# Disentangling Length from Quality in Direct Preference Optimization

**Anonymous ACL submission**

## Abstract

Reinforcement Learning from Human Feedback (RLHF) has been a crucial component in the recent success of Large Language Models. However, RLHF is know to exploit biases in human preferences, such as verbosity. A well-formatted and eloquent answer is often more highly rated by users, even when it is less helpful and objective. A number of approaches have been developed to control those biases in the classical RLHF literature, but the problem remains relatively under-explored for Direct Alignment Algorithms such as Direct Preference Optimization (DPO). Unlike classical RLHF, DPO does not train a separate reward model or use reinforcement learning directly, so previous approaches developed to control verbosity cannot be directly applied to this setting. Our work makes several contributions. For the first time, we study the length problem in the DPO setting, showing significant exploitation in DPO and linking it to out-of-distribution bootstrapping. We then develop a principled but simple regularization strategy that prevents length exploitation, while still maintaining improvements in model quality. We demonstrate these affects across datasets on summarization and dialogue, where we achieve up to 20% improvement in win rates when controlling for length, despite the GPT4 judge's well-known verbosity bias.

## 1 Introduction

Recently Large Language Models (LLMs) have seen significant improvements in capabilities, such as code-generation, mathematical reasoning, and tool use. Importantly, they can now fluently interact with users and follow their instructions, leading to their widespread adoption. Fine-tuning with Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2022) has been a significant component in those advances and is now a standard part of advanced LLM training pipelines (Ouyang et al., 2022; Bai et al., 2022a; Touvron et al., 2023; Jiang et al., 2024; Anil et al., 2023). Currently, all the leading LLMs deploy some sort of RLHF pipeline (Dubois et al., 2024; Zheng et al., 2023; Liang et al., 2023). The classical approach consists of three-stages. The first stage begins with a general model pre-trained with next-token prediction on a large corpus of text (Radford et al., 2019; Brown et al., 2020), which is then further-tuned for instruction-following purposes (Wei et al., 2022). In the second stage, the model is prompted with general requests, and generates multiple possible answers, which are then ranked by the user. These ratings are used to train a reward model, which represents human preferences (Christiano et al., 2017; Stiennon et al., 2022; Ziegler et al., 2020; Bai et al., 2022a; Touvron et al., 2023). In the final stage, the instruction-tuned LLM is further trained to maximize expected rewards from the reward model trained in the second stage (a proxy for user preferences) using general purpose reinforcement learning algorithms (Schulman et al., 2017; Mnih et al., 2016). While successful, this pipeline is quite technically complex, and

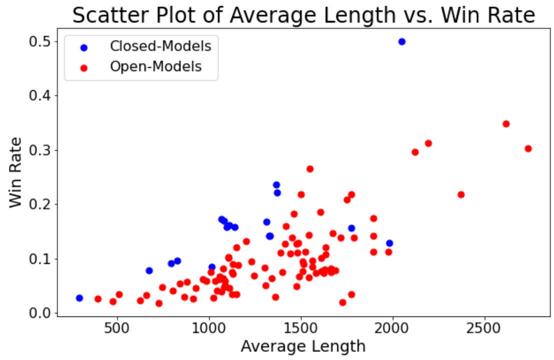


Figure 1: Average win rates versus generation length on the Alpaca Eval benchmark (Dubois et al., 2024). While the highest-scoring open-source models can match the overall performance of strong closed models, they lag significantly on length-corrected basis. Notable outliers are the Cohere Command and GPT4 models.

