MallowsPO: Fine-Tune Your LLM with Preference Dispersions

Anonymous authors
Paper under double-blind review

Abstract

Direct Preference Optimization (DPO) has recently emerged as a popular approach to improve reinforcement learning with human feedback (RLHF), leading to better techniques to fine-tune large language models (LLM). A weakness of DPO, however, lies in its lack of capability to characterize the diversity of human preferences. Inspired by Mallows' theory of preference ranking, we develop in this paper a new approach, the *MallowsPO*. A distinct feature of this approach is a *dispersion index*, which reflects the dispersion of human preference to prompts. We show that existing DPO models can be reduced to special cases of this dispersion index, thus unified with MallowsPO. More importantly, we demonstrate empirically how to use this dispersion index to enhance the performance of DPO in a broad array of benchmark tasks, from synthetic bandit selection to controllable generation and dialogues, while maintaining great generalization capabilities. MallowsPO is also compatible with other SOTA offline preference optimization methods, boosting nearly 2% extra LC win rate when used as a plugin for fine-tuning Llama3-Instruct.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Stiennon et al., 2020; Ziegler et al., 2019) has made significant contributions to the success of Large Language Models (LLMs) such as `ChatGPT` and `GPT4` (Achiam et al., 2023). Recently, Direct Preference Optimization (DPO) (Rafailov et al., 2023) proposes to bypass RL, thus leading to faster training and better resource efficiency. More importantly, DPO achieves comparable or superior performance against RLHF in downstream tasks such as fine-tuning LLMs in Llama3 (Dubey et al., 2024), Zephyr (Tunstall et al., 2023), Neural Chat, BTLM-DPO (Ivison et al., 2023), etc. DPO's success has attracted much research attention, leading to variants beyond pairwise ranking such as KTO (Ethayarajh et al., 2023; Song et al., 2024), unified perspectives on loss parameterization such as IPO (Azar et al., 2024), GPO (Tang et al., 2024), and reference-free alternatives such as CPO (Xu et al., 2024), ORPO (Hong et al., 2024), SimPO (Meng et al., 2024), etc.

Notwithstanding the successes by RLHF and DPO, both are limited by the assumption that preference follows the Bradley-Terry (BT) model (Bradley & Terry, 1952). Particularly, they do not account for varying degrees of agreement in responses to different prompts. For instance, people are more likely to agree on "$1+1=?$ // 2." as opposed to "What is the best city to live in the U.S.? // New York." In language models, this concerns the *dispersion* of next-token prediction, similar to *personalization* in recommendation systems (Chan et al., 2022; Fu et al., 2022). See Figure 1 for more examples.

The purpose of this paper is to formalize the idea of prompt dispersion in the design of DPO. We adapt Mallows' preference ranking theory (Diaconis, 1988; Mallows, 1957), a family of ranking models that provide a natural carrier for prompt dispersion, and propose the following decomposition/factorization of the (latent) reward function: $\text{reward}(\text{prompt}, \text{completion}) = \text{dispersion}(\text{prompt}) \times \text{scaled reward}(\text{completion} \mid \text{prompt})$, where "prompt" and "completion" correspond, respectively, to question and answer. This decomposition allows to specify the diverse level of prompt dispersions hidden in the DPO, which is translated into a prompt-dependent factor – the *dispersion index* in the preference likelihood. The scaled reward is given by the relative rank of the (possible) completions, which further enhances the model interpretability. We then leverage the change of variables technique to propose two models, MallowsPO-θ and MallowsPO-ϕ, by two choices of the discrepancy function in the Mallows model which we will elaborate in Section 3.1.

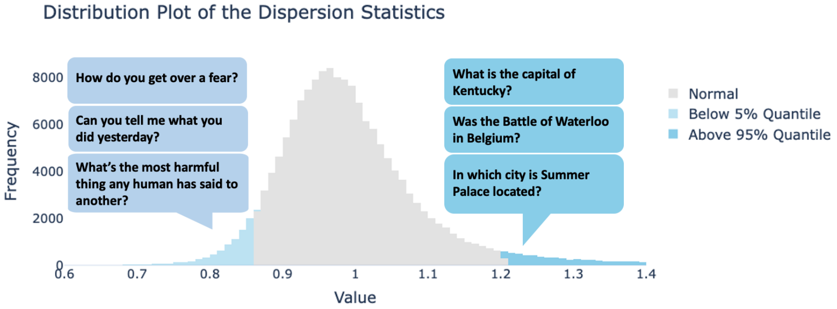


Figure 1: Prompts with low/high neg-log dispersion estimate values from Anthropic HH dataset.

The main contributions of this paper are four-fold.

(1) We formalize the idea of prompt dispersion in DPO, and develop the MallowsPO approach to implement this idea, so as to improve and generalize DPO. As far as we are concerned, this is the first work in preference optimization that considers a mathematically well-grounded preference ranking model, i.e., Mallows, beyond BT setting.

(2) We propose approximations to the dispersion index, a critical component of the Mallows model, so as to facilitate computation. We present a synthetic example to illustrate that our proposed estimate is aligned with the true dispersion.

(3) The Mallows model serves as a foundation for developing versatile preference optimization objectives, including MallowsPO-θ (a generalization of DPO) and MallowsPO-ϕ. We also provide various analytical results for MallowsPOs, uncovering various new insights on existing DPO models, and a generalized ΨPO model that unifies all DPO models (including MallowsPO).

(4) We conduct extensive experiments, from synthetic bandits, controllable generation, fine-tuning Pythia 2.8B on off-policy Anthropic HH dataset, to fine-tuning Llama3-8B-Instruct on a hybrid UltraFeedback dataset. Notably, we perform an exclusive hyperparameter search for a fair comparison, and repeat for different random seeds to justify the significance of the improvement. The results show clear advantages of MallowsPO over (BT-)DPO, which showcases the potential of this novel direction of considering preference/prompt dispersion.

Related Works. Existing work on personalization in dialogue generation such as Fu et al. (2022) and Li et al. (2016) has also paid attention to the diversity of human preferences ("there are a thousand Hamlets in a thousand people's eyes"); Munos et al. (2023) proposes a Nash game model to incorporate the diversity. There are also other DPO variants: f-DPO (Wang et al., 2023) considers general f-divergence in DPO; ODPO (Amini et al., 2024) adds a margin to account for the preference significance. Recent works propose to learn online preferences (Calandriello et al., 2024; Tajwar et al., 2024), or learn from AI feedbacks (Bai et al., 2022b; Chen et al., 2024; Lee et al., 2023). For classical RLHF, studies to improve the design and capabilities of RLHF include Dubois et al. (2024); Kirk et al. (2023); Wang et al. (2024a); Zhai et al. (2023); Zhao et al. (2023); Zheng et al. (2023), whose ideas can also benefit DPO. See Winata et al. (2024) for a survey on learning from preferences.

The remainder of the paper is organized as follows. Background materials on RLHF and DPO are highlighted in Section 2. Section 3 focuses on the development of MallowsPO, followed by more analytical results and various perspectives in Section 4. Experimental results are detailed in Section 5, and concluding remarks in Section 6.

2 Preliminaries

Both RLHF and DPO start with fine-tuning a pre-trained LLM by supervised learning on high-quality data for some downstream tasks of interest, to acquire a model π^{SFT}. This step is referred to as the supervised fine-tuning (SFT) phase. For instance, for training InstructGPT (Ouyang et al., 2022), GPT-3 (Brown et al., 2020) is first fine-tuned on the given input prompt distribution.

♢ **RLHF** (Ouyang et al., 2022; Stiennon et al., 2020; Ziegler et al., 2019). On top of π^{SFT}, RLHF is proposed to serve as the next step to conduct further fine-tuning to generate high-quality outputs as judged by humans. Given a generative model π, it is prompted with prompts x to produce pairs of answers (or, “completions”), $\{y_1, y_2\} \sim \pi(y \mid x)$, which are then presented to human labelers who express preferences for one completion over the other. Denote by $y_w \succ y_l \mid x$, meaning that $y_w \in \{y_1, y_2\}$ is preferred over $y_l \in \{y_1, y_2\}$. The preferences are assumed to be generated by some latent reward model $r^*(x, y)$. Based on the collected preference data $\{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$, RLHF consists of first learning a reward model $r(x, y)$, followed by learning a policy $\pi_r(y \mid x)$ in which the prompt x is the state, and the completion y is the action.

(a) **Reward model**. To capture the underlying human preferences, RLHF assumes the Bradley-Terry model (Bradley & Terry, 1952) that stipulates the pairwise preference distribution:

$$p^*\left(y_1 \succ y_2 \mid x\right) := \sigma\left(r^*\left(x, y_1\right) - r^*\left(x, y_2\right)\right), \tag{1}$$

where $\sigma(s) := \frac{1}{1+e^{-s}}$. Given access to a static dataset of comparisons $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1,\dots,N}$, RLHF seeks to approximate the latent reward $r^*(x, y)$ by a family of functions $\{r_\psi(x, y)\}_\psi$, and estimate the parameters by minimizing the (negative) log-likelihood loss: $\min_\psi \mathcal{L}\left(r_\psi, \mathcal{D}\right) := -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma\left(r_\psi\left(x, y_w\right) - r_\psi\left(x, y_l\right)\right)\right]$. Denote by $r_{\psi_*}(x, y)$ the solution to the problem.

(b) **RL**. The learned reward function $r_{\psi_*}(x, y)$ is then used to provide feedback to the language model. More precisely, the following KL-regularized RL problem is considered:

$$\max_\pi \mathbb{E}_{x \sim \mathcal{D}}\left[\mathbb{E}_{y \sim \pi(y|x)}\left[r_{\psi_*}(x, y)\right] - \beta \mathrm{KL}\left(\pi(\cdot \mid x) \| \pi_{\text{ref}}(\cdot \mid x)\right)\right], \tag{2}$$

where $\beta > 0$ is a hyperparameter controlling the deviation from the reference policy $\pi_{\text{ref}} = \pi^{\text{SFT}}$. In view of (2), RLHF uses the reward function $r(x, y) = r_\psi(x, y) - \beta\left(\log \pi(y \mid x) - \log \pi_{\text{ref}}\left(y \mid x\right)\right)$, and solves the RL problem by proximal policy optimization (PPO) (Schulman et al., 2017).

♢ **DPO** (Rafailov et al., 2023). One disadvantage of RLHF is that the RL step often requires substantial computational effort (e.g., to carry out PPO). The idea of DPO is to combine the two steps (a)–(b) in RLHF into a single one, bypassing the computation in the RL step.

The key idea is that given a reward function $r(x, y)$, the problem in (2) has a closed-form solution: $\pi_r(y \mid x) = \frac{1}{Z(x)} \pi_{\text{ref}}\left(y \mid x\right) \exp\left(r(x, y)/\beta\right)$, where $Z(x) = \sum_y \pi_{\text{ref}}\left(y \mid x\right) \exp\left(r(x, y)/\beta\right)$. Rewrite the above as: $r(x, y) = \beta \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)$. Through this change of variables, the latent reward $r^*(x, y)$ can be expressed in terms of the optimal policy $\pi^*(y \mid x)$, the reference policy $\pi_{\text{ref}}\left(y \mid x\right)$ and a constant $Z^*(x)$. Substituting this r^* expression into (1) yields:

$$p^*\left(y_1 \succ y_2 \mid x\right) = \sigma\left(\beta \log \frac{\pi^*\left(y_1 \mid x\right)}{\pi_{\text{ref}}\left(y_1 \mid x\right)} - \beta \log \frac{\pi^*\left(y_2 \mid x\right)}{\pi_{\text{ref}}\left(y_2 \mid x\right)}\right), \tag{3}$$

where $Z^*(x)$ cancels out. The expression in (3) motivates the DPO objective:

$$\min_\pi \mathcal{L}_{\text{DPO}}\left(\pi; \pi_{\text{ref}}\right) := -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi\left(y_w \mid x\right)}{\pi_{\text{ref}}\left(y_w \mid x\right)} - \beta \log \frac{\pi\left(y_l \mid x\right)}{\pi_{\text{ref}}\left(y_l \mid x\right)}\right)\right]. \tag{4}$$

3 DPO BASED ON MALLOWS RANKING MODELS

3.1 MALLOWS RANKING MODELS

The Mallows model is built upon the analysis of rankings, instead of scores or ratings that play the central role in BT models. Concretely, for a positive integer n that represents e.g., n possible items, let $\mathfrak{S}_n$ be the set of permutations of $[n] = \{1, \dots, n\}$ and the space of rankings. We consider the following probability of observing a ranking μ (which represents the preference of n items, e.g., the top ranked item is preferred over the others):

$$\mathbb{P}_{\phi, \mu_0, d}(\mu) := \phi^{d(\mu, \mu_0)}/Z(\phi, d) \quad \text{for } \mu \in \mathfrak{S}_n, \tag{5}$$

where $\phi \in (0, 1]$ is the dispersion parameter, μ_0 is the central ranking, (also known as the location parameter), $d(\cdot, \cdot)$ is a discrepancy function that is right invariant: $d(\mu_1, \mu_2) = d\left(\mu_1 \circ \mu_2^{-1}, id\right)$ for $\mu_1, \mu_2 \in \mathfrak{S}_n$, and $Z(\phi, d) := \sum_{\mu \in \mathfrak{S}_n} \phi^{d(\mu, \mu_0)}$ is the normalizing constant.

Notice that the dispersion indeed reflects how dispersed the probability distribution (5) on the space of rankings is: When $\phi \to 0$, it is concentrated on μ_0, and when $\phi = 1$, it is uniformly distributed. In an attempt to study ranking models (over n items) with pairwise preferences, Mallows (1957) further considered two specific cases of the discrepancy function in (5):

- Mallows-θ model: $d(\mu_1, \mu_2) = \sum_{i=1}^{n} (\mu_1(i) - \mu_2(i))^2$ is the Spearman's rho,
- Mallows-ϕ model: $d(\mu_1, \mu_2) = \operatorname{inv}\left(\mu_1 \circ \mu_2^{-1}\right)$ is the Kendall's tau,

where $\operatorname{inv}(\mu) := \#\left\{(i,j) \in [n]^2 : i < j \text{ and } \mu(i) > \mu(j)\right\}$ is the number of inversions of μ.

