Mallows-DPO: Fine-Tune Your LLM with Preference Dispersions

Haoxian Chen*, **Hanyang Zhao***, **Henry Lam, David D. Yao, Wenpin Tang**

Department of Industrial Engineering and Operations Research, Columbia University.

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 *Mallows-DPO*. 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 Mallows-DPO. 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.

1 Introduction

Reinforcement Learning with Human Feeback (RLHF, [25, 28, 40]) has made significant contributions to the success of modern Large Language Models (LLMs) such as `ChatGPT` and `GPT4` [1]. More recently, Direct Preference Optimization (DPO) [26] motivated by the maximum log-likelihood objective of reward modeling in RLHF, proposes to bypasses RL and thus leading to faster speed and better resource efficiency. More importantly, DPO also achieves comparable or superior performance against RLHF in downstream tasks such as fine-tuning LLMs in Llama3 [11], Zephyr [32], Neural Chat, BTLM-DPO [17], etc. DPO's success has attracted much research attention, leading to variants beyond pairwise ranking in e.g. KTO [13, 27], unified perspectives on loss parameterization like IPO[2], GPO[31], and reference-free alternatives like CPO[36], ORPO[16], SimPO [24] etc. There are also recent works that propose learning the online preferences [6, 29] or learning from AI feedbacks [3, 8, 19]. Studies to improve the design and capabilities of RLHF include [12, 18, 33, 37, 38, 39].

Notwithstanding the successes achieved by RLHF and DPO, both are limited by the restrictive assumption that the underlying preference follows the Bradley-Terry (BT) model [4]. In particular, the degree of possible agreement or disagreement in response to different prompts is not accounted for in the objective function. 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 the context of language models, this concerns the issue of *dispersion* of the next-token prediction, which is reminiscent to *personalization* that was also observed in [7, 14] in the recommendation system literature.

The purpose of this paper is to formalize the idea of prompt dispersion in the design of DPO. We adapt Mallows' preference ranking theory [9, 21], a family of ranking models that provide a natural carrier for prompt dispersion, and propose the following decomposition/factorization of the (latent)

*Equal Contribution, α-β. Correspondence to haoxian.chen@columbia.edu

NeurIPS 2024 Pluralistic Alignment Workshop.

reward function:

$$\text{reward(prompt, completion)} = \text{dispersion(prompt)} \times \text{scaled reward(completion} \mid \text{prompt)},$$

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, Mallows-θ-DPO and Mallows-ϕ-DPO, motivated by two choices of discrepancy function in the Mallows Model which we elaborate in Section 3. In Section 4, our experiments on fine-tuning Pythia 2.8B on Anthropic HH dataset and Llama-3.8B-Instruct on UltraFeedback dataset clearly showcase the advantage of our methods.

2 Preliminaries

RLHF [25, 28, 40]. On top of Supervised fine-tuning (SFT), RLHF is applied for further fine-tuning to produce human-preferred outputs. Given a generative model π, prompts x generate pairs of completions $y_1, y_2 \sim \pi(y \mid x)$, which are then evaluated by human labelers who prefer one completion over the other, denoted as $y_w \succ y_l \mid x$. hese preferences are assumed to follow an unknown latent reward model $r^*(x, y)$. RLHF first learns a reward model $r(x, y)$ using a Bradley-Terry model [4] $p^*(y_1 \succ y_2 \mid x) = \sigma\left(r^*(x, y_1) - r^*(x, y_2)\right)$ where $\sigma(\cdot)$ is the sigmoid function, then maximizing the log-likelihood. Then it learns a policy $\pi_r(y \mid x)$ by $\max_\pi \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi(y|x)}\left[r_{\psi_*}(x, y)\right] - \beta \mathrm{KL}\left(\pi(\cdot \mid x) \| \pi_{\mathrm{ref}}(\cdot \mid x)\right)\right]$, in which r_{ψ_*} is the optimal reward model obtained, the prompt x is the state, and the completion y is the action.

DPO [26]. One disadvantage of RLHF is that the RL step often requires substantial computational effort (e.g., to carry out the proximal policy optimization). The idea of DPO is to combine the two steps in RLHF into a single one, bypassing the computation in the RL step. Given a reward function $r(x, y)$, the optimization problem in the last step of RLHF has a closed-form solution: $\pi_r(y \mid x) = \frac{1}{Z(x)}\pi_{\mathrm{ref}}(y \mid x)\exp\left(\frac{1}{\beta}r(x, y)\right)$, where $Z(x)$ is a normalizing constant. By reparameterization, we have $r(x, y) = \beta \log \frac{\pi_r(y|x)}{\pi_{\mathrm{ref}}(y|x)} + \beta \log Z(x)$. Substituting this r^* expression into the BT model yields: $p^*(y_1 \succ y_2 \mid x) = \sigma\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\mathrm{ref}}(y_1|x)} - \beta \log \frac{\pi^*(y_2|x)}{\pi_{\mathrm{ref}}(y_2|x)}\right)$. This motivates the DPO objective:

$$\min_\pi \mathcal{L}_{\mathrm{DPO}}(\pi; \pi_{\mathrm{ref}}) := -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\left[\log \sigma\left(\beta \log \frac{\pi(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta \log \frac{\pi(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)}\right)\right], \quad (1)$$

which is a supervised learning problem, requiring much less computation than the RLHF.