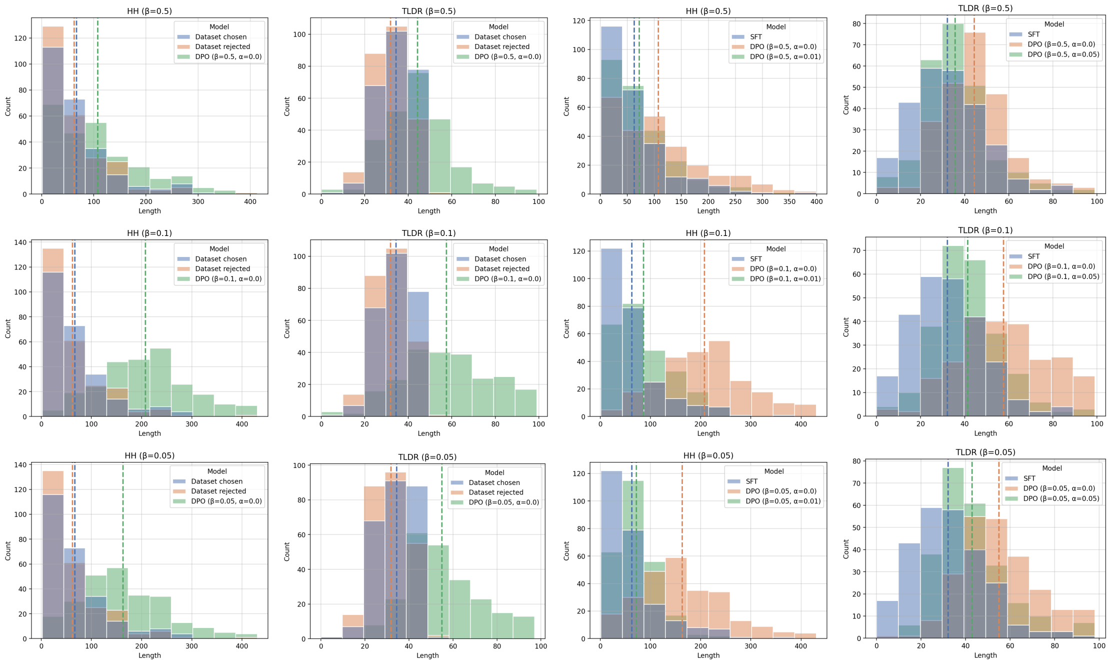


Figure 2: Distribution of response lengths of human feedback datasets, average length is marked by the dashed line. **First Column:** Statistics on Anthropic's Helpful and Harmless dialogue dataset (Bai et al., 2022b). **Second Column:** Statistics on the Reddit TL;DR summarization dataset (Stiennon et al., 2022). While both datasets exhibit a small bias in preference towards longer responses, the un-regularized DPO model produces answers twice as long on average, with lengths significantly out of distribution of the feedback dataset. **Third and Fourth Columns:** Comparison between the SFT, DPO and length-regularized DPO models on HH and TLDR respectively. While length-regularized DPO algorithm still generates longer answers on average, it stays closer to the SFT model.

computationally expensive, mainly due to the final stage of RL optimization.

The quality of the learned reward model is crucial for the RLHF process (Touvron et al., 2023). However, prior works have demonstrated that reward models can be exploited (Casper et al., 2023; Gao et al., 2023) due to a Goodhart's law effect (Clark and Amodei, 2016; Manheim and Garrabrant, 2019; Skalse et al., 2022; Lambert and Calandra, 2023). Under this phenomenon, the model can achieve high rewards during the RL training while generating undesirable behaviours (Gao et al., 2023; Dubois et al., 2024). A particular case of the reward exploitation phenomenon is the well-known verbosity issue - models fine-tuned with RLHF generate significantly longer answers, without necessarily improving the actual quality (Singhal et al., 2023; Kabir et al., 2023). This has been linked to an explicit bias in the preference data towards longer responses (Singhal et al., 2023), however, the statistical increase in verbosity of RLHF-trained models significantly outmatches the the difference of distribution lengths between the preferred and rejected answers. This effect is even observed in in strong propriety models, such as GPT4 (John Schulman et al., 2022), which is now frequently used to evaluate the performance of other LLMs (Dubois et al., 2024; Zheng et al., 2023; Zeng et al., 2023). However, even as an evaluator GPT4, exhibits strong preferences for length. Prior work (Wang et al., 2023) has noted that when evaluating 13B parameter models in head-to-head comparisons with the Davinci-003 model, win rates and the average number of unique tokens in the model's response have correlation of 0.96.

Recently Direct Preference Optimization (Rafailov et al., 2023) has emerged as an alternative to the standard RLHF pipeline. The key observation of DPO is that the reward model can directly be re-parameterized through the optimal LLM policy obtained in the reinforcement learning stage. This allows us to directly train the language model through the reward learning pipeline, eliminating the need for the reinforcement learning stage. This algorithm has become widely used, since it can train completely offline, yielding better simplicity of tuning, speed and resource efficiency, while maintaining performance (Dubois et al., 2024; Jiang et al., 2024). For these reasons it has also been widely adopted by the open-source community. At the time of this writing, 9 out of the top 10 models on the HuggingFace Open LLM Leaderboard use DPO as part of their training pipeline.

While the question of length exploitation has been extensively studied in the classical RLHF pipeline, it has not been explored in the DPO setting before. Moreover, recently concerns have been raised that open-source models have not improved significantly across automated benchmarks, but instead have been exploiting the verbosity bias of the evaluator (Liu, 2024). These statistics are demonstrated in Figure 1, as open-source models can match the overall performance of proprietary ones, but lag significantly on length-corrected basis.

**We make several contributions in our work:** First we study the length exploitation problem in the DPO setting and and show it is quite persistent, which we empirically link to out-of-distribution bootstrapping. Next, we derive a simple but efficient regularization approach, which we show can effectively control verbosity, without impacting model performance, even when evaluated by a biased judge, such as GPT4.

## 2 Preliminaries

In this section we will outline the core components of the standard RLHF pipeline Ziegler et al.; Stiennon et al.; Bai et al.; Ouyang et al.) and the Direct Preference Optimization algorithm (Rafailov et al., 2023), which is central to our analysis and regularization derivations.

### 2.1 Reinforcement Learning From Human Feedback

The standard RLHF pipeline consists of three stages: 1) We first pre-train a general LLM for instruction-following purposes with supervised fine-tuning (SFT); the Reward Modelling stage consists of gathering human feedback and training a parameterized reward model; finally during the final Reinforcement Learning stage, we further optimize the LLM in a reinforcement learning loop, uing the trained reward model from the previous stage.