The general form in (5) was suggested by Diaconis (1988) along with other discrepancy functions (e.g., Cayley, Ulam distances, etc.) See Critchlow (1985); Diaconis (1988; 1989) for the related group representation approach to ranked, or partially ranked data. Note that the Mallows models can be extended to infinite ranking models with $n = \infty$ (see Meila & Bao (2010); Pitman & Tang (2019); Tang (2019).) In the context of language models, this conforms to a possibly infinite number of completions given a prompt, and allows interpreting unseen completions .

3.2 MallowsPO

We adapt Mallows ranking models highlighted above to the setting of language models. First, denote by $\mu(\cdot \mid x)$ a ranking of completions given the prompt x, such that the preference distribution is:

$$p^*\left(y_1 \succ y_2 \mid x\right) = \mathbb{P}\left(\mu(y_1 \mid x) < \mu\left(y_2 \mid x\right)\right). \tag{6}$$

Next, for the preference probability in (5), given an input prompt x, we assume it induces a conditional central ranking $\mu_0(\cdot \mid x)$, and a dispersion index $\phi(x) \in (0, 1]$. As pointed out in Tang (2019), finding $\mu_0(\cdot \mid x)$ is computationally hard. Fortunately, we discover that in RLHF, this part can be "cleverly" circumvented. By representing $r^*(x, y)$ as the (negative) rank $-\mu_0(y \mid x)$, our goal now becomes:

$$\max_{\pi} \mathbb{E}_{x \sim \mathcal{D}} \left[\mathbb{E}_{y \sim \pi_\theta(y|x)} \left[-\mu_0(y \mid x)\right] - \beta \mathrm{KL}\left(\pi(\cdot \mid x) \| \pi_{\mathrm{ref}}(\cdot \mid x)\right)\right],$$

Note that a *smaller* rank is preferred as per (6). Hence, this provides a natural candidate for the scaled reward that enhances model interpretation. This perspective leads to the discovery of a novel family of preference optimization objectives, each of which corresponds to an instance of Mallows models.

Theorem 1 (MallowsPO) *Given a prompt x, suppose the completions y follow a Mallows preference distribution p^* as in (6), with chosen hyperparameters $d(\cdot, \cdot)$ and $\phi(x)$. Then for any underlying central rank $\mu_0(\cdot \mid x)$, the optimal RLHF policy $\pi_{\mu_0}(\cdot \mid x)$ satisfies $p^*\left(y_1 \succ y_2 \mid x\right) = g_{d,\phi(x)}(\beta \log(\frac{\pi_{\mu_0}(y_1|x)}{\pi_{\mathrm{ref}}(y_1|x)}) - \beta \log(\frac{\pi_{\mu_0}(y_2|x)}{\pi_{\mathrm{ref}}(y_2|x)}))$. Therefore, the policy optimization objective is:*

$$\min_{\pi} -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log\left(g_{d,\phi(x)}\left(\beta \log \frac{\pi_{\mu_0}(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta \log \frac{\pi_{\mu_0}(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)}\right)\right)\right]. \tag{7}$$

The proof is given in Appendix B. Theorem 1 implies that for any Mallows model characterized by $d(\cdot, \cdot)$ and $\phi(x)$, we can construct a corresponding MallowsPO objective. Building on this result, we introduce two instances of MallowsPO, derived from the two most widely used Mallows models.

Corollary 2 (MallowsPO-θ) *Suppose the preference distribution is Mallows-θ, i.e., $d(\mu_1, \mu_2) = \sum_{i=1}^{n} (\mu_1(i) - \mu_2(i))^2$. Given $\phi(x)$, we have*

$$p^*\left(y_1 \succ y_2 \mid x\right) = 1/(1 + \exp\left(-(-2\log\phi(x)) \cdot \left(\mu_0\left(y_2 \mid x\right) - \mu_0\left(y_1 \mid x\right)\right)\right)) := g_{\theta,\phi(x)}(s), \tag{8}$$

where $\log\phi(x) \in (-\infty, 0)$ and $s := \mu_0\left(y_2 \mid x\right) - \mu_0\left(y_1 \mid x\right)$. By Theorem 1, the MallowsPO-θ objective is in the form of (7) with $g_{d,\phi(x)}(\cdot) = g_{\theta,\phi(x)}(\cdot)$ in (8).

(8) shows that in the Mallows-θ ranking, the representation of $p^*(\cdot \mid x)$ resembles a sigmoid function but differs in that it is scaled by the term $-2\log\phi(x)$, which reflects the distribution's dispersion. Compared to BT, where the distribution is exactly in form of sigmoid, Mallows-θ allows for greater flexibility in controlling the spread of the distribution function. This is particularly important in language modeling, as the concept of dispersion provides insight into how diverse people's preferences are for different completions/responses. As $\phi(x) \to 0$, $p^*(\cdot \mid x)$ is getting closer to a step function (i.e., Dirac delta) (as shown in Fig. 2), corresponding to cases where the prompt x has a clear, standard answer. Conversely, as $\phi(x) \to 1$, $p^*(\cdot \mid x)$ approaches a constant value of 0.5 (i.e. uniform), indicating that any answer to the prompt x is equally reasonable.

In terms of the objective, by setting $-2\log\phi(x) = 1$, we recover the DPO in (4). When comparing the objective of the DPO with that of MallowsPO-θ, the key difference is the presence of an extra term $-\log\phi(x)$, which reflects the dispersion of the prompt x. Thus, MallowsPO-θ can be viewed as a generalized version of DPO that incorporates prompt dispersion. To see the effect of this additional term: When dispersion is high ($\phi(x) \approx 1$), the term $-2\log\phi(x)$ approaches 0, reducing the weight on preference pairs; when dispersion $\phi(x)$ decreases, $-2\log\phi(x)$ increases, assigning more weight to preference pairs.

Corollary 3 (MallowsPO-ϕ) *Suppose the preference distribution is Mallows-ϕ, i.e., $d(\mu_1, \mu_2) = \operatorname{inv}\left(\mu_1 \circ \mu_2^{-1}\right)$. Given $\phi(x)$, we have*

$$p^*\left(y_1 \succ y_2 \mid x\right) = g_{\phi,\phi(x)}(\mu_0(y_2 \mid x) - \mu_0(y_1 \mid x)), \tag{9}$$

where $\log\phi(x) \in (-\infty, 0)$, and $g_{\phi,\phi(x)}(s) := \frac{1-\operatorname{sgn}(s)}{2} + \operatorname{sgn}(s)\left(\frac{|s|+1}{1-\phi(x)^{|s|+1}} - \frac{|s|}{1-\phi(x)^{|s|}}\right)$. By Theorem 1, the MallowsPO-ϕ objective is in the form of (7) with $g_{d,\phi(x)}(\cdot) = g_{\phi,\phi(x)}(\cdot)$ in (9).

By specifying the underlying ranking to Mallows-ϕ, we get a different link function $g_{\phi,\phi(x)}$, which also contains the dispersion index $\phi(x)$, resulting in a new preference optimization objective.

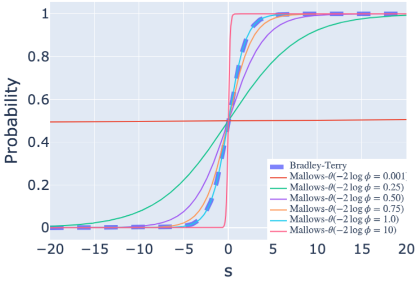


Figure 2: Distribution plot.

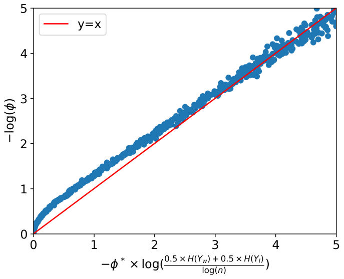


Figure 3: Our proposed estimate matches the true (neg log) dispersion under a Mallows model.

3.3 How to choose the dispersion index $\phi(x)$?

As the dispersion index $\phi(x) \in (0, 1]$ is unknown, computation or estimation of it requires learning via neural networks or other algorithms (Meila & Bao, 2010). Here, however, we propose a more direct approach to estimate $\phi(x)$ without any pretraining or learning. The idea is to qualitatively relate $\phi(x)$ to the empirical output distribution of the pretrained model, on which we propose an 'easy-to-compute' proxy to the negative log dispersion $-\log(\phi(x))$ for each prompt x.

Suppose the preference follows the Mallows-ϕ model. There are two extreme cases: When $-\log(\phi(x)) \to \infty$, we have: $p^*\left(y_1 \succ y_2 \mid x\right) = \begin{cases} 1, & \text{if } \mu_0(y_1 \mid x) < \mu_0(y_2 \mid x), \\ 0, & \text{if } \mu_0(y_1 \mid x) > \mu_0(y_2 \mid x). \end{cases}$ Thus, the probability distribution of the next token will concentrate on a point mass. When $-\log(\phi(x)) \to 0$, we have: $p^*\left(y_1 \succ y_2 \mid x\right) = \frac{1}{2}$, so the next token will be uniformly distributed.

The above observation motivates us to use Shannon's entropy $H(\cdot)$. Note that $H(X) = 0$ when X is a point mass, and $H(X) = \log n$ when X is uniform on n points. Thus, for a given constant $\phi^* > 0$, we propose:

$$-\phi^* \log\left(H(\pi(\cdot \mid x))/\log n\right), \tag{10}$$

as a proxy to $-\log\phi(x)$, where $\pi(\cdot \mid x)$ can be either the pretrained LM model π^{PRE} or the SFT model π^{SFT}. Furthermore, we approximate the entropy term in (10) via a realization of a sequence of $N = \max(|Y^w|, |Y^l|)$ tokens $\{Y_i^w, Y_i^l\}_{i=1,\dots,N}$ given the prompt X:

$$H(\pi(\cdot \mid X)) \approx \frac{1}{2}\sum_{i=1}^{N-1}\left[H(Y_{i+1} \mid Y_i = Y_i^w) + H(Y_{i+1} \mid Y_i = Y_i^l)\right], \tag{11}$$

which can be *easily* computed by the logits of the model given the output data. This is also related to the predictive entropy (Hernández-Lobato et al., 2014; MacKay, 1992) of the next-token predictions.

Accuracy of the estimate. To validate our proposed estimate (10), we consider the similar 'bandit' setup in Tang (2019). We draw rankings μ from a Mallows-ϕ ranking model, and then obtain a pair of winning/losing actions by choosing the highest/lowest ranked elements in the ranking μ. We plot $-\phi^* \cdot \log\left(\frac{H(Y^w)+H(Y^l)}{2\log n}\right)$, given the preferences data (x-axis) and the true dispersion (y-axis) that these data are generated from. Figure 3 shows that our proposed estimator indeed matches the true dispersion, which heuristically reflects the accuracy of our estimate.

3.4 Unify Mallows-θ and Mallows-ϕ for computation

Note that the link function $g_{\phi,\phi(x)}$ in MallowsPO-ϕ is not continuous (or smooth) at $x = 0$, with

$$g'_{\phi,\phi(x)}(s) = \begin{cases} \frac{1}{1-\phi(x)^{s+1}} + \frac{(s+1)\phi^{s+1}\log\phi(x)}{(1-\phi(x)^{s+1})^2} - \frac{1}{1-\phi(x)^s} - \frac{s\phi(x)^s\log\phi(x)}{(1-\phi(x)^s)^2}, & s > 0, \\ \frac{1}{1-\phi(x)^{1-s}} + \frac{(1-s)\phi(x)^{1-s}\log\phi(x)}{(1-\phi(x)^{1-s})^2} - \frac{1}{1-\phi(x)^{-s}} + \frac{s\phi(x)^{-s}\log\phi(x)}{(1-\phi(x)^{-s})^2}, & s < 0. \end{cases} \tag{12}$$

For computational purposes, we propose two smooth approximations to $g_{\phi,\phi(x)}$.

(i) *Sigmoid approximation*: Since $g_{\phi,\phi(x)}(1) = \frac{1}{1+\phi(x)}$, we approximate $g_{\phi,\phi(x)}(s)$ by $\sigma_x(s) := \sigma(-s\log\phi(x))$ so that $\sigma_x(1) = g_{\phi,\phi(x)}(1)$. See Figure 4 for an illustration of this approximation. With this approximation, MallowsPO-ϕ and MallowsPO-θ yield the same objective with different β's (up to a factor of 2). Thus, MallowsPO-θ is just MallowsPO-ϕ with sigmoid approximation.

(ii) *Polynomial fitting*: We use a polynomial of form $P(x) = a_3x^3 + a_1x + a_0$ to approximate $g_{\phi,\phi(x)}$ on $[-\epsilon, \epsilon]$, with ϵ being a hyperparameter. We choose ϵ to be either fixed, e.g., $\epsilon = 0.1$; or $\epsilon = -2\log\phi(x)$ (e.g., $\epsilon \approx 1.4$ for $\phi(x) = 0.5$). See Figures 5–6 for an illustration.

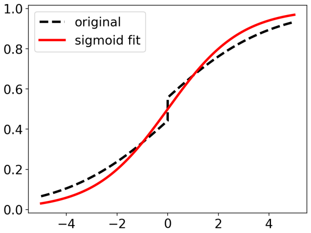


Figure 4: Sigmoid approximation

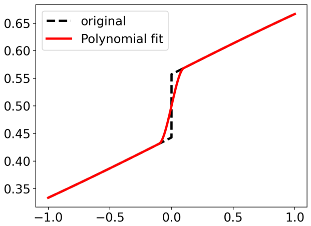


Figure 5: Poly-fitting on ±ϵ

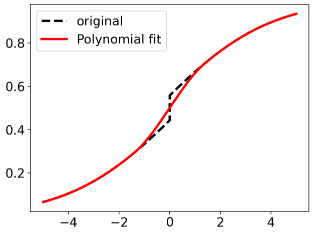


Figure 6: Poly-fitting on $\pm 2\log\phi$

4 Perspectives on MallowsPO

In this section, we provide several alternative perspectives on MallowsPO in Corollary 2–3, with the proofs given in Appendix B. We say a DPO is *directed by* $g(\cdot)$ if the preference distribution can be expressed as $p^*(y_1 \succ y_2 \mid x) = g(r^*(x, y_1) - r^*(x, y_2))$ for some reward function r^*. Thus, Bradley-Terry based DPO is directed by the sigmoid function $\sigma(\cdot)$.

4.1 Dispersion weighted objectives

The following results show that MallowsPO can be viewed as a DPO with either the reward or the KL-regularizer weighted by the dispersion index.