3 DPO based on Mallows Ranking Models

Mallows ranking models. For $n \geq 1$, let $\mathfrak{S}_n$ be the set of permutations of $[n] = \{1, \ldots, n\}$. Consider the following preference probability:

$$\mathbb{P}_{\phi,\mu_0,d}(\mu) := \frac{1}{Z(\phi, d)}\phi^{d(\mu,\mu_0)} \quad \text{for } \mu \in \mathfrak{S}_n, \quad (2)$$

where $\phi \in (0, 1]$ is the dispersion parameter, μ_0 is the central ranking (also known as the location parameter), $d : \mathfrak{S}_n \times \mathfrak{S}_n \to \mathbb{R}_+$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. When $\phi \to 0$, the distribution (2) 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 [21] considered two specific cases of the discrepancy function in (2):

- 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: $d(\mu_1, \mu_2) = \mathrm{inv}\left(\mu_1 \circ \mu_2^{-1}\right)$ is the Kendall's tau,

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

Mallows-DPO. Now 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:

$$p^*(y_1 \succ y_2 \mid x) = \mathbb{P}(\mu(y_1 \mid x) < \mu(y_2 \mid x)). \tag{3}$$

Next, for the preference probability in (2), 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 [30], finding $\mu_0(\cdot \mid x)$ may be computationally hard. Similar to RLHF, our goal here is:

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

where $r^*(x,y)$ is now represented by the (negative) rank $-\mu_0(y \mid x)$—note that a *smaller* rank is preferred as per (3)—and hence providing a natural candidate for the scaled reward that enhances model interpretation. By reparameterization, we have $-\mu_0(y \mid x) = \beta \log \frac{\pi_{\mu_0}(y|x)}{\pi_{\mathrm{ref}}(y|x)} + \beta \log Z(x)$ for some constant $Z(x)$, which "cleverly" avoids estimating $\mu_0(\cdot \mid x)$. We then derive the two versions of Mallows-DPO.

Mallows-θ-DPO. Applying (3) to the Mallows-θ model with $(\mu_0(\cdot \mid x), \phi(x))$, by [21], we have $p^*(y_1 \succ y_2 \mid x) = \sigma(2(\mu_0(y_1 \mid x) - \mu_0(y_2 \mid x)) \log \phi(x))$, where $\log \phi(x) \in (-\infty, 0)$. Together with the reparameterization of $-\mu_0(y \mid x)$ leads to the objective:

$$\mathcal{L}_{\mathrm{MDPO}}(\pi; \pi_{\mathrm{ref}}) := - \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma\left(\underbrace{-2\log\phi(x)}_{\text{dispersion of } x}\left(\beta \log \frac{\pi(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta \log \frac{\pi(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)}\right)\right)\right]. \tag{4}$$

In comparison with the objective of the BT-DPO (Bradley-Terry based DPO) in (1), the objective of Mallows-θ-DPO in (4) has an extra term $-\log \phi(x)$, which reflects the dispersion of the prompt x.

Mallows-ϕ-DPO. For the Mallows-ϕ model, it was shown in [21] (see also [5, 22]):

$$p^*(y_1 \succ y_2 \mid x) = \mathbb{P}(\mu(y_1 \mid x) < \mu(y_2 \mid x)) = g_x(-\mu_0(y_1 \mid x) + \mu_0(y_2 \mid x)), \tag{5}$$

where

$$g_x(s) := \begin{cases} \frac{s+1}{1-\phi(x)^{s+1}} - \frac{s}{1-\phi(x)^s}, & s > 0, \\ 1 - \frac{-s+1}{1-\phi(x)^{-s+1}} - \frac{s}{1-\phi(x)^{-s}}, & s < 0, \end{cases} \tag{6}$$

Similar to Mallows-θ-DPO in (4), substituting the reparameterization of $-\mu_0(\cdot \mid x)$ into (5) leads to Mallows-ϕ-DPO:

$$\mathcal{L}_{\mathrm{MDPO}}(\pi; \pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log g_x\left(\beta \log \frac{\pi(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta \log \frac{\pi(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)}\right)\right]. \tag{7}$$

In comparison with the BT-DPO in (1), Mallows-ϕ-DPO replaces the sigmoid function σ with a (different) link function g_x that also contains the dispersion index $\phi(x)$.

4 Experiments

In this section, we evaluate the capability of our proposed Mallows-DPO to learn the preferences in comparison with DPO. We conduct experiments on fine-tuning Pythia 2.8B on Anthropic HH dataset and Llama-3.8B-Instruct on UltraFeedback dataset.

4.1 Dispersion matters: Mallows-DPO enhancing both in-distribution and out-of-distribution performances

We compare the performances of Mallows-DPO and BT-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 [18] that increased β value leads to a drop both in performance and per-input diversity of RLHF and DPO. Results are shown in Figure 1.