**SFT**: During this stage, we use a dataset of prompts $\mathbf{x}$ and high-quality answers $\mathbf{y}$ to train an LLM with next-token prediction to obtain a model $\pi_{\text{SFT}}(\mathbf{y}|\mathbf{x})$. In our notation we treat the entire prompt and answer strings as a single variable.

**Reward Modelling Phase**: In the second phase the instruction-tuned model is given prompts $\mathbf{x}$ and produce pairs of answers $(\mathbf{y}_1, \mathbf{y}_2) \sim \pi_{\text{SFT}}(\mathbf{y}|\mathbf{x})$. Users then rank the answers, denoted as $\mathbf{y}_w \succ \mathbf{y}_l \mid \mathbf{x}$ where $\mathbf{y}_w$ and $\mathbf{y}_l$ are the preferred and dispreferred answer respectively. The rankings are usually assumed to be generated by the Bradley-Terry (BT) (Bradley and Terry, 1952), in which the preference distribution $p$ is assumed to be driven by an unobserved latent reward $r(\mathbf{x}, \mathbf{y})$ and the following parameterization:

$$p(\mathbf{y}_1 \succ \mathbf{y}_2 \mid x) = \frac{\exp\left(r(\mathbf{x}, \mathbf{y}_1)\right)}{\exp\left(r(\mathbf{x}, \mathbf{y}_1)\right) + \exp\left(r(\mathbf{x}, \mathbf{y}_2)\right)}. \tag{1}$$

Then given a dataset of user rankings $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{y}_w^{(i)}, \mathbf{y}_l^{(i)}\}_{i=1}^N$, we can train a parameterized reward model $r_\phi(\mathbf{x}, \mathbf{y})$ using maximum likelihood:

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) \sim \mathcal{D}}\big[\log \sigma(r_\phi(\mathbf{x}, \mathbf{y}_w) - r_\phi(\mathbf{x}, \mathbf{y}_l))\big] \tag{2}$$

where $\sigma$ is the logistic function.

**Reinforcement Learning Phase**: During the final phase, we use the learned reward function in an RL loop to where the LLM is treated as a policy. The most common optimization objective is the following:

$$\max_{\pi_\theta} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}, \mathbf{y} \sim \pi_\theta(\mathbf{y}|\mathbf{x})}\big[r_\phi(\mathbf{x}, \mathbf{y})\big] - \beta \mathbb{D}_{\text{KL}}\big[\pi_\theta(\mathbf{y} \mid \mathbf{x}) \mid\mid \pi_{\text{ref}}(\mathbf{y}|\mathbf{x})\big] \tag{3}$$

where $\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})$ is a reference distribution (usually taken to be $\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})$) and $\beta$ is a hyper-parameter. This objective trades-off maximizing the reward $r_\phi(\mathbf{x}, \mathbf{y})$ and a divergence term from a fixed reference distribution. The second term acts as a regularizer to prevent the policy $\pi_\theta$ from drifting too far away from the initialization $\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})$. This objective is then optimized using a general purpose RL algorithm, such as PPO (Schulman et al., 2017).

### 2.2 Direct Preference Optimization

Direct Preference Optimization (Rafailov et al., 2023) starts with the same objective as Eq. 3. However, DPO assumes we have access to the ground truth reward $r(\mathbf{x}, \mathbf{y})$ and derives an analytical transformation between the optimal reward and optimal policy. This can be substituted back into the reward optimization objective in Eq. 2, which allows us to train the optimal model directly on the feedback data using the following objective:

$$\mathcal{L}_{\text{DPO}}\left(\pi_\theta; \pi_{\text{ref}}\right) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta\left(\mathbf{y}_w \mid \mathbf{x}\right)}{\pi_{\text{ref}}\left(\mathbf{y}_w \mid \mathbf{x}\right)} - \beta\log\frac{\pi_\theta\left(\mathbf{y}_l \mid \mathbf{x}\right)}{\pi_{\text{ref}}\left(\mathbf{y}_l \mid \mathbf{x}\right)}\Big)\Big] \quad (4)$$

Here the parameter $\beta$ is the same as in Eq. 3 and similarly controls the trade-off between expected reward and divergence from the model initialization. The DPO objective is attractive as it allows us to recover the optimal model using a standard classification loss, without the need for on-policy sampling or significant amount of hyper-parameter tuning. Eq. 4 resembles the reward modelling objective in Eq. 2 under the parameterization

$$r_\theta(\mathbf{x},\mathbf{y}) = \beta\log\frac{\pi_\theta\left(\mathbf{y} \mid \mathbf{x}\right)}{\pi_{\text{ref}}\left(\mathbf{y} \mid \mathbf{x}\right)} \quad (5)$$

We will refer to this as the DPO "implicit reward". Theorem 1 in (Rafailov et al., 2023) shows that this is indeed a valid parameterization of a reward model without loss of generality. If we substitute this form of $r_\theta(\mathbf{x},\mathbf{y})$ into the RL objective 3 we can obtain the optimal solution in a closed form, which happens to be $\pi_\theta$. We will return to the interpretation of DPO as an implicit reward function later on in our analysis of out-of-distribution bootstrapping.