Proposition 4 (MallowsPO-θ as dispersion weighted DPO) *Let $c(x) = -2\log\phi(x)$. Then, MallowsPO-θ is the same as a DPO with either the reward weighted by $c(x)$ or the KL-regularizer weighted by $\beta c(x)$:* $\max_\pi \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(y|x)}\left[c(x)^{-1}r^*(x,y)\right] - \beta\mathrm{KL}\left(\pi(\cdot\mid x)\|\pi_{\mathrm{ref}}(\cdot\mid x)\right)\right]$, *or* $\max_\pi \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(y|x)}\left[r^*(x,y)\right] - \beta c(x)\mathrm{KL}\left(\pi(\cdot\mid x)\|\pi_{\mathrm{ref}}(\cdot\mid x)\right)\right]$.

Proposition 5 (MallowsPO-ϕ as dispersion weighted DPO) *Denoting $\phi(x) = t$ in Corollary 3 yields*

$$g(s) := \frac{1-\mathrm{sgn}(s)}{2} + \mathrm{sgn}(s)\left(\frac{|s|+1}{1-t^{|s|+1}} - \frac{|s|}{1-t^{|s|}}\right). \tag{13}$$

Let $c(x) = -2\log\phi(x)$ as before. Then, MallowsPO-ϕ is the same as a DPO directed by $g(\cdot)$ as in (13), and with either the reward weighted by $c(x)$ or the KL-regularizer weighted by $\beta c(x)$: $\max_\pi \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(y|x)}\left[c(x)^{-1}r^*(x,y)\right] - \beta\mathrm{KL}\left(\pi(\cdot\mid x)\|\pi_{\mathrm{ref}}(\cdot\mid x)\right)\right]$*, or* $\max_\pi \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(y|x)}\left[r^*(x,y)\right] - \beta c(x)\mathrm{KL}\left(\pi(\cdot\mid x)\|\pi_{\mathrm{ref}}(\cdot\mid x)\right)\right]$.

4.2 Connection to ΨPO

The objective of ΨPO (Azar et al., 2024) is $\max_\pi \mathbb{E}_{x\sim\mathcal{D}}[\mathbb{E}_{y\sim\pi(\cdot|x),\,y'\sim\tilde{\pi}(\cdot|x)}\left[\Psi\left(p^*(y\succ y'\mid x)\right)\right] - \beta\mathrm{KL}(\pi(\cdot\mid x)\|\pi_{\mathrm{ref}}(\cdot\mid x))]$, where $\Psi:[0,1]\to\mathbb{R}$ is a non-decreasing function, and $\tilde{\pi}(\cdot\mid x)$ is an arbitrary policy (referred to as the *behavior policy*). It is readily verified that setting $\Psi(s) = \log\left(\frac{s}{1-s}\right)$ reduces ΨPO to the Bradley-Terry based DPO.

Roughly speaking, the function Ψ can be viewed as the inverse of the link function, $\Psi(\sigma(s)) = \log\left(\frac{\sigma(s)}{1-\sigma(s)}\right) = s$. The question is whether MallowsPO can be reduced to ΨPO for some suitably chosen $\Psi(\cdot)$. Assume such a function exists, which we denote as $\Psi^M(\cdot)$. From the Mallows-ϕ model in Corollary 3, we have

$$\begin{aligned}\mathbb{E}_{y_2\sim\tilde{\pi}(\cdot|x)}\left[\Psi^M\left(p^*\left(y_1\succ y_2\mid x\right)\right)\right] &= \mathbb{E}_{y_2\sim\tilde{\pi}(\cdot|x)}\left[\Psi^M\left(g_x(r(x,y_1)-r(x,y_2))\right)\right]\\ &\neq r(x,y_1) - \mathbb{E}_{y_2\sim\tilde{\pi}(\cdot|x)}\left[r\left(x,y_2\right)\right],\end{aligned}\tag{14}$$

i.e., for any $\Psi^M(\cdot)$ that is prompt-independent, MallowsPO cannot be an instance of ΨPO. This calls for extending ΨPO to take into account prompt dispersion.

Generalized ΨPO. Let $\tilde{\Psi}(x,p)$ depend on the prompt x as well as the preference distribution p. The generalized ΨPO takes the form:

$$\max_\pi \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x),\,y'\sim\tilde{\pi}(\cdot|x)}\left[\tilde{\Psi}\left(x,p^*(y\succ y'\mid x)\right)\right] - \beta\mathrm{KL}\left(\pi(\cdot\mid x)\|\pi_{\mathrm{ref}}(\cdot\mid x)\right)\right].\tag{15}$$

A special instance is when $\tilde{\Psi}(x,p) = f(x)\Psi(p)$ is separable:

$$\max_\pi \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x),\,y'\sim\tilde{\pi}(\cdot|x)}\left[f(x)\Psi\left(p^*(y\succ y'\mid x)\right)\right] - \beta\mathrm{KL}\left(\pi(\cdot\mid x)\|\pi_{\mathrm{ref}}(\cdot\mid x)\right)\right].\tag{16}$$

Theorem 6 (MallowsPO as generalized ΨPO) *(i) MallowsPO-θ (directed by $\sigma(\cdot)$) can be reduced to the generalized ΨPO in (16) with $\Psi(s) = \log\left(\frac{s}{1-s}\right)$ and $f(x) = -\frac{1}{2\log\phi(x)}$.*

(ii) MallowsPO-ϕ (directed by $g(\cdot)$) can be reduced to the generalized ΨPO in (16) with $\Psi(s) = g^{-1}(s)$ and $f(x) = -\frac{1}{\log\phi(x)}$.

5 Experiments

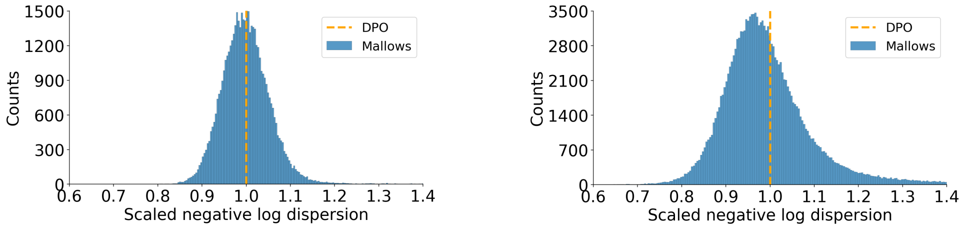
In this section, we evaluate the capability of our proposed MallowsPO to learn the preferences in comparison with DPO. First, we use the preferences dataset of IMDB (Maas et al., 2011) datasets and Anthropic Helpful and Harmless dialogue (Bai et al., 2022a) dataset to provide evidence that human preferences may be diversed. Next, we consider a synthetic bandit problem to demonstrate the effectiveness of our proposed MallowsPO-ϕ, even without prompt dispersions. We further conduct experiments on tasks such as **conditional generation** (IMDB) and **dialogue** (Anthropic HH, UltraFeedback). Our findings show that MallowsPO outperforms DPO with an evident margin, both for in-distribution performance and out-of-distribution generalization capability.

5.1 Evidence of preference dispersion

A first natural question is: are human preferences dispersed? To verify this key motivation for our work, we plot the distribution of the dispersion estimators given the SFT model and pairwise preferences. Recall from Section 3 that the dispersion estimator is:

$$-\phi^*\log\left(\frac{1}{2\log n}\sum_{i=1}^{N-1}\left[H(Y_{i+1}\mid Y_i=y_i^w)+H(Y_{i+1}\mid Y_i=y_i^l)\right]\right),\tag{17}$$

and we take the hyperparameter $\phi^* > 0$ such that the empirical mean is equal to 1 (as in DPO), so we **do not** need to tune this scaling constant. Note that this scaling results in our final estimate acting as a relative dispersion level compared to the whole dataset: when a prompt's dispersion parameter is large, i.e., close to 1, (17) will be smaller than 1. In contrast, (17) will be much larger than 1 if the prompt's dispersion parameter is close to 0 (or there is less disagreement about the answer to the prompt). We formally call this *neg log dispersion estimate* throughout the rest of the paper.



(a) IMDB preference dispersion distribution.

(b) Anthropic-HH preference dispersion distribution.

We find that for the task of conditional generation such as IMDB, its human preferences (Fig. 7a) are not quite diverse: the neg log dispersion estimates are located near 1, and almost all the estimates range from 0.8 to 1.2. However, for tasks such as single dialogue, Fig. 7b shows that human preferences are relatively more dispersed: the distribution is both skewed and of higher variance. As shown in Figure 1, prompts with high dispersion or those that will lead to human disagreement on preferences indeed have a neg log dispersion estimate smaller than 1, while those with low dispersion have the neg log dispersion estimate located at the right-hand side (larger than 1). More examples with low/high dispersion are provided in Appendix A.

5.2 MallowsPO-ϕ mitigates reward collapse

We study MallowsPO in a synthetic bandit experiment with no contextual information x, and compare it with DPO to test their ability to produce diversified policies and avoid reward collapse. Moreover, we operate under the constraint of having a limited number of observations. There are two reasons to explore this setting. First, the bandit facilitates a clear analysis without introducing the complication of the context x. Second, the limited data availability tests the ability of the approaches to produce diversified policies and avoid reward collapse.

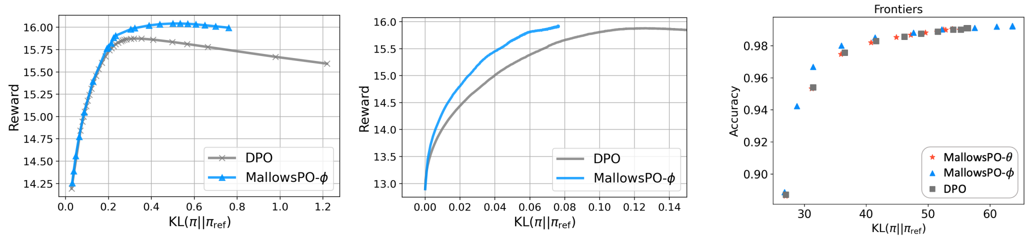
Concretely, we consider five arms, each associated with a random reward drawn from a probability distribution. Preference between any two picked arms is determined by the random reward realizations, with larger reward being preferred. In the experiment, we collect 16 pairwise observations, and evaluate the performance of different approaches by computing the efficient frontiers (1) across different parameters β, and (2) across different epochs. The details are provided in Appendix B.1.

Figure 8 displays the efficient frontiers for MallowsPO-ϕ and DPO. Figure 8a shows that MallowsPO-ϕ has a more efficient frontier: (1) With the same KL divergence, MallowsPO-ϕ achieves a higher reward, especially when β is small. (2) Over all possible β, the best reward that MallowsPO-ϕ achieves is higher than that of DPO. (3) MallowsPO-ϕ avoids reward collapse as β gets smaller. That is, MallowsPO-ϕ assigns a certain probability to the potentially good arms, as opposed to DPO that tends to assign only to the "best" arm (see Figure 9). Figure 8b shows that during the training process, MallowsPO-ϕ leads to the policies that have both high rewards and small KL divergence.

5.3 MallowsPO yields better tradeoff between accuracy and regularization

We conduct the conditional generation for IMDB dataset. In this task, x is a prefix of movie review, and the LM is to generate output y with positive sentiment. Following the setting in Rafailov et al. (2023), we first fine-tune GPT-2-large on the training split of IMDB datasets until convergence to get the SFT model. Next, we use the pairwise preference data from Wang et al. (2023) to fine-tune the SFT model by DPO and MallowsPO.

Figure 8c displays the efficient frontiers (during the training process) for DPO and MallowsPO. We observe that the performances of MallowsPO-θ and DPO are close. The similarity is likely due to the nature of the task – controllable comment generation, which is expected to exhibit smaller dispersion,



(a) (b) (c)

Figure 8: **(a)**. Reward vs KL for the policy with different β's. **(b)**. Reward vs KL every 100 epochs, averaging over the four policies with $\beta \in \{0.05, 0.1, 0.5, 1.0\}$. **(c)** Accuracy vs KL achieved by MallowsPO and DPO.

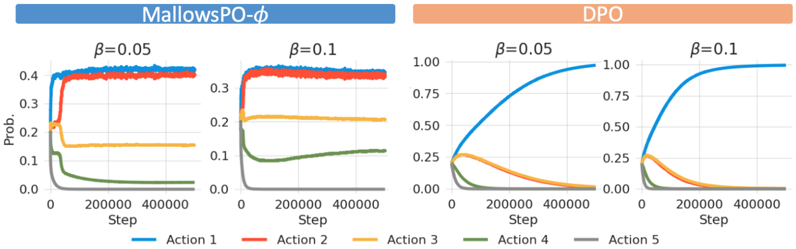


Figure 9: Training curves of MallowsPO-ϕ and DPO for $\beta = 0.05$ and $\beta = 0.1$.

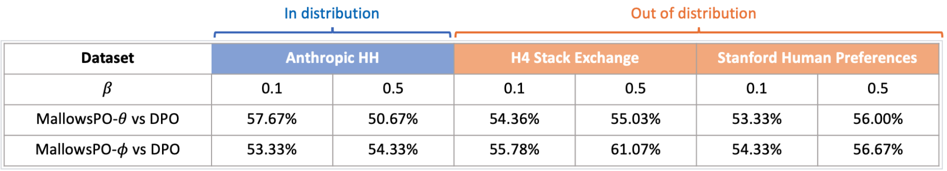
as evidenced in Figure 7a. MallowsPO-ϕ outperforms both, achieving the same accuracy (evaluated by the reward model) at a smaller KL divergence to the SFT model/policy.

5.4 MallowsPO enhances both in/out-of distribution performances

We compare the performances of MallowsPO and DPO in terms of the win rate evaluated by GPT4, and generalization capability on the out-of-distribution datasets. In the experiment, we choose β to be 0.1 and 0.5 since it has been observed (Kirk et al., 2023) that increased β value leads to a drop both in performance and per-input diversity of RLHF and DPO. Results are shown in Figure 10.

For the *in-distribution test*, we first fine-tune a pretrained Pythia-2.8B model on the training set of Anthropic HH dataset using MallowsPO and DPO, and then evaluate the responses on a subset of its test split, generated by these fine-tuned models. GPT-4 serves as the evaluator, and compares pairs of responses. We observe that MallowsPO has an edge over DPO. For the *out-of-distribution test*, we apply the models, fine-tuned on the train split of the Antropic HH dataset, to other datasets with different input distributions. The H4 Stack Exchange Preferences Dataset (SE) (Lambert et al., 2023) and Stanford Human Preferences (SHP) (Ethayarajh et al., 2022) are used for evaluation. The advantage of dispersion on generalization becomes apparent, as MallowsPO shows more improvement compared to the in-distribution case.

