR.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. arXiv:2112.09332.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction tuning with gpt-4. arXiv:2304.03277.

Plackett, R. L. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2): 193–202.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

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

Snell, C.; Kostrikov, I.; Su, Y.; Yang, M.; and Levine, S. 2022. Offline rl for natural language generation with implicit language q learning. arXiv:2206.11871.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. arXiv:2302.13971.

Wang, P.; Li, L.; Chen, L.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023. Large language models are not fair evaluators. arXiv:2305.17926.

Warnell, G.; Waytowich, N.; Lawhern, V.; and Stone, P. 2018. Deep TAMER: Interactive Agent Shaping in High-Dimensional State Spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. arXiv:2306.01693.

Xue, W.; An, B.; Yan, S.; and Xu, Z. 2023. Reinforcement Learning from Diverse Human Preferences. arXiv:2301.11774.

Yuan, Z.; Yuan, H.; Tan, C.; Wang, W.; Huang, S.; and Huang, F. 2023. Rrhf: Rank responses to align language models with human feedback without tears. arXiv:2304.05302.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2023. Lima: Less is more for alignment. arXiv:2305.11206.

Zhu, B.; Jiao, J.; and Jordan, M. I. 2023. Principled Reinforcement Learning with Human Feedback from Pairwise or $K$-wise Comparisons. arXiv:2301.11270.

Zhu, B.; Sharma, H.; Frujeri, F. V.; Dong, S.; Zhu, C.; Jordan, M. I.; and Jiao, J. 2023. Fine-Tuning Language Models with Advantage-Induced Policy Alignment. arXiv:2306.02231.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. arXiv:1909.08593.