## 3 Building in Explicit Regularization in DPO

Prior works have explicitly considered length-regularization in the classical RLHF pipeline (Singhal et al., 2023), however these methods do not transfer directly to direct alignment algorithms, such as DPO. We will derive a length-regularized version of the algorithm from first principles, by adding a regularized term in the RL problem in Eq. 3. The below considerations hold for a general regularizer, but we will focus on a length term $\alpha|y|$, where $\alpha$ is a hyper-parameter and $|y|$ denotes the token-length of the answer $y$. We then formulate the regularized RL problems in the following objective:

$$\max_{\pi_\theta}\mathbb{E}_{\mathbf{x}\sim\mathcal{D},\mathbf{y}\sim\pi_\theta(\mathbf{y}|\mathbf{x})}\left[r(\mathbf{x},\mathbf{y})\right] - \alpha|\mathbf{y}| - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta(\mathbf{y} \mid \mathbf{x}) \mid\mid \pi_{\text{ref}}(\mathbf{y}|\mathbf{x})\right] \quad (6)$$

where we assume that $r(\mathbf{x},\mathbf{y})$ is still the same latent reward driving human preferences. We can follow the same derivations in (Rafailov et al., 2023) for the reward function $r(\mathbf{x},\mathbf{y})] - \alpha|\mathbf{y}|$ and obtain the optimal solution to Eq. 6 as

$$\pi^*(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\pi_{\text{ref}}e^{\frac{1}{\beta}(r(\mathbf{x},\mathbf{y})-\alpha|\mathbf{y}|)} \quad (7)$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}}\pi_{\text{ref}}e^{\frac{1}{\beta}(r(\mathbf{x},\mathbf{y})-\alpha|\mathbf{y}|)}$. With some simple algebra, we can then obtain the equivalent regularized reward re-formulation:

$$r(\mathbf{x},\mathbf{y}) = \beta\log\frac{\pi^*(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})} + \beta\log Z(\mathbf{x}) - \alpha|\mathbf{y}| \quad (8)$$

We can then plug in Eq. 8 into the reward modelling stage in Eq. 2, which yields the following regularized DPO objective:

$$\mathcal{L}_{\text{R-DPO}}\left(\pi_\theta; \pi_{\text{ref}}\right) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta\left(\mathbf{y}_w \mid \mathbf{y}\right)}{\pi_{\text{ref}}\left(\mathbf{y}_w \mid \mathbf{x}\right)} - \beta\log\frac{\pi_\theta\left(\mathbf{y}_l \mid \mathbf{x}\right)}{\pi_{\text{ref}}\left(\mathbf{y}_l \mid \mathbf{x}\right)} - \left(\alpha|\mathbf{y}_w| - \alpha|\mathbf{y}_l|\right)\Big)\Big] \quad (9)$$

This is similar to the standard DPO objective, except for the an additional regularization term within $\alpha|y_w| - \alpha|y_l|$ in the logit of the binary classification loss.

Concurrent work (Chen et al., 2024) also consider the length exploitation problem in the classical RLHF pipeline. They suggest a similar regularization in the reward modelling stage in Eq. 2 to disentangle the answer's quality from the length bias and show meaningful improvement in length-controlled model performance. Our derivations can be seen as the DPO implicit reward counterpart to that classical RLHF approach, explicitly linking the regularized reward modelling problem to an equivalent regularized RL setup.

Similar to the original DPO formulation, the regularized objective still aims to increase the likelihood along the preferred answer, while decreasing the likelihood along the dis-preferred answer, modulated by a weighting term. This term is equivalent to the original DPO formulation with the addition of the regularization margin $\alpha|y_w| - \alpha|y_l|$. We can interpret this as an additional per-example learning rate, which up-weighs the gradient on feedback pairs, in which the selected answer is shorter and

| Dataset | Preferred Length | | | Dispreferred Length | | |
|---|---|---|---|---|---|---|
| | Mean | Median | Std. | Mean | Median | Std. |
| Anthropic RLHF HH | **79.6** | **57.0** | **74.0** | 75.7 | 51.0 | 73.3 |
| Reddit TL;DR | **37.9** | **36.0** | **13.9** | 35.2 | 34.0 | 13.4 |

Table 1: Summary statistics across preference datasets. Bold indicates maximum between preferred and dispreferred statistic for a particular dataset. Statistics do not exclude long tails.

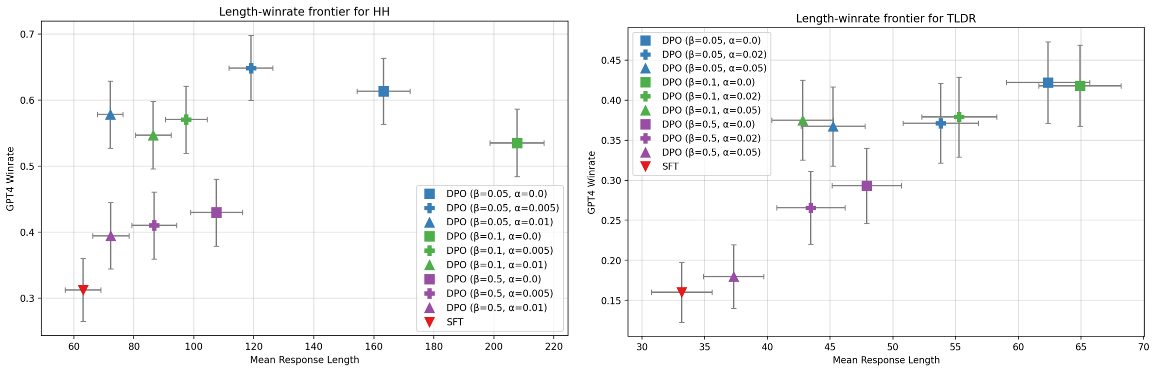


Figure 3: Sampled lengths vs. GPT4 winrates for HH and TLDR test sets. 256 samples evaluated for length and winrates. gpt4-0613 used as judge with prompt similar to (Rafailov et al., 2023), with random position flipping.