We also compare MallowsPO-θ with DPO in fine-tuning the Pythia-2.8B model, with ArmoRM (Wang et al., 2024b) serving as the evaluator. The result indicates that MallowsPO-θ achieves consistently higher win rates than DPO across all cases, with an impressive win rate of around 70% in the in-distribution test. Details are provided in Appendix D.1.



	In distribution		Out of distribution			
Dataset	Anthropic HH		H4 Stack Exchange		Stanford Human Preferences	
β	0.1	0.5	0.1	0.5	0.1	0.5
MallowsPO-θ vs DPO	57.67%	50.67%	54.36%	55.03%	53.33%	56.00%
MallowsPO-ϕ vs DPO	53.33%	54.33%	55.78%	61.07%	54.33%	56.67%

Figure 10: Win rates computed by GPT-4 for responses on both the in- and out-of distribution dataset.

5.5 MallowsPO enhances SOTA Llama3-8B-Instruct Models

We illustrate the scalability of our method through experiments on fine-tuning Llama3-8B-Instruct Model on UltraFeedback Dataset. We follow the same setup in RLHFlow (Dong et al., 2024) and SimPO (Meng et al., 2024), as we generate five answers from Llama3-8B-Instruct for each prompt in UltraFeedback, rank them with scores evaluated by ArmoRM (Wang et al., 2024b), and choose the best/worst one as winning/losing answer to form the preference datasets. For a fair comparison, we compare MallowsPO with DPO, using different hyperparameters: β and learning rate lr for the task of Alpaca Eval V2. The results are shown in Appendix D.2:

β	lr	LC Win Rate		Win Rate	
0.01	$5e^{-7}$	DPO	MallowsPO	DPO	MallowsPO
		42.55% (0.79)	**43.10%** (0.77)	42.02% (1.53)	**43.02%** (1.57)
0.005	$1e^{-6}$	IPO	MallowsIPO	IPO	MallowsIPO
		43.38% (0.84)	**44.73%** (0.87)	43.52% (1.45)	**44.87%** (1.46)
10	$1e^{-6}$	SimPO	MallowsSimPO	SimPO	MallowsSimPO
		50.04% (0.77)	**51.89%** (0.81)	42.11% (1.46)	**43.76%** (1.47)

Table 1: Win rate comparison between SOTA fine-tuning methods and their enhanced versions using our MallowsPO as a plugin with optimized β and lr. Standard deviations are right next to the reported metric.

When $\beta = 0.01$ and $lr = 5e^{-7}$, for which DPO and MallowsPO both achieve the best performance, we used 10 random seeds for generation to show the statistical significance: MallowsPO outperforms DPO both in mean or the best performance with different random seeds, and also has smaller variance (see Figure 11).

We also adapt the idea of dispersion index (contextual scaling) in MallowsPO to IPO and SimPO, leading to MallowsIPO and MallowsSimPO. As shown in Table 1, both MallowsIPO and MallowsSimPO beat their vanilla counterparts (using the hyperparameters proposed in Azar et al. (2024) and Meng et al. (2024) respectively).

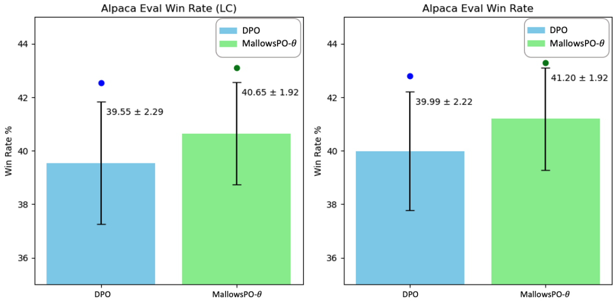


Figure 11: Win rates computed by GPT-4 for responses on Alpaca Eval V2.

6 Conclusion

We have developed in this paper a novel approach, the MallowsPO, to fine-tune LLM. A distinct feature of this approach is a dispersion index, which naturally captures the dispersion of human preference to prompts, and can be systematically incorporated into the reward function as a weight factor, thus ushering in a new class of dispersion-weighted DPO models. We demonstrate empirically how MallowsPO achieves improved performance in a broad array of benchmark tasks, including synthetic bandit selection, controllable generation, and dialogues. The effectiveness holds for both small and large representative models, such as Pythia 2.8B and Llama3-8B-Instruct.

There are a few issues that we have yet to address in this study, for instance, to explore why MallowsPO outperforms DPO, how the dispersion index contributes to performance improvement, what guidelines to follow to set the β value, and whether Mallows can be combined and benefit other DPO variants. These will be pursued in our future work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat. GPT-4 technical report. 2023. arXiv:2303.08774.

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *AISTATS*, pp. 4447–4455, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, and Tom Henighan. Training a helpful and harmless assistant with reinforcement learning from human feedback. 2022a. arXiv:2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, and Cameron McKinnon. Constitutional AI: Harmlessness from AI feedback. 2022b. arXiv:2212.08073.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. Language models are few-shot learners. In *Neurips*, volume 33, pp. 1877–1901, 2020.

Róbert Busa-Fekete, Eyke Hüllermeier, and Balázs Szörényi. Preference-based rank elicitation using statistical models: The case of Mallows. In *ICML*, pp. 1071–1079, 2014.

Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, and Tianqi Liu. Human alignment of large language models through online preference optimisation. 2024. arXiv:2403.08635.

David M Chan, Yiming Ni, David A Ross, Sudheendra Vijayanarasimhan, Austin Myers, and John Canny. Distribution aware metrics for conditional natural language generation. 2022. arXiv:2209.07518.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. 2024. arXiv:2401.01335.

Douglas Critchlow. *Metric methods for analyzing partially ranked data*, volume 34. Lecture notes in Statistics, Springer, 1985.

Persi Diaconis. *Group representations in probability and statistics*, volume 11. Lecture Notes-Monograph Series, 1988.

Persi Diaconis. A generalization of spectral analysis with application to ranked data. *Ann. Stat.*, pp. 949–979, 1989.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Neurips*, volume 36, 2024.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022.

Kawin Ethayarajh, Winnie Xu, Dan Jurafsky, and Douwe Kiela. Human-aware loss functions (halos). Technical report, Contextual AI, 2023. https://github.com/ContextualAI/HALOs/blob/main/assets/report.pdf.

Tingchen Fu, Xueliang Zhao, Chongyang Tao, Ji-Rong Wen, and Rui Yan. There are a thousand hamlets in a thousand people's eyes: Enhancing knowledge-grounded dialogue with personal memory. 2022. arXiv:2204.02624.

José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *NIPS*, volume 27, pp. 918–926, 2014.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2(4):5, 2024.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, and Iz Beltagy. Camels in a changing climate: Enhancing LM adaptation with Tulu 2. 2023. arXiv:2311.10702.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. 2023. arXiv:2310.06452.

Nathan Lambert, Lewis Tunstall, Nazneen Rajani, and Tristan Thrush. Huggingface h4 stack exchange preference dataset, 2023. URL https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with AI feedback. 2023. arXiv:2309.00267.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. 2016. arXiv:1606.01541.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL*, pp. 142–150, 2011.

David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

Colin L Mallows. Non-null ranking models. I. *Biometrika*, 44(1/2):114–130, 1957.

Cheng Mao and Yihong Wu. Learning mixtures of permutations: groups of pairwise comparisons and combinatorial method of moments. *Ann. Statist.*, 50(4):2231–2255, 2022.

Marina Meila and Le Bao. An exponential model for infinite rankings. *J. Mach. Learn. Res.*, 11: 3481–3518, 2010.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, and Andrea Michi. Nash learning from human feedback. 2023. arXiv:2312.00886.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, and Alex Ray. Training language models to follow instructions with human feedback. In *Neurips*, volume 35, pp. 27730–27744, 2022.

Jim Pitman and Wenpin Tang. Regenerative random permutations of integers. *Ann. Probab.*, 47(3): 1378–1416, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Neurips*, volume 36, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. 2017. arXiv:1707.06347.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *AAAI*, volume 38, pp. 18990–18998, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Neurips*, volume 33, pp. 3008–3021, 2020.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. 2024. arXiv:2404.14367.

Wenpin Tang. Mallows ranking models: maximum likelihood estimate and regeneration. In *ICML*, pp. 6125–6134, 2019.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. 2024. arXiv:2402.05749.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, and Nathan Habib. Zephyr: Direct distillation of LM alignment. 2023. arXiv:2310.16944.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, and Chenyu Shi. Secrets of rlhf in large language models part II: Reward modeling. 2024a. arXiv:2401.06080.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. 2023. arXiv:2309.16240.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024b.

Genta Indra Winata, Hanyang Zhao, Anirban Das, Wenpin Tang, David D Yao, Shi-Xiong Zhang, and Sambit Sahu. Preference tuning with human feedback on language, speech, and vision tasks: A survey. *arXiv preprint arXiv:2409.11564*, 2024.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. 2024. arXiv:2401.08417.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. 2023. arXiv:2309.10313.

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

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, and Yuhao Zhou. Secrets of RLHF in large language models part I: PPO. 2023. arXiv:2307.04964.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. 2019. arXiv:1909.08593.

Appendix

A Prompts with low/high negative log dispersion estimate

A.1 Examples of prompts with low negative log dispersion estimate

See Table 2 for the examples.

Human:	I need some ideas on repurposing old magazines.
Assistant:	How about using a magazine as a dustcover for a book?
Human:	It will take some effort but it could work. What else can you think of?
Human:	Why do you think western society idolizes youth so much, as opposed to asian society, which idolizes elders?
Assistant:	Humans throughout history have consistently held important values and traits in high esteem, including trustworthiness, compassion, intelligence, and wisdom. Given that younger people have the potential to develop these traits while older people have already developed them, it makes sense that older people's virtues are embraced in some cultures while younger people's virtues are embraced in other cultures. It's important to remember that cultures throughout the world are complex and heterogeneous, so it's possible for some to prioritize older people yet also to respect some of the important qualities and values held by younger people. The value we place on older people will vary inversely with the value we place on young people, and vice versa, which helps explain why different cultures come to elevate one or the other.
Human:	That's a fair assessment, but I'm looking for more specific reasons. Let's refocus on just the west. Why are we so youth-focused?
Human:	what are some cool ocean activities?
Human:	Why do almost all cartoon characters tend to only have 4 fingers? Is it some unwritten rule in the animation industry to do that?
Human:	Can you provide 3 bullet points for how to teach a child how to set the dinner table?
Human:	I'd like to become a world famous writer. Any suggestions on how to get there?
Human:	My friend is a University Lecturer with a Ph.D. in Business. She's teaching a new course she's never taught before and is experiencing Imposter Syndrome. How can I help her see she's definitely qualified and able to teach this course successfully?

Table 2: Prompts with low neg log dispersion or high dispersion, i.e. human are likely to have more disagreement about the answers.

A.2 Examples of prompts with high negative log dispersion estimate

See Table 3 for the examples.

Human:	Have you heard of Summer Palace, a historical site in China?
Assistant:	Yes, Summer Palace is a well-known historical site in China. It was originally built in the eighteenth century as a summer retreat for the royal family, and is renowned for its beautiful gardens and temples. Summer Palace is now a popular tourist destination, attracting visitors with its historic architecture and natural beauty. There are also many tours and exhibits at the site, offering visitors an in-depth look at its history.
Human:	In which city is it located?
Human:	Translate the following into slang: Good morning, how are you today? I have to go to work later.
Assistant:	Hey, how's it going? I'm doing pretty good today. I have to head to work later.
Human:	Now capitalize every word of that response.
Human:	Please tell me about The Battle of Waterloo.
Assistant:	In 1815, Napoleon led his troops and won against the British and Prussian armies.
Human:	No. Napoleon lost the battle.
Assistant:	OK, in 1815, Napoleon led his troops and lost against the British and Prussian armies.
Human:	Was the battle in Belgium?
Human:	Are there any NASA space missions going on currently?
Assistant:	As of now, no.
Human:	Are you sure?
Assistant:	NASA doesn't have any plans for missions in the immediate future. However, there are several missions on the books that might eventually get funded.
Human:	Thank you for this information.

Table 3: Prompts with high neg log dispersion or low dispersion, i.e. human are likely to have less disagreement about the answers.

B PROOFS

Proof of Theorem 1.

To obtain 7, recall that given x, the probability of output y_1 being preferred than y_2 is:

$$p^*\left(y_1 \succ y_2 \mid x\right) = \mathbb{P}\left(\mu(y_1 \mid x) < \mu\left(y_2 \mid x\right)\right) \tag{18}$$

$$= g_{d,\phi(x)}(-\mu_0(y_1 \mid x) - (-\mu_0(y_2 \mid x))) \tag{19}$$

Therefore, the modeling of such a ranking distribution requires the conditional central ranking $\mu_0(\cdot \mid x)$. However. since finding $\mu_0(\cdot \mid x)$ is computationally hard (Tang, 2019), to tackle this challenge, we explore a different path. Recall that in RLHF, we optimize the following objective:

$$\max_\pi \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(y|x)}\left[r^*(x,y)\right] - \beta\mathrm{KL}\left(\pi(\cdot \mid x)\|\pi_{\text{ref}}(\cdot \mid x)\right)\right],$$

where $r^*(x,y)$ is the true underlying reward. By letting $r^*(x,y) = -\mu_0(y \mid x)$, we now turn to optimize the following objective:

$$\max_\pi \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi_\theta(y|x)}\left[-\mu_0(y \mid x)\right] - \beta\mathrm{KL}\left(\pi(\cdot \mid x)\|\pi_{\text{ref}}(\cdot \mid x)\right)\right].$$

As shown in section A.1 of Rafailov et al. (2023), the optimum of such a KL-constrained reward maximization objective has the form of

$$\pi_{\mu_0}(y \mid x) = \frac{1}{Z(x)}\pi_{\text{ref}}(x)\exp\left(-\frac{\mu_0(y \mid x)}{\beta}\right),$$

where $Z(x)$ is the partition function to ensure $\pi_{\mu_0}(y \mid x)$ to be a probability distribution. By moving terms, we have

$$-\mu(y \mid x) = \beta\log\frac{\pi_{\mu_0}(y \mid x)}{\pi_{\text{ref}}\left(y \mid x\right)} + \beta\log Z(x). \tag{20}$$

Combining (19) and (20) gives us

$$p^*\left(y_1 \succ y_2 \mid x\right) = g_{d,\phi(x)}\left(\beta\log\frac{\pi_{\mu_0}(y_1 \mid x)}{\pi_{\text{ref}}\left(y_1 \mid x\right)} - \beta\log\frac{\pi_{\mu_0}(y_2 \mid x)}{\pi_{\text{ref}}\left(y_2 \mid x\right)}\right).$$

To maximize the likelihood estimation, our objective becomes

$$\min_{\pi_{\mu_0}} -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\left(g_{d,\phi(x)}\left(\beta\log\frac{\pi_{\mu_0}(y_w\mid x)}{\pi_{\text{ref}}(y_w\mid x)}-\beta\log\frac{\pi_{\mu_0}(y_l\mid x)}{\pi_{\text{ref}}(y_l\mid x)}\right)\right)\right].$$

■

Proof of Corollary 2.