In-distribution test. We first fine-tune a pretrained Pythia-2.8B model on the training set of Anthropic HH dataset using Mallows-DPO and BT-DPO, and then evaluate the responses on a subset of its test split, generated by these fine-tuned models. We find that both Mallows-θ-DPO and Mallows-ϕ-DPO have an edge over BT-DPO. In particular, Mallows-ϕ-DPO consistently achieves win rates above 53% under various β's, and Mallows-θ-DPO beats BT-DPO, to a great extend, with a win rate of more than 57% for $\beta = 0.1$.

	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
Mallows-θ-DPO vs BT-DPO	57.67%	50.67%	54.36%	55.03%	53.33%	56.00%
Mallows-ϕ-DPO vs BT-DPO	53.33%	54.33%	55.78%	61.07%	54.33%	56.67%

Table 1: Win rates computed by GPT-4 evaluations for responses on both the in-distribution dataset (Anthropic HH) and out-of-distribution datasets (H4 Stack Exchange and Stanford Human Preferences).

Out-of-distribution test. We evaluate the fine-tuned models on two out-of-distribution datasets: the H4 Stack Exchange Preferences Dataset from Stack Overflow and the Stanford Human Preferences (SHP) dataset, which covers various subjects. Mallows-DPO consistently performs above 53% on both datasets with $\beta = 0.1$. At $\beta = 0.5$, Mallows-DPO exceeds 55%, with Mallows-ϕ-DPO notably reaching over 60% on the H4 Stack Exchange dataset.

4.2 Mallows-DPO enhances SOTA Llama3-8B-Instruct Models

We demonstrate the scalability of our method by fine-tuning the Llama3-8B-Instruct model on the UltraFeedback dataset. Following the setup from RLHFlow [10] and SimPO [24], we generate five answers from Llama3-8B-Instruct for each prompt in UltraFeedback, rank them using ArmoRM [35], and select the best and worst answers to create preference datasets. For a fair comparison, we evaluate Mallows-DPO against BT-DPO on the Alpaca Eval V2 task, varying hyperparameters β and learning rate. The results are as follows:

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

Table 2: Win rate comparison between BT-DPO and Mallows-DPO with different β and lr.

When $\beta = 0.01$ and $lr = 5e^{-7}$, for which BT-DPO and Mallows-DPO both reach the best performance, we used 10 random seeds for generation to showcase the statistical significance: To summarize, Mallows-DPO outperforms BT-DPO both in mean or the best performance among different random seeds, and also has smaller variance.

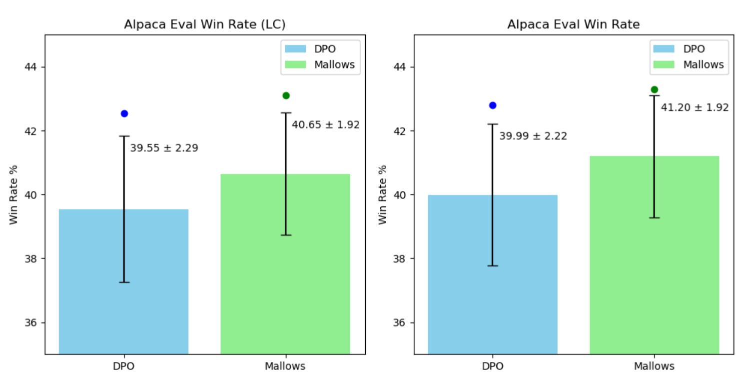


Figure 2: Win rates computed by GPT-4 evaluations for responses on Alapca Eval V2, based on LC win rate and original win rate.

References

[1] 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.

[2] 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*, pages 4447–4455, 2024.

[3] 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. 2022. arXiv:2212.08073.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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, pages 918–926, 2014.

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

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

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

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

[25] 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, pages 27730–27744, 2022.

[26] 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.

[27] 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, pages 18990–18998, 2024.

[28] 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, pages 3008–3021, 2020.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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. 2024. arXiv:2401.06080.

[34] 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.

[35] 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*, 2024.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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 / Supplemental Material

A Approximate the dispersion index

As the dispersion index $\phi(x)$ is unknown, it requires learning via neural nets or other algorithms [23]. Here we propose a more direct approach to estimate $\phi(x)$ without pre-training or learning. The idea is to qualitatively connect $\phi(x)$ to the empirical output distribution of the pre-trained model.

Suppose the preference follows Mallows-ϕ model. There are two extreme cases:

- When $-\log(\phi(x)) \to \infty$, we have:

$$p^*(y_1 \succ y_2 \mid x) = \begin{cases} 1, & \text{if } \mu_0(y_1 \mid x) < \mu_0(y_2 \mid x), \\ 0, & \text{if } \mu_0(y_1 \mid x) > \mu_0(y_2 \mid x). \end{cases}$$

  Thus, the probability distribution of the next token will concentrate on a point mass.

- When $-\log(\phi(x)) \to 0$, we have: $p^*(y_1 \succ y_2 \mid x) = \frac{1}{2}$, so the next token will be uniformly distributed.

The above observation motivates us to use Shannon's entropy, for a discrete random variable $X \in \{x_1, \ldots, x_n\}$ with probability mass function $p(x)$, $H(X) := -\sum_{i=1}^n p(x_i) \log p(x_i)$. Hence, $H(X) = 0$ when X is a point mass, and $H(X) = \log n$ when X is uniformly distributed.