down-weights the gradient on pairs in which the selected answer is longer, proportional to the difference in length.

## 4 Experiments

In this section we will empirically investigate the verbosity exploitation issues in DPO, the effectiveness of our regularization strategy and the potential causes of these effects. We beging with a description of our evaluation tasks and models.

### 4.1 Datasets and Models

We utilize three different setups in our experimental setting based on summarization, dialogue and general instruction-following.

**Summarization** We use the standard Reddit TL;DR (TL;DR) summarization dataset from (Stiennon et al., 2022), which consists of a Reddit post and several short summaries, judged for quality and informativeness by human evaluators.

**Dialogue:** For our dialogue experiment we use the Anthropic Helpful and Harmless (HH) datasets (Bai et al., 2022b), which consists of general conversations with a language model assistants, which are also ranked by human annotators.

Datasets statistics are included in Table 1 where exhibit a small length bias in the preferred response. Following (Rafailov et al., 2023) we use the Pythia 2.8B (Biderman et al., 2023) for both the dialogue and summarization tasks and carry out full-parameter fine-tuning, using the DPO original codebase[1] with default hyperparameters, except when noted otherwise. All experiments were carried out on 4 A40 GPUs for a total of about 2000 GPU hours.

### 4.2 Length Exploitation in DPO and Effectiveness of Regularization

We first consider the Anthropic Helpful and Harmless and Reddit TL;DR datasets. For both tasks, we train models with three parameter values $\beta \in [0.5, 0.1, 0.05]$ and then sample 256 answers using prompts fron the evaluation dataset. The length histograms are shown in Fig. 2. The first two columns show the answer length distribution for the set of preferred, rejected and DPO-generated answer, with each row corresponding to a different value of the $\beta$ parameter. We see that the DPO generated answers are, on average, significantly longer than both the preferred and rejected answers. Models trained with smaller values of $\beta$ generate longer responses on average, which is expected as this parameter controls the deviation from the initial policy. Not only does the DPO model generate longer answers, it also generates answers that are significantly out-of-distribution in terms of length

[1]https://github.com/eric-mitchell/direct-preference-optimization

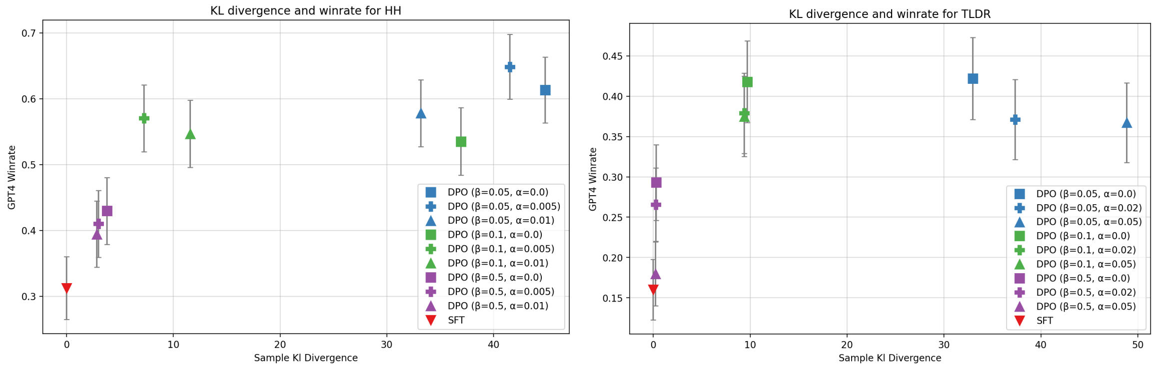


Figure 4: KL divergence vs. sampled lengths for HH and TL;DR, where KL divergence is calculated as the expected reward across the 256 samples generated from the test prompts in both datasets. At most 512 new tokens sampled.

from the offline preference dataset.

The third and fourth column in Fig .2 show results for the SFT, DPO the length-regularized DPO model introduced in Section 3. We use parameters of $\alpha = 0.01$ and $\alpha = 0.05$ for the Anthropic Helpful and Harmless and Reddit TL;DR datasets respectively. While the length-regularized models still show mild increase in average length, they match the SFT model much more closely. Moreover, they do not generate answers with significantly out-of-distribution lengths. This indicates that the proposed algorithm can efficiently regularize the verbosity of the trained model.