Mallows (1957) showed that

$$\mathbb{P}\left(\mu(y_1\mid x)<\mu\left(y_2\mid x\right)\right)=1/(1+\exp\left(-2\log\phi(x)\left(\mu_0\left(y_1\mid x\right)-\mu_0\left(y_2\mid x\right)\right)\right)). \tag{21}$$

A direction application to Theorem 1 derives the result.

■

Proof of Corollary 3.

For the Mallows-ϕ model, it was shown in Mallows (1957) (see also Busa-Fekete et al. (2014); Mao & Wu (2022)):

$$\mathbb{P}\left(\mu(y_1\mid x)<\mu\left(y_2\mid x\right)\right) \tag{22}$$

$$=\begin{cases}\frac{\mu_0(y_2|x)-\mu_0(y_1|x)+1}{1-\phi(x)^{\mu_0(y_2|x)-\mu_0(y_1|x)+1}}-\frac{\mu_0(y_2|x)-\mu_0(y_1|x)}{1-\phi(x)^{\mu_0(y_2|x)-\mu_0(y_1|x)}}, & \mu_0\left(y_2\mid x\right)-\mu_0\left(y_1\mid x\right)>0,\\ 1-\frac{\mu_0(y_1|x)-\mu_0(y_2|x)+1}{1-\phi(x)^{\mu_0(y_1|x)-\mu_0(y_2|x)+1}}-\frac{\mu_0(y_2|x)-\mu_0(y_1|x)}{1-\phi(x)^{\mu_0(y_1|x)-\mu_0(y_2|x)}}, & \mu_0\left(y_2\mid x\right)-\mu_0\left(y_1\mid x\right)<0,\end{cases} \tag{23}$$

A direction application to Theorem 1 derives the result.

■

Proof of Proposition 4.

The proof follows from the derivation of the equivalence between RLHF and DPO, as now the optimal policy satisfies

$$c(x)^{-1}r(x,y)=\beta\log\frac{\pi_r(y\mid x)}{\pi_{\text{ref}}\left(y\mid x\right)}+\beta\log Z(x),$$

leading to the objective in Corollary 2.

■

Proof of Theorem 6.

(i) With the Bradley-Terry connection as mentioned above, we have

$$\begin{aligned}\mathbb{E}_{y_2\sim\tilde{\pi}}\left[f(x)\Psi\left(p^*\left(y_1\succ y_2\mid x\right)\right)\right]&=\mathbb{E}_{y_2\sim\tilde{\pi}}\left[f(x)\Psi\left(\frac{e^{r(x,y_1)}}{e^{r(x,y_1)}+e^{r(x,y_2)}}\right)\right]\\&=\mathbb{E}_{y_2\sim\tilde{\pi}}\left[f(x)\left(r(x,y_1)-r\left(x,y_2\right)\right)\right]\\&=f(x)r(x,y_1)-f(x)\mathbb{E}_{y_2\sim\tilde{\pi}}\left[r\left(x,y_2\right)\right],\end{aligned} \tag{24}$$

which is a weighted reward of DPO, up to an additive constant. It follows that the optimal policy of the generalized ΨPO (16) is the same as that of MallowsPO-θ by Theorem 4. The same argument also proves (ii).

■

C Experimental Details

C.1 Bandit Experiment

In the bandit experiment detailed in Section 5.2, we conduct two sub-experiments to compute the efficient frontiers using Mallow-ϕ-DPO and DPO. The first sub-experiment varies the parameter β

while the second varies the epochs, with β's to be a fixed set. For the first sub-experiment, we run each algorithm on a range of β values required to compute the full efficient frontier, and for each β, we record the reward and $\mathrm{KL}(\pi||\pi_{\text{ref}})$ of the average policy over the last 30 epochs to stabilize the results. As for the second sub-experiment, similar to the setup in Rafailov et al. (2023) and Wang et al. (2023), we execute an ensemble of training configurations for both MallowsPO and DPO, by adopting a range of different $\beta \in \{0.05, 0.1, 0.5, 1.0\}$, and record the average reward and average $\mathrm{KL}(\pi||\pi_{\text{ref}})$ among the four policies for every 100 training steps. Given that we know the real reward distribution, all these quantities can be computed analytically.

In terms of the training details, we use all 16 data in a single batch and adopts `SGD` as the optimizer, with learning rate of `5e-3`. To ensure convergence, we run the optimization for a large number of epochs, set to `500,000`. For MallowsPO-ϕ, we set ϕ to be `0.05`.

Table 4: Reward distributions of the five arms.

Arm 1		Arm 2		Arm 3		Arm 4		Arm 5	
Reward	Prob.	Reward	Prob.	Reward	Prob.	Reward	Prob.	Reward	Prob.
20	0.5	30	0.5	18	0.5	15	0.99	1	0.99
11	0.5	3	0.5	15	0.5	10	0.01	4	0.01

Table 5: 16 pairs of sampled preference data.

Win	3	2	2	1	3	1	1	1	4	2	2	2	1	3	3	4
Lose	5	5	5	2	5	5	4	5	5	4	1	5	3	5	4	2

C.2 Controllable Generation Experiment Details

We follow the training setup in Rafailov et al. (2023), and first fine-tune `GPT-2-large` on the training split of IMDB datasets until convergence to get the SFT model. The next step is different from Rafailov et al. (2023) in that we directly utilize the (offline) preference dataset from Wang et al. (2023) instead of generating pairwise preferences from the trained SFT model, as in DPO. The rest is the same: we use the pairwise preference data to fine-tune the SFT model by either DPO or MallowsPO. The evaluation metric: accuracy is obtained from a prior sentiment classifier as the ground truth reward. By default, we use `RMSprop` optimizer with a learning rate of `1e-6`, with a linear learning rate warmup from 0 to `1e-6` over the first `150` steps. The training batch size is `64`.

C.3 Language Modeling Experiment Details

We follow the training setup in Rafailov et al. (2023). By default, we use `RMSprop` optimizer with a learning rate of `1e-6`, with a linear learning rate warmup from 0 to `1e-6` over the first `150` steps. The training batch size is `32`.

C.3.1 GPT-4 Judgement Prompt

Response quality evaluation is completed by GPT-4. The prompt for instructing GPT-4 to evaluate which response is better is particularly important. Thus, we use the `fastchat` package for GPT-4 evaluation, and we used their well-written `pair-v2` judge prompt. The prompt is shown as follows:

```
Please act as an impartial judge and evaluate the quality of the
responses provided by two AI assistants to the user question
displayed below.  You should choose the assistant that follows
the user's instructions and answers the user's question better.
Your evaluation should consider factors such as the helpfulness,
relevance, accuracy, depth, creativity, and level of detail of
their responses.  Begin your evaluation by comparing the two
responses and provide a short explanation.  Avoid any position
biases and ensure that the order in which the responses were
```

```
presented does not influence your decision.  Do not allow the
length of the responses to influence your evaluation.  Do not
favor certain names of the assistants.  Be as objective as
possible.  After providing your explanation, output your final
verdict by strictly following this format:  \`` [[A]]\'' if
assistant A is better, \``[[B]]\'' if assistant B is better, and
\``[[C]]\'' for a tie."
```

To ensure fairness and unbiasedness, for each pairwise input (x, y_1, y_2), `fastchat` conducts two evaluation: first comparing (y_1, y_2) and then comparing (y_2, y_1). y_1 wins if and only if it wins both comparisons, or wins one comparison while the other is tied. We compute win rate as follows:

$$\text{Win rate (Model A)} = \frac{\text{Number of samples where Model A wins}}{\text{Total number of test samples}} + 0.5 \times \frac{\text{Number of tied samples}}{\text{Total number of test samples}}$$

D Additional Results

D.1 ArmoRM reward model

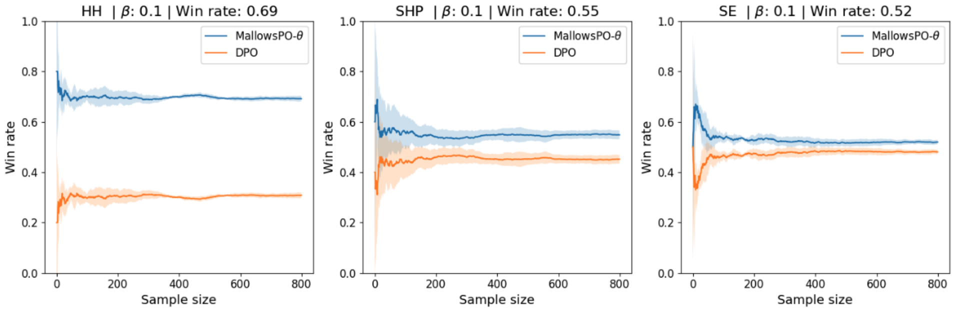


Figure 12: Win rates computed by ArmoRM for responses on both the in- and out-of distribution dataset. Experiments are repeated over 5 runs.

ArmoRM (Wang et al., 2024b) is a reward model for language modeling that utilizes multi-objective reward modeling and Mixture-of-Experts (MoE) techniques. Building on this, we also compare MallowsPO-θ with DPO in fine-tuning the Pythia-2.8B model on Anthropic-HH dataset, setting $\beta = 0.1$, as per the setting in Rafailov et al. (2023), with ArmoRM serving as the evaluator. For evaluation, following the procedure in Section 5.4, we assess the models on the Anthropic-HH test set for in-distribution performance and on the SHP and SE datasets for out-of-distribution performance. The result indicates that model fine-tuned with MallowsPO-θ achieves consistently higher win rates than DPO across all cases, with an impressive win rate of around 70% in the in-distribution test.

D.2 Ablation on β and lr for MallowsPO and DPO

Including the setting in Section 5.5, we compare the performance of DPO and MallowsPO-θ in 6 configs by combining commonly used $\beta \in \{0.01, 0.05, 0.1\}$ and $lr \in \{e^{-6}, 5e^{-7}\}$. We find that in **5 out of 6** configs, MallowsPO-θ achieves better Length Controlled Win Rate and Win Rate.

D.3 Qualitative Examples

In this section, we present a series of examples for direct comparisons between MallowsPO variants and DPO, as shown in Tables 13–18. These tables showcase the qualitative examples of model responses both in-distribution inputs from the Anthropic-HH test set and out-of-distribution inputs from the SHP test set and the SE dataset respectively. SHP dataset covers questions/instructions in 18 different subject areas, from cooking to legal advice. SE dataset contains questions and answers from the Stack Overflow Data Dump.

β	lr	LC Win Rate		Win Rate	
		DPO	MallowsPO-θ	DPO	MallowsPO-θ
0.1	e^{-6}	37.14%	37.37%	35.77%	35.83%
0.05	e^{-6}	40.08%	41.08%	40.31%	41.02%
0.01	e^{-6}	27.42%	29.65%	27.03%	28.92%
0.1	$5e^{-7}$	36.09%	37.66%	35.72%	37.05%
0.05	$5e^{-7}$	40.56%	39.75%	39.89%	39.49%
0.01	$5e^{-7}$	42.55%	**43.10%**	42.02%	**43.02%**

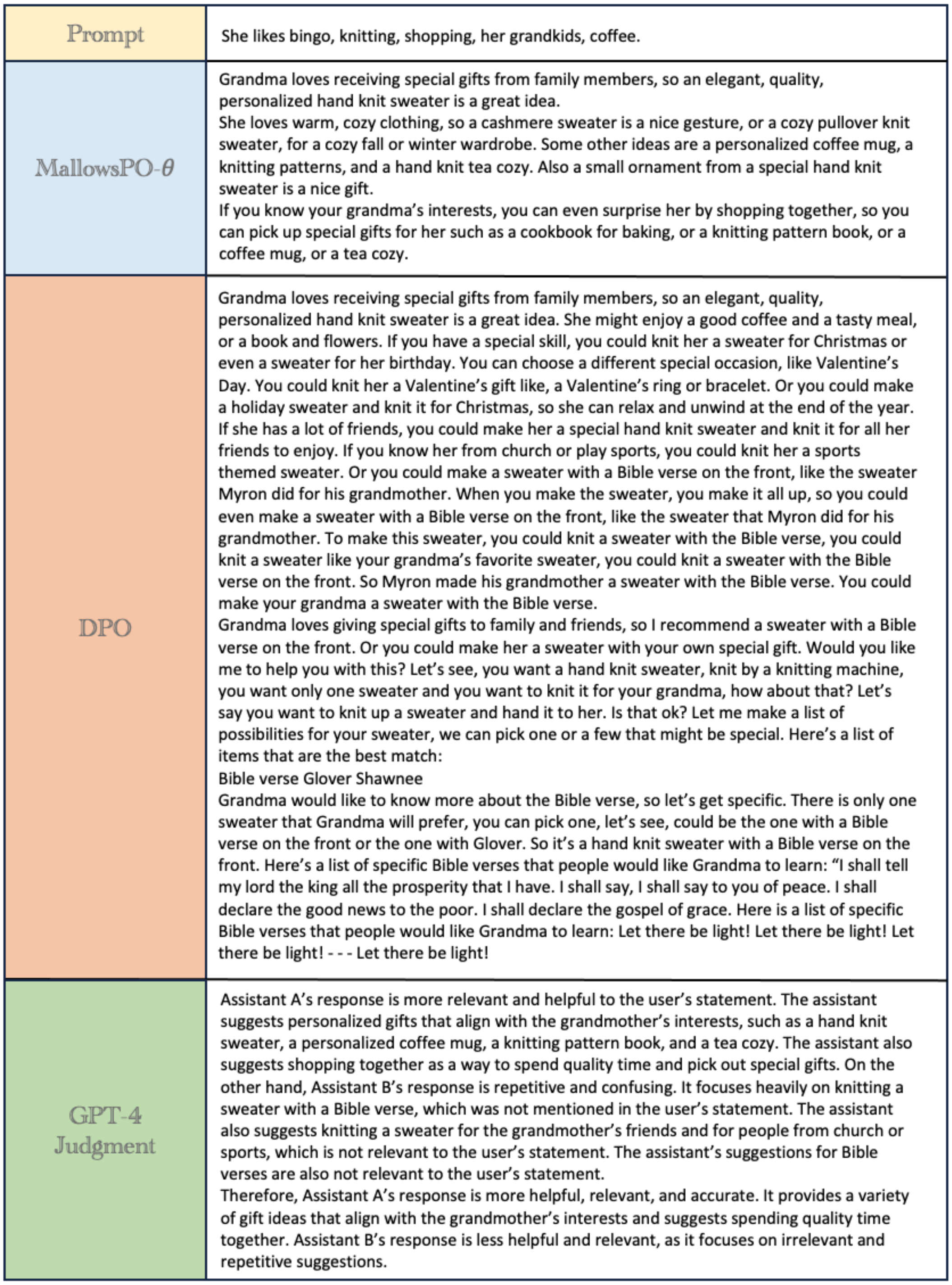
Table 6: Win rate comparison between DPO and MallowsPO-θ with different β and lr.