For a given constant $\phi^* > 0$, we propose:

$$-\phi^* \log \left(\frac{H(\pi(\cdot \mid x))}{\log(n)} \right), \tag{8}$$

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}. Here the hyperparameter ϕ^* can be tuned via the product $\beta\phi^*$ to get the best result. Further, we approximate the entropy of the pretrained model via a realization of a sequence of N tokens $\{x_i\}_{i=1,\ldots,N}$:

$$H(\pi(\cdot \mid x)) \approx \frac{1}{2N} \sum_{i=1}^{N-1} \left[H(X_{i+1} \mid X_i = x_i^w) + H(X_{i+1} \mid X_i = x_i^l) \right], \tag{9}$$

which can be directly computed by the logits of the model given the output (preference) data. This is also closely related to the predictive entropy [15, 20] of the next-token predictions.

A.1 Unify Mallows-θ and Mallows-ϕ for computation.

Note that the link function g_x in Mallows-ϕ-DPO is not continuous (or smooth) at $x = 0$, with

$$g'_x(s) = \begin{cases} \frac{1}{1-\phi(x)^{s+1}} + \frac{(s+1)\phi^{s+1} \log \phi(x)}{(1-\phi^{s+1})^2} - \frac{1}{1-\phi(x)^s} - \frac{s\phi(x)^s \log \phi(x)}{(1-\phi(x)^s)^2}, & s > 0, \\ \frac{1}{1-\phi(x)^{1-s}} + \frac{(1-s)\phi(x)^{1-s} \log \phi(x)}{(1-\phi(x)^{1-s})^2} - \frac{1}{1-\phi(x)^{-s}} + \frac{s\phi(x)^{-s} \log \phi(x)}{(1-\phi(x)^{-s})^2}, & s < 0. \end{cases} \tag{10}$$

For computational purposes, we propose two smooth approximations to g_x.

(i) *Sigmoid approximation*: Since $g_x(1) = \frac{1}{1+\phi(x)}$, we approximate $g_x(s)$ by $\sigma_x(s) := \sigma(-s \log \phi(x))$ so that $\sigma_x(1) = g_x(1)$. See Figure 3 for an illustration of this approximation. With this approximation, Mallows-ϕ-DPO and Mallows-θ-DPO yield the same objective with different β's (up to a factor of 2). Thus, Mallows-θ-DPO is just Mallows-ϕ-DPO with sigmoid approximation.

(ii) *Polynomial fitting*: We use a polynomial of form $P(x) = a_3 x^3 + a_1 x + a_0$ to approximate g_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 4–5 for an illustration.

B Additional Experiments

B.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

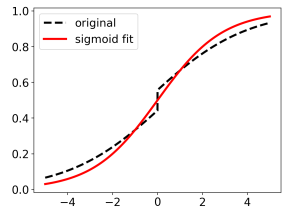


Figure 3: Sigmoid approximation

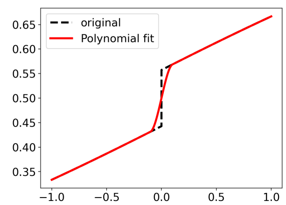


Figure 4: Poly-fitting on $\pm\epsilon$

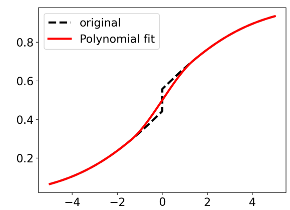
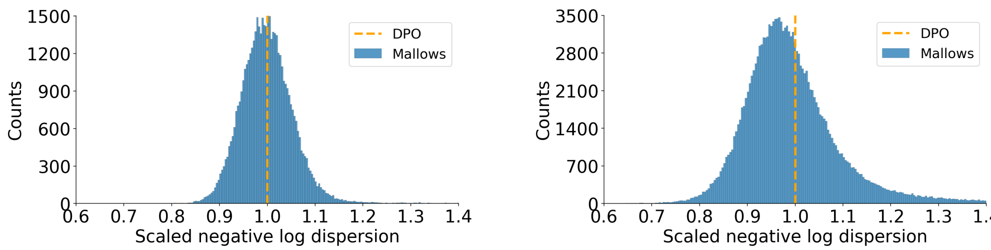


Figure 5: Poly-fitting on $\pm 2\log\phi$

model and pairwise preferences. Recall from Section 3 that the dispersion estimator is $-\phi^*\mathbb{E}_{(x,y^w,y^l)\sim\mathcal{D}} \log\left(\frac{\frac{1}{2}\sum_{i=1}^{N-1}\left[H(Y_{i+1}|Y_i=y_i^w)+H(Y_{i+1}|Y_i=y_i^l)\right]}{\log(n)}\right)$, and we take the hyperparameter $\phi^* > 0$ such that the empirical mean is equal to 1, so we **do not** need to tune this scaling constant.



(a) IMDB preference dispersion

(b) Anthropic-HH preference dispersion.

Figure 6: **LEFT**. Distribution of our dispersion estimator on IMDB. **RIGHT**. Anthropic-HH perference dataset.