### 4.3 Length Versus Quality Trade-Offs

In this section we evaluate the length versus quality model trade-offs. For the Anthropic Helpful and Harmless and Reddit TL;DR datasets we use the answers generated in the previous section and compare them head-to-head against the dataset preferred answer, using GPT 4 as an evaluator. For the UltraFeedback dataset, we evaluate the model on MTBench (Zheng et al., 2023), which also uses GPT 4 to directly provide numerical scores to the model-generated answers. Our main results are shown in Fig. 3, which plots model win rates against average answer length, with 90% confidence intervals. We again evaluate three different values for the beta parameter $\beta \in [0.05, 0.1.0.5$ and three values of $\alpha$ with $\alpha \in [0, 0.005, 0.01]$ for HH and $\alpha \in [0, 0.2, 0.5]$ for TL;DR respectively ($\alpha = 0$ is the standard DPO algorithm). Similar to before, we see that the length-regularized training can efficiently control verbosity, significantly decreasing the average length of the answers as compared to the standard DPO training. Moreover, on the HH task, regularization also leads to mild improvement in win rates, but a slight decrease on TL;DR although both of these are not statistically significant. These results are quite promising, as GPT4 is known to have a significant length bias in its preferences (Wang et al., 2023; Singhal et al., 2023). On both the HH and TL;DR, the length-regularized experiments with $\beta = 0.05$ and $beta = 0.01$ match the average lengths of the corresponding $\beta = 0.5$ runs, but achieve statistically significant higher corresponding win rates with close to 20% improvement on HH and 15% improvement on TL;DR.

### 4.4 Is Length a Proxy for KL-Divergence?

In the constrained RL problem in Eq. 3 and the corresponding DPO objective in Eq. 4, the $\beta$ parameter controls the degree of policy divergence from the initial reference model. In Fig. 2 and Fig. 3, we see that average length of the model generated answers is inversely proportional to the $\beta$ parameter. In this section, we investigate the relationship between the length-regularized DPO objective in Eq. 9 and the KL divergence from the initial policy. In Fig .4, we plot the trained policy KL divergence from the initialization $\pi_{\text{ref}}$ for the different values of $\beta$ and $\alpha$ parameters. We see only a weak correlation between KL divergence and length. For both HH and TL;DR, length-regularized models trained with $\beta = 0.05$ and $\beta = 0.01$ match the average length of train runs with $\beta = 0.5$ (Fig. 3). At the same time, these runs have statistically significant higher KL divergences and win rates as shown in Fig. 3. We hypothesize that this indicates the existence of different factors driving human preference, with length being only a partial one.

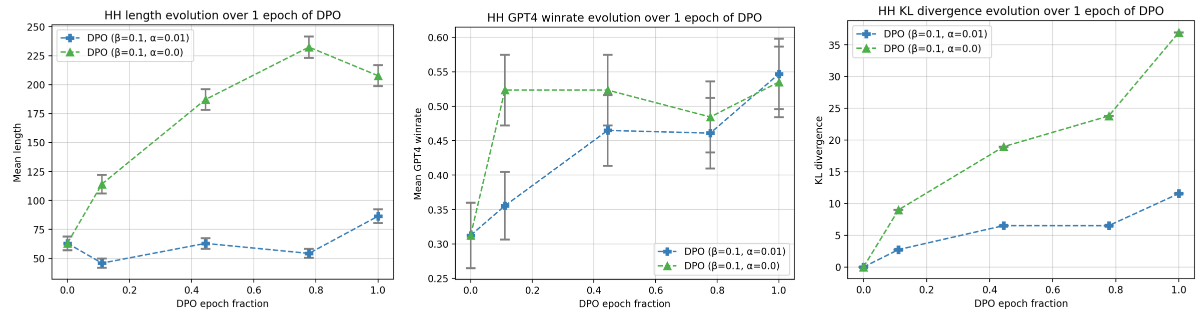


Figure 5: Evolution of HH sample length, GPT4 winrates, and KL divergence along equally-spaced intervals within one epoch (170K steps) of DPO training. Error bars indicate 90% confidence intervals.

### 4.5 DPO and Early Convergence

In (Rafailov et al., 2023), the authors show early convergence of the DPO algorithm on the HH dataset. DPO achieves its best performance within a few hundred gradient steps, and does not improve with further training. Similar observations have also been made within the open-source community. We claim that this effect is likely due to length-exploitation and the biased GPT4 evaluator. In Fig. 6, we consider the training progression on the HH dataset with $\beta = 0.1$. We compare the regular DPO run ($\alpha = 0$) with the length-regularized one $\alpha = 0.1$. We train for a single epoch and evaluate intermediate checkpoints on the same set of prompts for average answer length, win rates, and KL divergence. We see that already within the first 10% of the epoch, the standard DPO run produces answers almost twice as long as the SFT model. Unregularized DPO achieves its highest win rate here, with only KL divergence and average length increasing steadily with further training. In contrast, the length-regularized run sees little to no intermediate increase in length, but steady improvement in win rates throughout training and slow increases in divergence from the reference policy. We hypothesize that the regular DPO training quickly increases length, which exploits the evaluator's bias, but does not capture the more complex features of preferences. On the other hand, the length-regularized training run is able to disentangle the verbosity component and fit other, more difficult quality features over a longer training period.