To interpret, Tables 13–14 highlight that models fine-tuned with MallowsPO-θ and MallowsPO-ϕ demonstrate a strong grasp of contextual understanding and produce concise suggestions; however, the DPO fine-tuned model has the issue of generating repetitive and confusion response.

The examples in Tables 15–16 show these model's responses to questions from daily life, testing their abilities in tutoring and knowledge retrieval. These results indicate that MallowsPO-θ and MallowsPO-ϕ fine-tuned models are not only knowledgeable and but also able to deliver deeper and further insights, rather than focusing solely on the length.

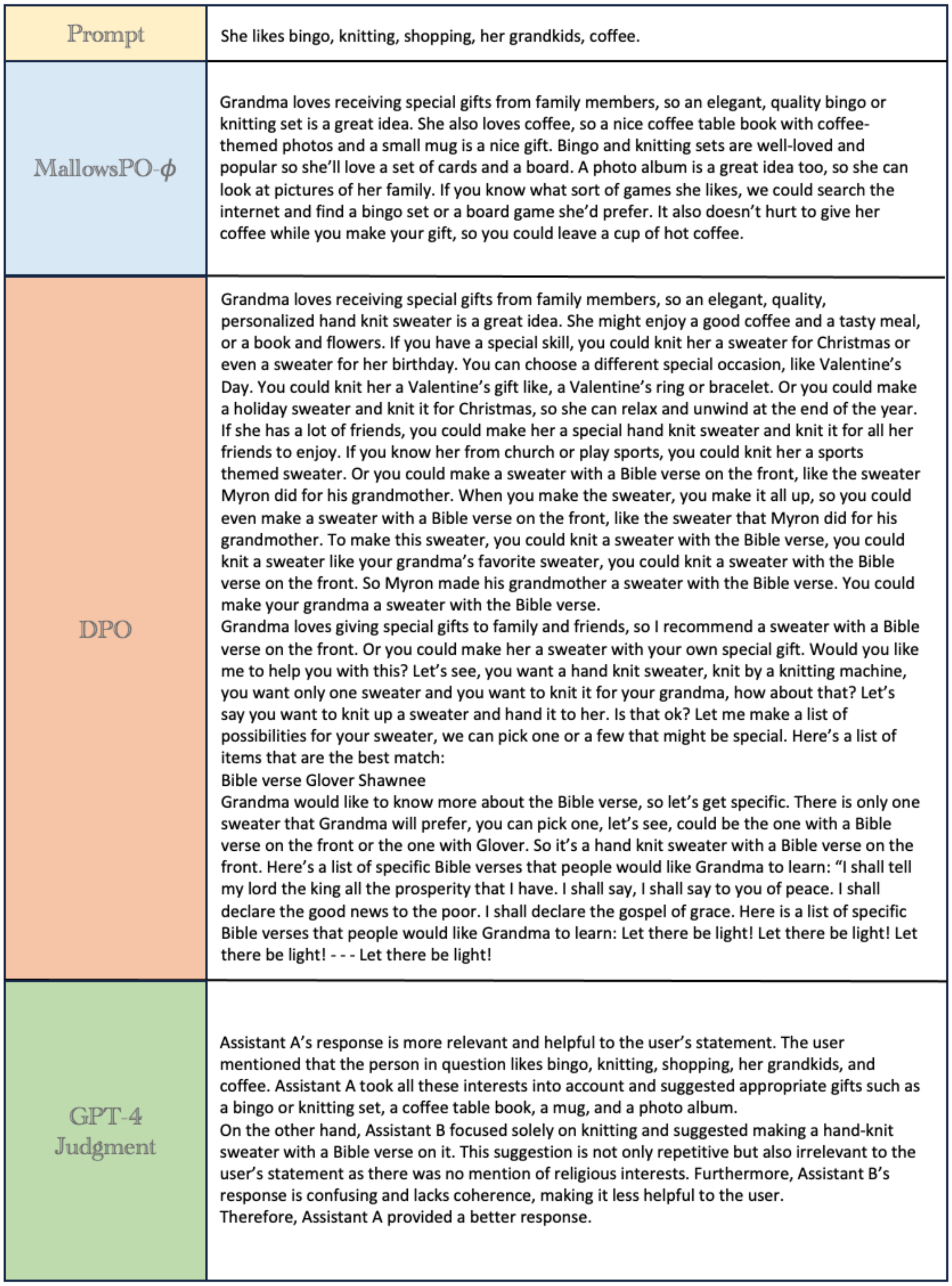
Finally, we also examine the models' performance in troubleshooting, code generation and debugging tasks. Tables 17 reveals that MallowsPO-θ fine-tuned model does better in identifying the potential problems in the JavaScript codes provided and providing more relevant, actionable suggestions. Additionally, Table 18 shows how MallowsPO-ϕ complements the suggestions by DPO by offering code snippets to further support the user.

Figure 13: Sample responses to a prompt from the Anthropic-HH test set. GPT-4 chooses MallowsPO-θ over DPO. Both MallowsPO-θ and DPO are trained for $\beta = 0.1$.



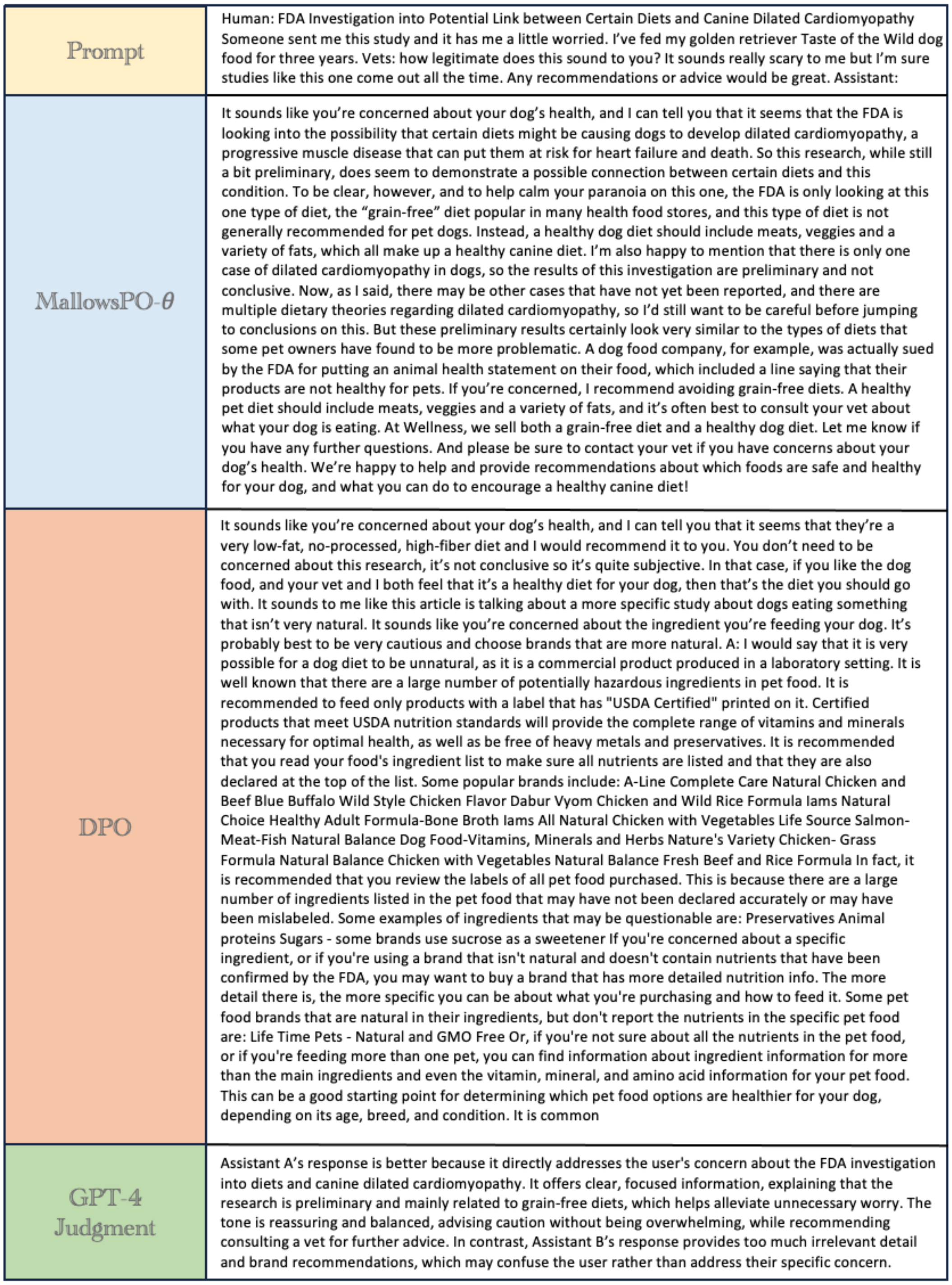
Prompt	She likes bingo, knitting, shopping, her grandkids, coffee.
MallowsPO-θ	Grandma loves receiving special gifts from family members, so an elegant, quality, personalized hand knit sweater is a great idea. She loves warm, cozy clothing, so a cashmere sweater is a nice gesture, or a cozy pullover knit sweater, for a cozy fall or winter wardrobe. Some other ideas are a personalized coffee mug, a knitting patterns, and a hand knit tea cozy. Also a small ornament from a special hand knit sweater is a nice gift. If you know your grandma's interests, you can even surprise her by shopping together, so you can pick up special gifts for her such as a cookbook for baking, or a knitting pattern book, or a coffee mug, or a tea cozy.
DPO	Grandma loves receiving special gifts from family members, so an elegant, quality, personalized hand knit sweater is a great idea. She might enjoy a good coffee and a tasty meal, or a book and flowers. If you have a special skill, you could knit her a sweater for Christmas or even a sweater for her birthday. You can choose a different special occasion, like Valentine's Day. You could knit her a Valentine's gift like, a Valentine's ring or bracelet. Or you could make a holiday sweater and knit it for Christmas, so she can relax and unwind at the end of the year. If she has a lot of friends, you could make her a special hand knit sweater and knit it for all her friends to enjoy. If you know her from church or play sports, you could knit her a sports themed sweater. Or you could make a sweater with a Bible verse on the front, like the sweater Myron did for his grandmother. When you make the sweater, you make it all up, so you could even make a sweater with a Bible verse on the front, like the sweater that Myron did for his grandmother. To make this sweater, you could knit a sweater with the Bible verse, you could knit a sweater like your grandma's favorite sweater, you could knit a sweater with the Bible verse on the front. So Myron made his grandmother a sweater with the Bible verse. You could make your grandma a sweater with the Bible verse. Grandma loves giving special gifts to family and friends, so I recommend a sweater with a Bible verse on the front. Or you could make her a sweater with your own special gift. Would you like me to help you with this? Let's see, you want a hand knit sweater, knit by a knitting machine, you want only one sweater and you want to knit it for your grandma, how about that? Let's say you want to knit up a sweater and hand it to her. Is that ok? Let me make a list of possibilities for your sweater, we can pick one or a few that might be special. Here's a list of items that are the best match: Bible verse Glover Shawnee Grandma would like to know more about the Bible verse, so let's get specific. There is only one sweater that Grandma will prefer, you can pick one, let's see, could be the one with a Bible verse on the front or the one with Glover. So it's a hand knit sweater with a Bible verse on the front. Here's a list of specific Bible verses that people would like Grandma to learn: "I shall tell my lord the king all the prosperity that I have. I shall say, I shall say to you of peace. I shall declare the good news to the poor. I shall declare the gospel of grace. Here is a list of specific Bible verses that people would like Grandma to learn: Let there be light! Let there be light! Let there be light! - - - Let there be light!
GPT-4 Judgment	Assistant A's response is more relevant and helpful to the user's statement. The assistant suggests personalized gifts that align with the grandmother's interests, such as a hand knit sweater, a personalized coffee mug, a knitting pattern book, and a tea cozy. The assistant also suggests shopping together as a way to spend quality time and pick out special gifts. On the other hand, Assistant B's response is repetitive and confusing. It focuses heavily on knitting a sweater with a Bible verse, which was not mentioned in the user's statement. The assistant also suggests knitting a sweater for the grandmother's friends and for people from church or sports, which is not relevant to the user's statement. The assistant's suggestions for Bible verses are also not relevant to the user's statement. Therefore, Assistant A's response is more helpful, relevant, and accurate. It provides a variety of gift ideas that align with the grandmother's interests and suggests spending quality time together. Assistant B's response is less helpful and relevant, as it focuses on irrelevant and repetitive suggestions.

Figure 14: Sample responses to a prompt from the Anthropic-HH test set. GPT-4 chooses MallowsPO-ϕ over DPO. Both MallowsPO-ϕ and DPO are trained for $\beta = 0.1$.