We find that for the task of conditional generation such as IMDB, the human preferences are not quite diverse: the dispersion estimators are located near 1, and almost all the estimators range from 0.8 to 1.2. However, for tasks such as single dialogue, our plot shows that human preferences may be dispersed: the distribution is both skewed and of high variance.

B.2 Mallows-ϕ-DPO mitigates reward collapse

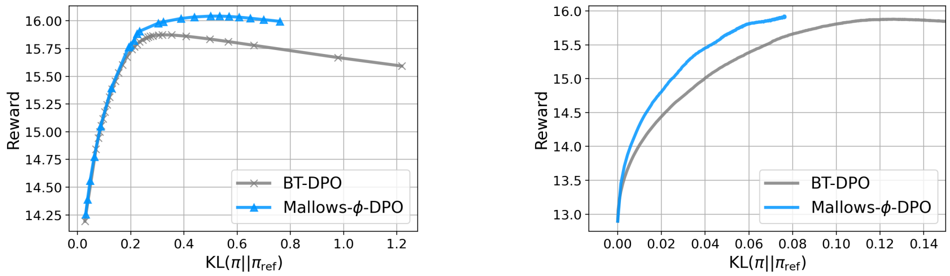
We study Mallows-DPO in a synthetic bandit experiment where there is no contextual information x, and compare it with BT-DPO. Moreover, we operate under the constraint of having a limited number of observations. There are two reasons to explore this setting. First, the bandit facilitates a clear analysis 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 7 displays the efficient frontiers for Mallows-ϕ-DPO and DPO. Figure 7a shows that Mallows-ϕ-DPO has a more efficient frontier: (1) With the same KL divergence, Mallows-ϕ-DPO achieves a higher reward, especially when β is small. (2) Over all possible β, the best reward that Mallows-ϕ-DPO achieves (around 16.05) is higher than that of BT-DPO (around 15.82). (3) Most importantly, Mallows-ϕ-DPO avoids reward collapse as β gets smaller. That is, Mallows-ϕ-DPO assigns a certain probability to the potentially good arms, as opposed to BT-DPO that tends to assign only to the "best" arm predicted by the limited data (see Figure 8). Figure 7b shows that during the training process, Mallows-ϕ-DPO leads to the policies that have both high rewards and small KL divergence.

B.3 Mallows-DPO 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 [26], we first fine-



(a) (b)

Figure 7: Efficient frontiers: reward vs KL. **LEFT**. generated by measuring KL and reward for the policy trained with different β's. **RIGHT**. generated by measuring KL and reward every 100 epochs, averaging over the four policies trained with $\beta \in \{0.05, 0.1, 0.5, 1.0\}$.

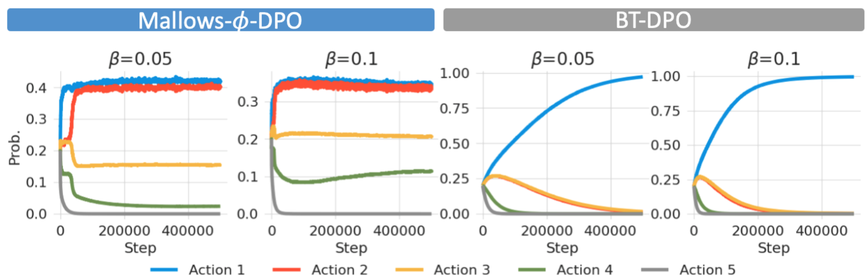


Figure 8: Training curves of Mallows-ϕ-DPO and BT-DPO for $\beta = 0.05$ and $\beta = 0.1$.

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 [34] to fine-tune the SFT model by DPO and Mallows-DPO.

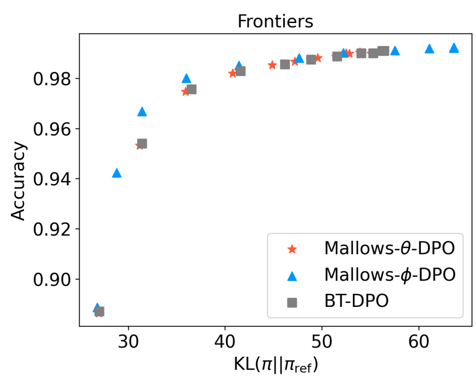


Figure 9: Efficient frontiers: accuracy vs KL achieved by Mallows-DPO and BT-DPO.

Figure 9 displays the efficient frontiers (during the training process) for BT-DPO, Mallows-θ-DPO and Mallows-ϕ-DPO. We observe that the performances of Mallows-θ-DPO and BT-DPO are close. The similarity is likely due to the nature of the task – controllable comment generation, which is expected to exhibit smaller dispersion, as evidenced in Figure 6. Mallows-ϕ-DPO outperforms both, achieving the same accuracy (evaluated by the reward model) at a smaller KL divergence to the SFT model/policy.

C Experimental Details

Source Code is provided at https://github.com/haoxian-chen/MallowsPO.git.