### 4.6 What Drives Length Exploitation?

Excessive model verbosity (John Schulman et al., 2022) has been well understood under classical RLHF as a reward exploitation problem (Gao et al., 2023; Casper et al., 2023; Lambert and Calandra, 2023) driven by a bias in the feedback datasets for longer answers. In particular, in the classical RLHF pipeline as outlined in Section 2.1, the reward model is continuously queried on new data generated by the model, which can create an out-of-distribution robustness issue. These results do not directly transfer to the DPO algorithm, as it does not train a separate reward model and only uses the offline feedback dataset for training. Surprisingly we find that the exploding length issue in DPO training is similarly driven by out-of-distribution exploitation. We consider the DPO algorithm as an implicit reward training method, as outlined in Section 2.2. We investigate the behaviour of the implicit reward $r_\theta$ as defined in Eq. 5. Since the DPO policy $\pi_\theta$ is the optimal solution to the constrained RL problem in Eq. 3 corresponding to $r_\theta$, any exploitation behaviour from the policy must be driven by the reward function. We evaluate $r_\theta$ trained with $\beta = 0.1$ and different $\alpha$ parameters on the offline feedback dataset (within its training distribution) and on answers generated by the corresponding DPO policy (out of distribution). Surprisingly, within distribution, the corresponding implicit reward models exhibit weak to no length correlation (and even negative length correlation with strong $\alpha$ regularization). However, they all show significant length bias on out-of-distribution samples, with length explaining 0.3-0.46 of the reward variance.

## 5 Related Work

**Reward Exploitation in RLHF:** RLHF reward exploitation, also known as reward over-optimization, is a well-known issue (Skalse et al., 2022; Pan et al., 2022; Casper et al., 2023; Lambert and Calandra, 2023) in which during the reinforcement learning

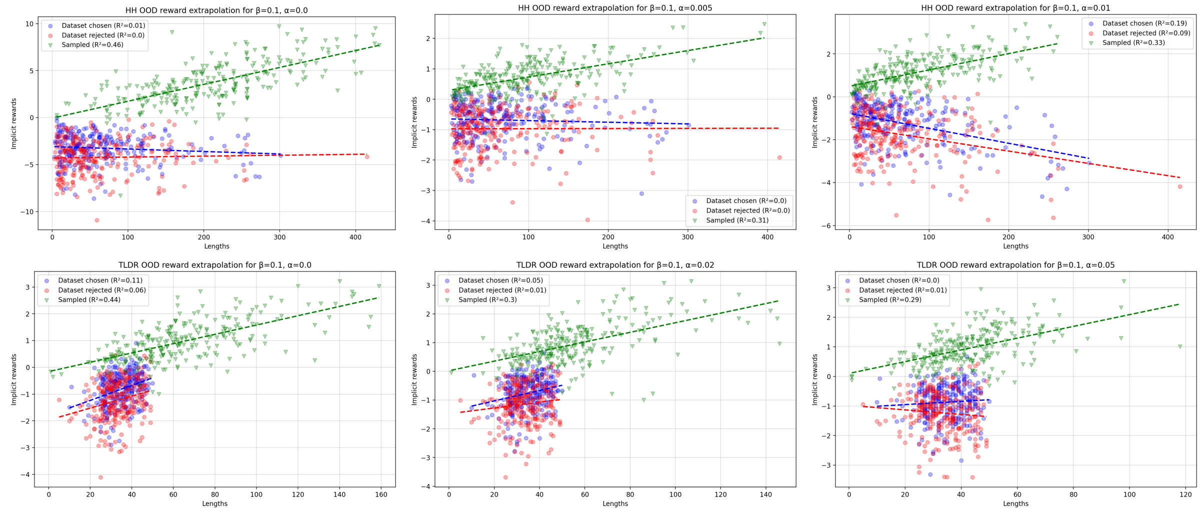


Figure 6: Evaluation of the DPO implicit reward model as defined in Section 2.2 on in-distribution preffered (blue) and rejected (red) answers, as well as OOD answers generated from the corresponding policy. The reward model exhibits little to no length bias in distribution, but significant length correlation outside its training distribution.

stage, the expected reward keeps improving, but the quality of the model begins to degrade after some point. These effects were confirmed analytically in controlled experiments (Gao et al., 2023), as well as empirically in user studies (Dubois et al., 2024). Increased model verbosity has been explicitly linked to this phenomenon (John Schulman et al., 2022). A number approaches have been proposed to mitigate this issue, such as penalizing epistemic uncertainty (Coste et al., 2023; Zhai et al., 2023) or using mixture reward models (Moskovitz et al., 2023), but they do not explicitly target the length issue.

**Mitigating Length Biases in RLHF:** A number of works have sought to explicitly address length biases in RLHF policies. (Ramamurthy et al., 2023) suggest setting a simple discount factor, which improves naturalness of the generated language, (Singhal et al., 2023) carry out an extensive study of length correlations in classical RLHF and suggest a number of mitigating approaches. The closes to our approach are the works of (Shen et al., 2023) and the concurrent work of (Chen et al., 2024), which propose to disentangle length-biases from quality during the reward modelling stage. Our work can be seen as a DPO equivalent counter-part to these approaches.