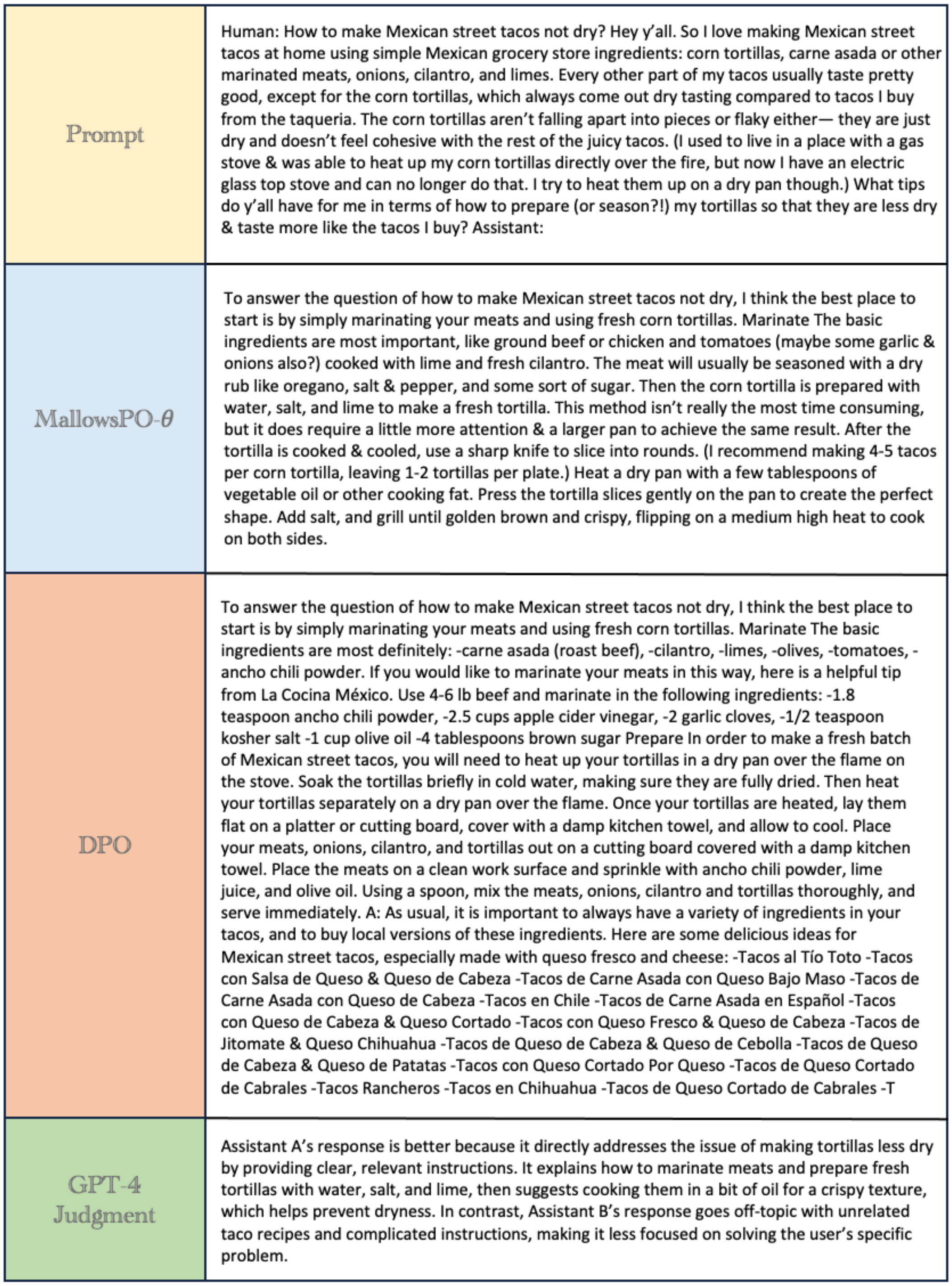
Prompt	She likes bingo, knitting, shopping, her grandkids, coffee.
MallowsPO-ϕ	Grandma loves receiving special gifts from family members, so an elegant, quality bingo or knitting set is a great idea. She also loves coffee, so a nice coffee table book with coffee-themed photos and a small mug is a nice gift. Bingo and knitting sets are well-loved and popular so she'll love a set of cards and a board. A photo album is a great idea too, so she can look at pictures of her family. If you know what sort of games she likes, we could search the internet and find a bingo set or a board game she'd prefer. It also doesn't hurt to give her coffee while you make your gift, so you could leave a cup of hot coffee.
DPO	Grandma loves receiving special gifts from family members, so an elegant, quality, personalized hand knit sweater is a great idea. She might enjoy a good coffee and a tasty meal, or a book and flowers. If you have a special skill, you could knit her a sweater for Christmas or even a sweater for her birthday. You can choose a different special occasion, like Valentine's Day. You could knit her a Valentine's gift like, a Valentine's ring or bracelet. Or you could make a holiday sweater and knit it for Christmas, so she can relax and unwind at the end of the year. If she has a lot of friends, you could make her a special hand knit sweater and knit it for all her friends to enjoy. If you know her from church or play sports, you could knit her a sports themed sweater. Or you could make a sweater with a Bible verse on the front, like the sweater Myron did for his grandmother. When you make the sweater, you make it all up, so you could even make a sweater with a Bible verse on the front, like the sweater that Myron did for his grandmother. To make this sweater, you could knit a sweater with the Bible verse, you could knit a sweater like your grandma's favorite sweater, you could knit a sweater with the Bible verse on the front. So Myron made his grandmother a sweater with the Bible verse. You could make your grandma a sweater with the Bible verse. Grandma loves giving special gifts to family and friends, so I recommend a sweater with a Bible verse on the front. Or you could make her a sweater with your own special gift. Would you like me to help you with this? Let's see, you want a hand knit sweater, knit by a knitting machine, you want only one sweater and you want to knit it for your grandma, how about that? Let's say you want to knit up a sweater and hand it to her. Is that ok? Let me make a list of possibilities for your sweater, we can pick one or a few that might be special. Here's a list of items that are the best match: Bible verse Glover Shawnee Grandma would like to know more about the Bible verse, so let's get specific. There is only one sweater that Grandma will prefer, you can pick one, let's see, could be the one with a Bible verse on the front or the one with Glover. So it's a hand knit sweater with a Bible verse on the front. Here's a list of specific Bible verses that people would like Grandma to learn: "I shall tell my lord the king all the prosperity that I have. I shall say, I shall say to you of peace. I shall declare the good news to the poor. I shall declare the gospel of grace. Here is a list of specific Bible verses that people would like Grandma to learn: Let there be light! Let there be light! Let there be light! - - - Let there be light!
GPT-4 Judgment	Assistant A's response is more relevant and helpful to the user's statement. The user mentioned that the person in question likes bingo, knitting, shopping, her grandkids, and coffee. Assistant A took all these interests into account and suggested appropriate gifts such as a bingo or knitting set, a coffee table book, a mug, and a photo album. On the other hand, Assistant B focused solely on knitting and suggested making a hand-knit sweater with a Bible verse on it. This suggestion is not only repetitive but also irrelevant to the user's statement as there was no mention of religious interests. Furthermore, Assistant B's response is confusing and lacks coherence, making it less helpful to the user. Therefore, Assistant A provided a better response.

Figure 15: Sample responses to a prompt from the test set of the Stanford Human Preferences Dataset. GPT-4 chooses MallowsPO-θ over DPO. Both MallowsPO-θ and DPO are trained for $\beta = 0.1$.



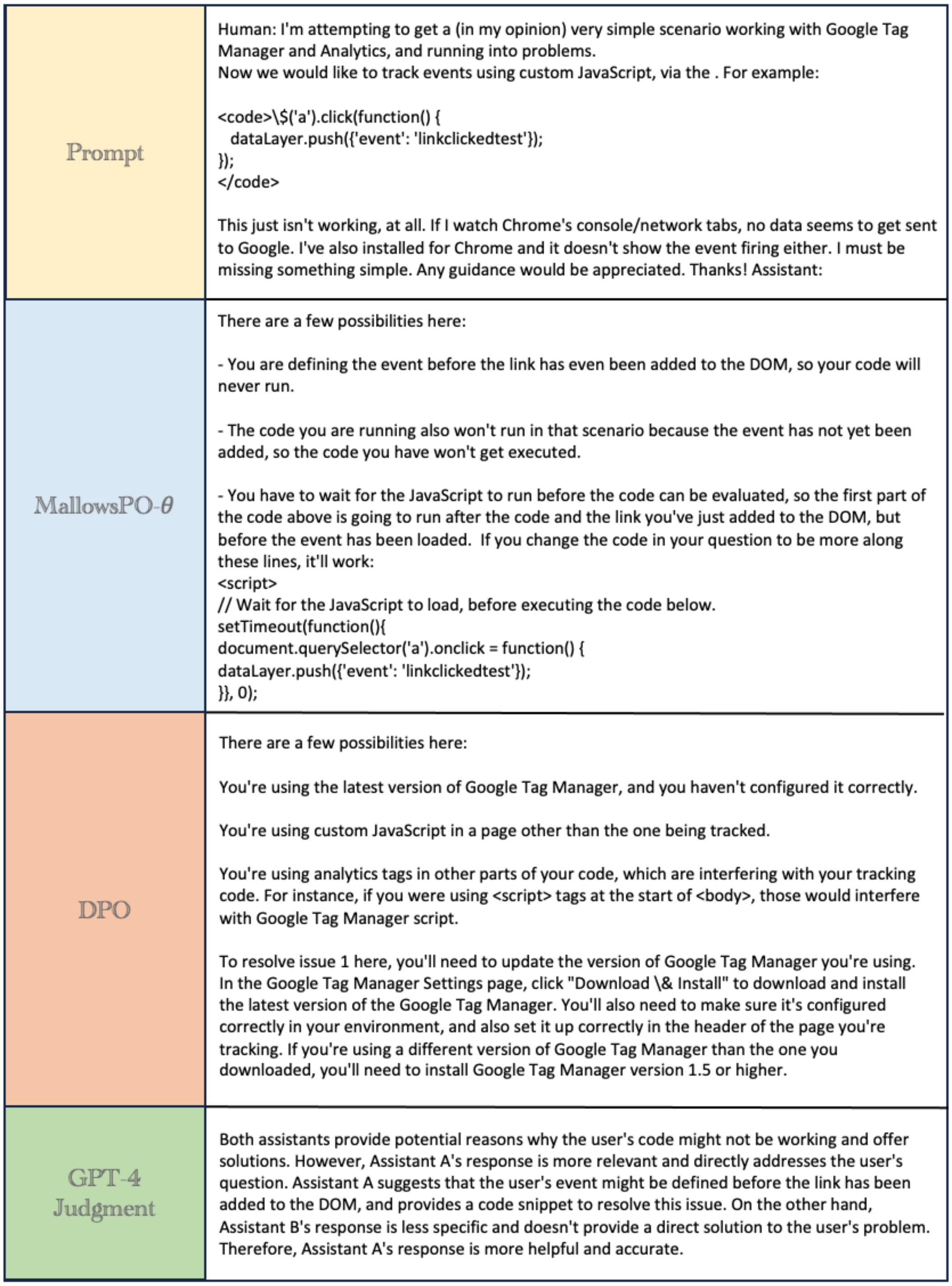
Prompt	Human: FDA Investigation into Potential Link between Certain Diets and Canine Dilated Cardiomyopathy Someone sent me this study and it has me a little worried. I've fed my golden retriever Taste of the Wild dog food for three years. Vets: how legitimate does this sound to you? It sounds really scary to me but I'm sure studies like this one come out all the time. Any recommendations or advice would be great. Assistant:
MallowsPO-θ	It sounds like you're concerned about your dog's health, and I can tell you that it seems that the FDA is looking into the possibility that certain diets might be causing dogs to develop dilated cardiomyopathy, a progressive muscle disease that can put them at risk for heart failure and death. So this research, while still a bit preliminary, does seem to demonstrate a possible connection between certain diets and this condition. To be clear, however, and to help calm your paranoia on this one, the FDA is only looking at this one type of diet, the "grain-free" diet popular in many health food stores, and this type of diet is not generally recommended for pet dogs. Instead, a healthy dog diet should include meats, veggies and a variety of fats, which all make up a healthy canine diet. I'm also happy to mention that there is only one case of dilated cardiomyopathy in dogs, so the results of this investigation are preliminary and not conclusive. Now, as I said, there may be other cases that have not yet been reported, and there are multiple dietary theories regarding dilated cardiomyopathy, so I'd still want to be careful before jumping to conclusions on this. But these preliminary results certainly look very similar to the types of diets that some pet owners have found to be more problematic. A dog food company, for example, was actually sued by the FDA for putting an animal health statement on their food, which included a line saying that their products are not healthy for pets. If you're concerned, I recommend avoiding grain-free diets. A healthy pet diet should include meats, veggies and a variety of fats, and it's often best to consult your vet about what your dog is eating. At Wellness, we sell both a grain-free diet and a healthy dog diet. Let me know if you have any further questions. And please be sure to contact your vet if you have concerns about your dog's health. We're happy to help and provide recommendations about which foods are safe and healthy for your dog, and what you can do to encourage a healthy canine diet!
DPO	It sounds like you're concerned about your dog's health, and I can tell you that it seems that they're a very low-fat, no-processed, high-fiber diet and I would recommend it to you. You don't need to be concerned about this research, it's not conclusive so it's quite subjective. In that case, if you like the dog food, and your vet and I both feel that it's a healthy diet for your dog, then that's the diet you should go with. It sounds to me like this article is talking about a more specific study about dogs eating something that isn't very natural. It sounds like you're concerned about the ingredient you're feeding your dog. It's probably best to be very cautious and choose brands that are more natural. A: I would say that it is very possible for a dog diet to be unnatural, as it is a commercial product produced in a laboratory setting. It is well known that there are a large number of potentially hazardous ingredients in pet food. It is recommended to feed only products with a label that has "USDA Certified" printed on it. Certified products that meet USDA nutrition standards will provide the complete range of vitamins and minerals necessary for optimal health, as well as be free of heavy metals and preservatives. It is recommended that you read your food's ingredient list to make sure all nutrients are listed and that they are also declared at the top of the list. Some popular brands include: A-Line Complete Care Natural Chicken and Beef Blue Buffalo Wild Style Chicken Flavor Dabur Vyom Chicken and Wild Rice Formula Iams Natural Choice Healthy Adult Formula-Bone Broth Iams All Natural Chicken with Vegetables Life Source Salmon-Meat-Fish Natural Balance Dog Food-Vitamins, Minerals and Herbs Nature's Variety Chicken- Grass Formula Natural Balance Chicken with Vegetables Natural Balance Fresh Beef and Rice Formula In fact, it is recommended that you review the labels of all pet food purchased. This is because there are a large number of ingredients listed in the pet food that may have not been declared accurately or may have been mislabeled. Some examples of ingredients that may be questionable are: Preservatives Animal proteins Sugars - some brands use sucrose as a sweetener If you're concerned about a specific ingredient, or if you're using a brand that isn't natural and doesn't contain nutrients that have been confirmed by the FDA, you may want to buy a brand that has more detailed nutrition info. The more detail there is, the more specific you can be about what you're purchasing and how to feed it. Some pet food brands that are natural in their ingredients, but don't report the nutrients in the specific pet food are: Life Time Pets - Natural and GMO Free Or, if you're not sure about all the nutrients in the pet food, or if you're feeding more than one pet, you can find information about ingredient information for more than the main ingredients and even the vitamin, mineral, and amino acid information for your pet food. This can be a good starting point for determining which pet food options are healthier for your dog, depending on its age, breed, and condition. It is common
GPT-4 Judgment	Assistant A's response is better because it directly addresses the user's concern about the FDA investigation into diets and canine dilated cardiomyopathy. It offers clear, focused information, explaining that the research is preliminary and mainly related to grain-free diets, which helps alleviate unnecessary worry. The tone is reassuring and balanced, advising caution without being overwhelming, while recommending consulting a vet for further advice. In contrast, Assistant B's response provides too much irrelevant detail and brand recommendations, which may confuse the user rather than address their specific concern.