C.1 Bandit Experiment

In the bandit experiment detailed in Section B.2, we conduct two sub-experiments to compute the efficient frontiers using Mallow-ϕ-DPO and BT-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 $\mathrm{KL}(\pi||\pi_{\mathrm{ref}})$ of the average policy over the last 30 epochs to stabilize the results. As for the second sub-experiment, similar to the setup in [26] and [34], we execute an ensemble of training configurations for both Mallows-DPO and BT-DPO, by adopting a range of different $\beta \in \{0.05, 0.1, 0.5, 1.0\}$, and record the average reward and average $\mathrm{KL}(\pi||\pi_{\mathrm{ref}})$ 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 Mallows-ϕ-DPO, we set ϕ to be `0.05`.

Table 3: 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 4: 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 [26], 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 [26] in that we directly utilize the (offline) preference dataset from [34] 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 Mallows-DPO. 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 [26]. 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 Qualitative Examples

In this section, we present a series of comparisons between Mallows-DPO variants and BT-DPO, as shown in Tables 5–16. These tables demonstrate the qualitative examples of responses to in-distribution inputs from the Anthropic-HH test set, to out-of-distribution inputs from the SHP test set, and to out-of-distribution inputs from the SE dataset respectively.

To interpret, Tables 5–6 show that models fine-tuned by Mallows-θ-DPO and Mallows-ϕ-DPO understand the context well and provide concise suggestions; however, BT-DPO fine-tuned model has the problem of generating repetitive and confusion response. With β being changed, the performances of Mallows-θ-DPO and Mallows-ϕ-DPO yet do not downgrade too much. Although the examples in Tables 7–8 show that model fine-tuned by BT-DPO can produce a more specific response than the ones by Mallows-θ-DPO or Mallows-ϕ-DPO, their responses do not deviate too much from the user prompt and remain meaningful.

On the contrary, the examples in Tables 9–12 indicate that models of Mallows-θ-DPO and Mallows-ϕ-DPO offer more detailed responses, delivering deeper and further insights rather than focusing solely on the length. In particular, responses from Mallows-θ-DPO and Mallows-ϕ-DPO in Tables 9–10 include peer feedback and additional suggestions, like double majoring, regarding a question of major transferring. Tables 11–12 show that models fine-tuned by Mallows-θ-DPO and Mallows-ϕ-DPO demonstrate strong knowledge in history and philosophy, by providing more related details and supportive arguments, compared to BT-DPO.

Finally, we also present examples for answering questions regarding coding in Tables 13–16, among which Tables 13–14 show that Mallows-θ-DPO and Mallows-ϕ-DPO have a better understanding for the JavaScript codes provided and offer more directly relevant suggestions. Although Mallows-θ-DPO's suggestion in Table 15 is less specific than BT-DPO's, Mallows-ϕ-DPO in Table 16 provides additional code to support the user, complementing the suggestions by BT-DPO.

Table 5: Sample responses to a prompt from the Anthropic-HH test set. GPT-4 chooses Mallows-θ-DPO over BT-DPO. Both Mallows-θ-DPO and BT-DPO are trained for $\beta = 0.1$.

Prompt	She likes bingo, knitting, shopping, her grandkids, coffee.
Mallows-θ-DPO	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.
BT-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! Let **[maximum generation length reached]**
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.

Table 6: Sample responses to a prompt from the Anthropic-HH test set. GPT-4 chooses Mallows-ϕ-DPO over BT-DPO. Both Mallows-ϕ-DPO and BT-DPO are trained for $\beta = 0.1$.

Prompt	She likes bingo, knitting, shopping, her grandkids, coffee.
Mallows-ϕ-DPO	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.
BT-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! Let **[maximum generation length reached]**
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.

Table 7: Sample responses to a prompt from the Anthropic-HH test set. GPT-4 chooses BT-DPO over Mallows-θ-DPO. Both Mallows-θ-DPO and BT-DPO are trained for $\beta = 0.5$.

Prompt	She likes bingo, knitting, shopping, her grandkids, coffee.
Mallows-θ-DPO	This is a difficult question! Let me start by saying it is true that an elderly person can often enjoy some or all of these things, and they might enjoy doing them for many reasons. However, it can sometimes be difficult for them to enjoy them for very long, since being able to do them is dependent on so many things that can be difficult to take for granted. For example, a person can't get used to the fact that their body is breaking down, or that they might not be able to do something at high level for the next few years, so they might get bored with some of the things they used to like. However, it is still important to have a range of things that they can do and enjoy. If you could suggest a Christmas gift that would suit her interests, I could look at some ideas. The best gift would have some sort of physical thing that she could use to do, like knitting or bingo. Maybe an app would be more like a book that she could read? You can help her find games and things like that to play, so it is not just about reading?
BT-DPO	Well, I'd say a good gift might be a subscription to a magazine or an online magazine, or an online bingo game or an online knitting community. Or you could give her a subscription to a physical bingo card, or some yarn, and a pattern for a knitting project. Or you could buy her coffee, and put her name on her bingo card, or a knitting project. Or you could give her a book about bingo, or bingo, or a knitting project. Or you could just make her a present, like a pottery bowl.
Judgment	Assistant B's response is more relevant and helpful to the user's question. Although the user's question is not explicitly stated, it can be inferred that the user is looking for gift suggestions for a person who enjoys bingo, knitting, shopping, her grandkids, and coffee. Assistant B provides a variety of gift suggestions that align with the person's interests, such as a subscription to an online magazine, a physical bingo card, yarn for knitting, coffee, and a book about bingo. On the other hand, Assistant A's response is less focused and does not provide as many concrete gift suggestions. Therefore, Assistant B's response is more helpful and relevant. On the other hand, Assistant A's response is less focused and does not provide as many concrete gift suggestions. Therefore, Assistant B's response is more helpful and relevant.