As far as we are aware, this is the first work to study the length exploitation problem for direct alignment algorithms, such as DPO.

## 6 Limitations

Our work addresses the particular issue of length exploitation in Direct Preference Optimization. Our regularization objective requires explicit penalty function (such as length) and may not be suitable to avoid general exploitation issues along axes separate from verbosity. Furthermore, we only study the DPO objective, which might behave differently from other direct alignment algorithms, which use different objective functions.

## 7 Conclusion

In this work, we study the problem of length exploitation in the Direct Preference Optimization algorithm for the first time. On two standard human feedback datasets, we empirically show that DPO exhibits significant length hacking across a range of hyperparameters. We then link this phenomenon to out-of-distribution bootstrapping. We derive an analytical length-regularized version of the DPO algorithm and show empirically that we can maintain model performance, as evaluated by GPT4 without significant increases in verbosity, boosting length-corrected win rates by up to 15-20%. Given the strong length bias in public feedback datasets and the prominence of DPO in the open source community, we hypothesize that a lot of open source models suffer from similar length-exploitation issues, driving the observations of Fig. 1. Our results are encouraging, suggesting that open-source models could match proprietary ones on automated evaluations on a length corrected basis as well.

## Ethics Statement

Scientific work published at ACL 2023 must comply with the ACL Ethics Policy.[2] We encourage all authors to include an explicit ethics statement on the broader impact of the work, or other ethical considerations after the conclusion but before the references. The ethics statement will not count toward the page limit (8 pages for long, 4 pages for short papers).

## Acknowledgements

This document has been adapted by Jordan Boyd-Graber, Naoaki Okazaki, Anna Rogers from the style files used for earlier ACL, EMNLP and NAACL proceedings, including those for EACL 2023 by Isabelle Augenstein and Andreas Vlachos, EMNLP 2022 by Yue Zhang, Ryan Cotterell and Lea Frermann, ACL 2020 by Steven Bethard, Ryan Cotterell and Rui Yan, ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

## References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Di-

[2] https://www.aclweb.org/portal/content/acl-code-ethics

ana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan

Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models.

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. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and

Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël 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 Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback.

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Odin: Disentangled reward mitigates hacking in rlhf.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jack Clark and Dario Amodei. 2016. Faulty reward functions in the wild.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. 2023. Reward model ensembles help mitigate overoptimization.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Barret Zoph John Schulman, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Michael Pokorny Luke Metz, Rapha Gontijo Lopes, Shengjia Zhao, Arun Vijayvergiya, Eric Sigler, Adam Perelman, Chelsea Voss, Mike Heaton, Joel Parish, Dave Cummings, Rajeev Nayak, Valerie Balcom, David Schnurr, Tomer Kaftan, Chris Hallacy, Nicholas Turley, Noah Deutsch, Vik Goel, Jonathan Ward, Aris Konstantinidis, Wojciech Zaremba, Long Ouyang, Leonard Bogdonoff, Joshua Gross, David Medina, Sarah Yoo, Teddy Lee, Ryan Lowe, Dan Mossing, Joost Huizinga, Roger Jiang, Carroll Wainwright amd Diogo Almeida, Steph Lin, Marvin Zhang, Kai Xiao, Katarina Slama, Steven Bills, Alex Gray, Jan Leike, Jakub Pachocki, Phil Tillet, Shantanu Jain, Greg Brockman, Nick Ryder, Alex Paino, Qiming Yuan, Clemens Winter, Ben Wang, Mo Bavarian, Igor Babuschkin, Szymon Sidor, Ingmar Kanitscheider, Mikhail Pavlov, Matthias Plappert, Nik Tezak, Heewoo Jun, William Zhuk, Vitchyr Pong, Lukasz Kaiser, Jerry Tworek, Andrew Carr, Lilian Weng, Sandhini Agarwal, Karl Cobbe, Vineet Kosaraju, Alethea Power, Stanislas Polu, Jesse Han, Raul Puri, Shawn Jain, Benjamin Chess, Christian Gibson, Oleg Boiko, Emy Parparita, Amin Tootoonchian, Kyle Kosic, and Christopher Hesse. 2022. Introducing chatgpt.

Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2023. Who answers it better? an in-depth analysis of chatgpt and stack overflow answers to software engineering questions.

Nathan Lambert and Roberto Calandra. 2023. The alignment ceiling: Objective mismatch in reinforcement learning from human feedback.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

Peter Liu. 2024. [link].

David Manheim and Scott Garrabrant. 2019. Categorizing variants of goodhart's law.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. *International Conference on Machine Learning*.

Ted Moskovitz, Aaditya K. Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca D. Dragan, and Stephen McAleer. 2023. Confronting reward model overoptimization with constrained rlhf.

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. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. *International Conference on Learning Representations*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward hacking.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *International Conference on Learning Representations*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Yuanzhao Zhai, Han Zhang, Yu Lei, Yue Yu, Kele Xu, Dawei Feng, Bo Ding, and Huaimin Wang. 2023. Uncertainty-penalized reinforcement learning from human feedback with diverse reward lora ensembles.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Conference on Neural Information Processing Systems Track on Datasets and Benchmarks*.

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