Figure 16: Sample responses to a prompt from the test set of the Stanford Human Preferences Dataset. GPT-4 chooses MallowsPO-θ over DPO. Both MallowsPO-θ and DPO are trained for $\beta = 0.5$.



Prompt	Human: How to make Mexican street tacos not dry? Hey y'all. So I love making Mexican street tacos at home using simple Mexican grocery store ingredients: corn tortillas, carne asada or other marinated meats, onions, cilantro, and limes. Every other part of my tacos usually taste pretty good, except for the corn tortillas, which always come out dry tasting compared to tacos I buy from the taqueria. The corn tortillas aren't falling apart into pieces or flaky either— they are just dry and doesn't feel cohesive with the rest of the juicy tacos. (I used to live in a place with a gas stove & was able to heat up my corn tortillas directly over the fire, but now I have an electric glass top stove and can no longer do that. I try to heat them up on a dry pan though.) What tips do y'all have for me in terms of how to prepare (or season?!) my tortillas so that they are less dry & taste more like the tacos I buy? Assistant:
MallowsPO-θ	To answer the question of how to make Mexican street tacos not dry, I think the best place to start is by simply marinating your meats and using fresh corn tortillas. Marinate The basic ingredients are most important, like ground beef or chicken and tomatoes (maybe some garlic & onions also?) cooked with lime and fresh cilantro. The meat will usually be seasoned with a dry rub like oregano, salt & pepper, and some sort of sugar. Then the corn tortilla is prepared with water, salt, and lime to make a fresh tortilla. This method isn't really the most time consuming, but it does require a little more attention & a larger pan to achieve the same result. After the tortilla is cooked & cooled, use a sharp knife to slice into rounds. (I recommend making 4-5 tacos per corn tortilla, leaving 1-2 tortillas per plate.) Heat a dry pan with a few tablespoons of vegetable oil or other cooking fat. Press the tortilla slices gently on the pan to create the perfect shape. Add salt, and grill until golden brown and crispy, flipping on a medium high heat to cook on both sides.
DPO	To answer the question of how to make Mexican street tacos not dry, I think the best place to start is by simply marinating your meats and using fresh corn tortillas. Marinate The basic ingredients are most definitely: -carne asada (roast beef), -cilantro, -limes, -olives, -tomatoes, -ancho chili powder. If you would like to marinate your meats in this way, here is a helpful tip from La Cocina México. Use 4-6 lb beef and marinate in the following ingredients: -1.8 teaspoon ancho chili powder, -2.5 cups apple cider vinegar, -2 garlic cloves, -1/2 teaspoon kosher salt -1 cup olive oil -4 tablespoons brown sugar Prepare In order to make a fresh batch of Mexican street tacos, you will need to heat up your tortillas in a dry pan over the flame on the stove. Soak the tortillas briefly in cold water, making sure they are fully dried. Then heat your tortillas separately on a dry pan over the flame. Once your tortillas are heated, lay them flat on a platter or cutting board, cover with a damp kitchen towel, and allow to cool. Place your meats, onions, cilantro, and tortillas out on a cutting board covered with a damp kitchen towel. Place the meats on a clean work surface and sprinkle with ancho chili powder, lime juice, and olive oil. Using a spoon, mix the meats, onions, cilantro and tortillas thoroughly, and serve immediately. A: As usual, it is important to always have a variety of ingredients in your tacos, and to buy local versions of these ingredients. Here are some delicious ideas for Mexican street tacos, especially made with queso fresco and cheese: -Tacos al Tío Toto -Tacos con Salsa de Queso & Queso de Cabeza -Tacos de Carne Asada con Queso Bajo Maso -Tacos de Carne Asada con Queso de Cabeza -Tacos en Chile -Tacos de Carne Asada en Español -Tacos con Queso de Cabeza & Queso Cortado -Tacos con Queso Fresco & Queso de Cabeza -Tacos de Jitomate & Queso Chihuahua -Tacos de Queso de Cabeza & Queso de Cebolla -Tacos de Queso de Cabeza & Queso de Patatas -Tacos con Queso Cortado Por Queso -Tacos de Queso Cortado de Cabrales -Tacos Rancheros -Tacos en Chihuahua -Tacos de Queso Cortado de Cabrales -T
GPT-4 Judgment	Assistant A's response is better because it directly addresses the issue of making tortillas less dry by providing clear, relevant instructions. It explains how to marinate meats and prepare fresh tortillas with water, salt, and lime, then suggests cooking them in a bit of oil for a crispy texture, which helps prevent dryness. In contrast, Assistant B's response goes off-topic with unrelated taco recipes and complicated instructions, making it less focused on solving the user's specific problem.

Figure 17: Sample responses to a prompt from the test set of the H4 Stack Exchange Preferences Dataset. GPT-4 chooses MallowsPO-θ over DPO. Both MallowsPO-θ and DPO are trained for $\beta = 0.1$ on Anthropic-HH dataset.

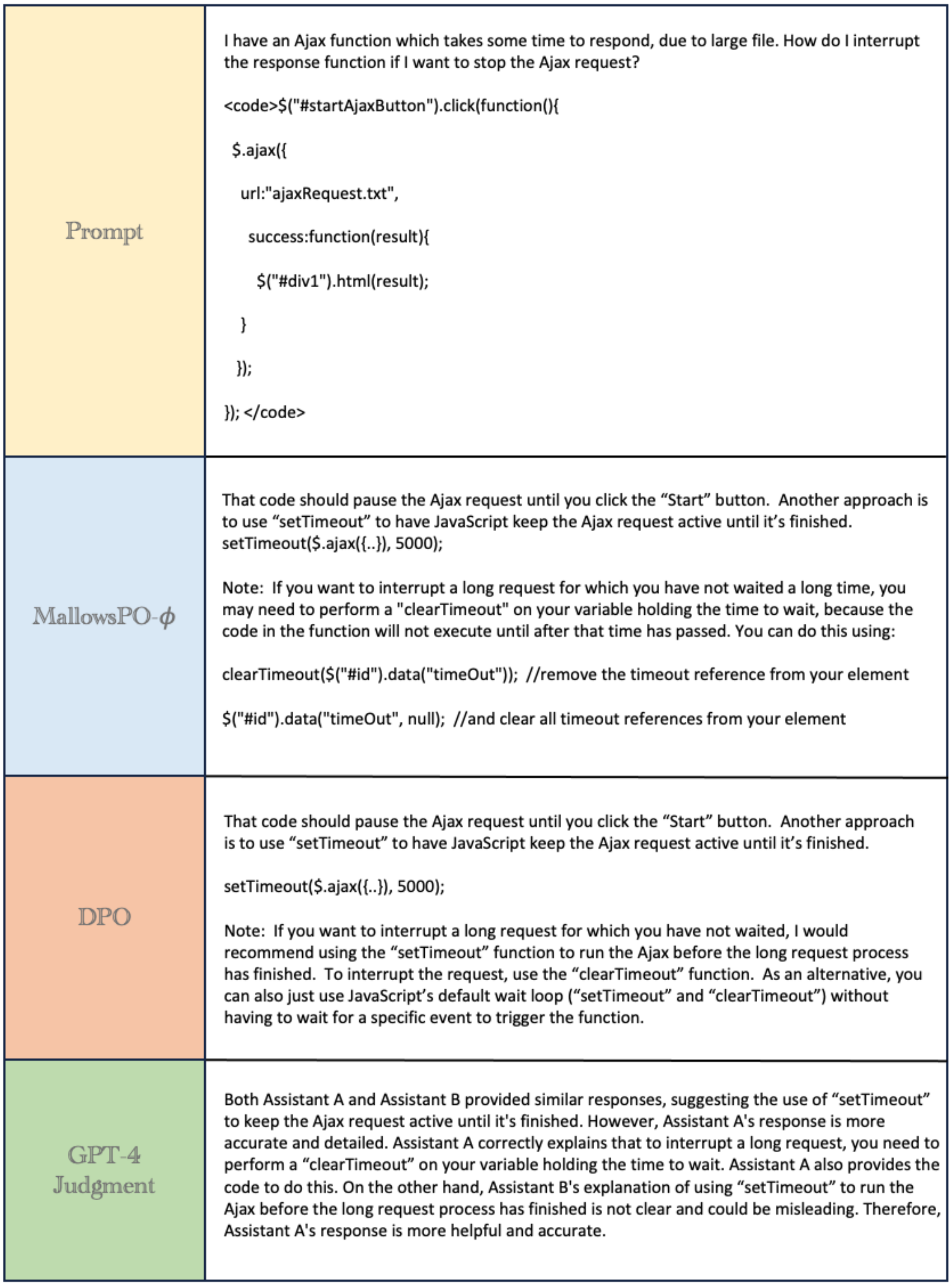


Prompt	Human: I'm attempting to get a (in my opinion) very simple scenario working with Google Tag Manager and Analytics, and running into problems.
	Now we would like to track events using custom JavaScript, via the . For example:
	```
	\$('a').click(function() {
	  dataLayer.push({'event': 'linkclickedtest'});
	});
	```
	
	This just isn't working, at all. If I watch Chrome's console/network tabs, no data seems to get sent to Google. I've also installed for Chrome and it doesn't show the event firing either. I must be missing something simple. Any guidance would be appreciated. Thanks! Assistant:

MallowsPO-θ	There are a few possibilities here:
	
	- You are defining the event before the link has even been added to the DOM, so your code will never run.
	
	- The code you are running also won't run in that scenario because the event has not yet been added, so the code you have won't get executed.
	
	- You have to wait for the JavaScript to run before the code can be evaluated, so the first part of the code above is going to run after the code and the link you've just added to the DOM, but before the event has been loaded. If you change the code in your question to be more along these lines, it'll work:
	<script>
	// Wait for the JavaScript to load, before executing the code below.
	setTimeout(function(){
	document.querySelector('a').onclick = function() {
	dataLayer.push({'event': 'linkclickedtest'});
	}}, 0);

DPO	There are a few possibilities here: You're using the latest version of Google Tag Manager, and you haven't configured it correctly. You're using custom JavaScript in a page other than the one being tracked. You're using analytics tags in other parts of your code, which are interfering with your tracking code. For instance, if you were using <script> tags at the start of <body>, those would interfere with Google Tag Manager script. To resolve issue 1 here, you'll need to update the version of Google Tag Manager you're using. In the Google Tag Manager Settings page, click "Download \& Install" to download and install the latest version of the Google Tag Manager. You'll also need to make sure it's configured correctly in your environment, and also set it up correctly in the header of the page you're tracking. If you're using a different version of Google Tag Manager than the one you downloaded, you'll need to install Google Tag Manager version 1.5 or higher.
GPT-4 Judgment	Both assistants provide potential reasons why the user's code might not be working and offer solutions. However, Assistant A's response is more relevant and directly addresses the user's question. Assistant A suggests that the user's event might be defined before the link has been added to the DOM, and provides a code snippet to resolve this issue. On the other hand, Assistant B's response is less specific and doesn't provide a direct solution to the user's problem. Therefore, Assistant A's response is more helpful and accurate.

Figure 18: Sample responses to a prompt from the test set of the H4 Stack Exchange Preferences Dataset. GPT-4 chooses MallowsPO-ϕ over DPO. Both MallowsPO-ϕ and DPO are trained for $\beta = 0.5$ on Anthropic-HH dataset.



Prompt	I have an Ajax function which takes some time to respond, due to large file. How do I interrupt the response function if I want to stop the Ajax request?
	```
	$("#startAjaxButton").click(function(){
	$.ajax({
	url:"ajaxRequest.txt",
	success:function(result){
	$("#div1").html(result);
	}
	});
	});
	```

MallowsPO-ϕ	That code should pause the Ajax request until you click the “Start” button. Another approach is to use “setTimeout” to have JavaScript keep the Ajax request active until it’s finished. setTimeout($.ajax({..}), 5000);
	Note: If you want to interrupt a long request for which you have not waited a long time, you may need to perform a "clearTimeout" on your variable holding the time to wait, because the code in the function will not execute until after that time has passed. You can do this using:
	clearTimeout($("#id").data("timeOut")); //remove the timeout reference from your element
	$("#id").data("timeOut", null); //and clear all timeout references from your element

DPO	That code should pause the Ajax request until you click the “Start” button. Another approach is to use “setTimeout” to have JavaScript keep the Ajax request active until it’s finished. setTimeout($.ajax({..}), 5000); Note: If you want to interrupt a long request for which you have not waited, I would recommend using the “setTimeout” function to run the Ajax before the long request process has finished. To interrupt the request, use the “clearTimeout” function. As an alternative, you can also just use JavaScript’s default wait loop (“setTimeout” and “clearTimeout”) without having to wait for a specific event to trigger the function.
GPT-4 Judgment	Both Assistant A and Assistant B provided similar responses, suggesting the use of “setTimeout” to keep the Ajax request active until it's finished. However, Assistant A's response is more accurate and detailed. Assistant A correctly explains that to interrupt a long request, you need to perform a “clearTimeout” on your variable holding the time to wait. Assistant A also provides the code to do this. On the other hand, Assistant B's explanation of using “setTimeout” to run the Ajax before the long request process has finished is not clear and could be misleading. Therefore, Assistant A's response is more helpful and accurate.