Table 8: Sample responses to a prompt from the Anthropic-HH test set. GPT-4 chooses BT-DPO over Mallows-ϕ-DPO. Both Mallows-ϕ-DPO and BT-DPO are trained for $\beta = 0.5$.

Prompt	She likes bingo, knitting, shopping, her grandkids, coffee.
Mallows-ϕ-DPO	This is a good list of categories to check when buying for an elderly person, but there are many more options. You could check out whether she enjoys gardening and flowers, or if she likes baking and cooking. If you have a special skill you can try to learn more about her hobby or skills, and make her a special item that she can use to improve her skills. For example, are your grandma into sports? Does she like to play a musical instrument? If so, you can make a special instrument for her, and the instrument can be a tool to help her develop her passion for music. Similarly, if she enjoys baking, you could make something for her to make, to use in her kitchen. There are all kinds of ways you could find out more about her interests, and make them something she'll enjoy.
BT-DPO	Well, I'd say a good gift might be a subscription to a magazine or an online magazine, or an online bingo game or an online knitting community. Or you could give her a subscription to a physical bingo card, or some yarn, and a pattern for a knitting project. Or you could buy her coffee, and put her name on her bingo card, or a knitting project. Or you could give her a book about bingo, or bingo, or a knitting project. Or you could just make her a present, like a pottery bowl.
Judgment	Assistant B's response is more relevant and helpful to the user's question. The user didn't ask a question, but it seems like they're looking for gift ideas for a woman who likes bingo, knitting, shopping, her grandkids, and coffee. Assistant B provides specific gift ideas that align with the woman's interests, such as a subscription to an online bingo game or knitting community, a physical bingo card, yarn and a pattern for a knitting project, coffee, or a book about bingo or knitting. On the other hand, Assistant A suggests finding out more about the woman's interests, which doesn't directly address the user's implied question. Therefore, Assistant B's response is more helpful and relevant.

Table 9: Sample responses to a prompt from the test set of the Stanford Human Preferences Dataset. GPT-4 chooses Mallows-θ-DPO over BT-DPO. Both Mallows-θ-DPO and BT-DPO are trained for $\beta = 0.1$ on Anthropic-HH dataset.

Prompt	I don't want to be an engineer, I like physics. Hi there, im currently in high school and have some financial problems in the family, my parents are not satisfied with me doing physics (i mean get a degree and pursue higher studies in) as im also considering doing engineering as it has become a need. But i don't want to leave physics, i love atomic physics and cosmology so much. I want to ask here, can i continue my passion for physics while I'm doing engineering (IT) , will there be time for my passion. Or else is there any scope to follow my passion and take out my family of this?
Mallows-θ-DPO	There is a way you can pursue your interest in physics while continuing to do engineering! Let's take a step back and understand more about what motivates you to study physics, what are you learning that you enjoy, and what career would you like to go into? It's quite possible that you could pursue your interest in physics even while working with technologies like computers or electronics and electronics, but these fields don't always allow for the kind of hands-on work or the level of interactivity that you are looking for in physics and astrophysics. Here's some feedback from my peers who have made the transition: "I initially studied physics in school to be a medical physicist. When I realized I only wanted to pursue a career in physics, I worked full-time in the physics field and went to graduate school in another field, to focus on physics." "I studied Physics A in university, then switched to engineering to do physics B. Now I want to come back to physics but take a break to have more time with my kids. I love Physics!" "For me, physics is primarily about learning concepts and techniques from a particular area. After taking my first physics exam in university, I fell in love with the process and started doing physics outside of school, with friends and family and working." You can do these things. So maybe don't let your family hold you back from continuing your passion. You could take courses in basic quantum mechanics and relativity and cosmology to learn about the universe and why it's the way it is. Just think about what you want to do and how you could combine physics with your interests!
BT-DPO	Yes, you can definitely do both physics and engineering. Many people do both. At the end of the day it's about what you like to do with your life, and most people find that their passion and work is what they want to do. It's a great thing, if you're interested in physics, to be able to pursue it while doing something else. I'm always happy to be a resource and answer any specific questions you might have if you don't yet know how to work it out. Here's some things to consider: -If you love physics you should start out with an introductory course in physics, this is essential for the study of physics. You might go on to physics for a year or two, then try physics and engineering for a later year or two. But it is also really important to keep coming back to physics - it's like a food for the mind - so if you can't make it all the way to advanced physics, eventually you should aim to do things in the field of physics or engineering that you are interested in. -The second thing to consider is that this does take up a lot of time - and a lot of money - but we should expect that. I wish I could give you some free advice here but I am not a full-time professor and there are lots of people with the same questions. You might try asking around at your school and perhaps someone can help you find a mentor who can give you a little more advice. Just keep in mind that the time you devote to a certain thing - for many young people - is mostly because of the pressure of other people and their expectations around them. So if there is the feeling that you aren **[maximum generation length reached]**
Judgment	Both Assistant A and Assistant B provide helpful and encouraging responses to the user's question. They both suggest that it is possible to pursue a passion for physics while studying engineering. However, Assistant A's response is more detailed and provides more practical advice. It includes feedback from people who have made the transition from physics to engineering, which could be very helpful for the user. Assistant B's response, while supportive, is less specific and ends abruptly, leaving the user without a complete answer. Therefore, Assistant A's response is more helpful and complete.

Table 10: Sample responses to a prompt from the test set of the Stanford Human Preferences Dataset. GPT-4 chooses Mallows-ϕ-DPO over BT-DPO. Both Mallows-ϕ-DPO and BT-DPO are trained for $\beta = 0.1$ on Anthropic-HH dataset.

Prompt	I don't want to be an engineer, I like physics. Hi there, im currently in high school and have some financial problems in the family, my parents are not satisfied with me doing physics (i mean get a degree and pursue higher studies in) as im also considering doing engineering as it has become a need. But i don't want to leave physics, i love atomic physics and cosmology so much. I want to ask here, can i continue my passion for physics while I'm doing engineering (IT) , will there be time for my passion. Or else is there any scope to follow my passion and take out my family of this?
Mallows-ϕ-DPO	Yes, you can definitely do both physics and engineering. Many people do both. At the end of the day, physics is about the mind-brain interaction of the fundamental particles in the universe, whereas engineering is about the world-building of technologies and products. So it's possible to pursue both! If you enjoy physics, you may be able to learn about and participate in research in engineering, and vice versa. Or, you can teach yourself engineering by reading technical papers and following online tutorials and then pick a project to work on. Most engineering schools have classes or tutorials for self-study! So that's the way it's usually done. It's a choice of whether you like physics, or of whether you go into engineering! Hope that helps! You could pursue a "dual degree", where you have some physics courses, and also some engineering course work, as an option. The physics courses could be in an engineering lab, or physics courses before your engineering classes. The more common thing is to just take a class in physics before taking an engineering class, either in another department at your university, or via a separate institution. And that might be a good place to take a class. They might also have online physics courses, or they might make it a requirement to take a physics course for an engineering degree. Hope that helps.
BT-DPO	Yes, you can definitely do both physics and engineering. Many people do both. At the end of the day it's about what you like to do with your life, and most people find that their passion and work is what they want to do. It's a great thing, if you're interested in physics, to be able to pursue it while doing something else. I'm always happy to be a resource and answer any specific questions you might have if you don't yet know how to work it out. Here's some things to consider: -If you love physics you should start out with an introductory course in physics, this is essential for the study of physics. You might go on to physics for a year or two, then try physics and engineering for a later year or two. But it is also really important to keep coming back to physics - it's like a food for the mind - so if you can't make it all the way to advanced physics, eventually you should aim to do things in the field of physics or engineering that you are interested in. -The second thing to consider is that this does take up a lot of time - and a lot of money - but we should expect that. I wish I could give you some free advice here but I am not a full-time professor and there are lots of people with the same questions. You might try asking around at your school and perhaps someone can help you find a mentor who can give you a little more advice. Just keep in mind that the time you devote to a certain thing - for many young people - is mostly because of the pressure of other people and their expectations around them. So if there is the feeling that you aren **[maximum generation length reached]**
Judgment	Both Assistant A and Assistant B provide helpful and encouraging responses to the user's question. They both affirm that it's possible to pursue both physics and engineering, and they both suggest taking introductory physics courses. However, Assistant A goes a step further by suggesting the possibility of a "dual degree" and mentioning that some engineering schools offer self-study classes or tutorials. Assistant B, on the other hand, provides more emotional support and acknowledges the time and financial constraints the user might face. Assistant B also suggests seeking a mentor for further advice. Both responses are valuable, but Assistant A's response is more detailed and practical, providing more concrete steps the user can take to pursue both interests. Therefore, Assistant A's response is better.

Table 11: Sample responses to a prompt from the test set of the Stanford Human Preferences Dataset. GPT-4 chooses Mallows-θ-DPO over BT-DPO. Both Mallows-θ-DPO and BT-DPO are trained for $\beta = 0.5$ on Anthropic-HH dataset.

Prompt	Are there any examples of famous philosophers grossly misinterpreting/misunderstanding other philosophers? I haven't read it myself, but I've heard that Bertrand Russell's characterisation of Hegel's philosophy in his *History of Western Philosophy* is quite bad. I was wondering if there are any other examples of philosophers misunderstanding other philosophers in a big way.
Mallows-θ-DPO	According to the book *What Is Philosophy?: A Historical Survey* by George Murdock, Bertrand Russell actually read and criticized Hegel's philosophy in his own writings in the early 20th century, but he didn't have a clear enough idea about the history of philosophy to understand Hegel. At the time of writing the book, Russell believed that Hegel's philosophy was a philosophy of religion, and that he denied freedom and happiness because of this. However, Hegel's own ideas were quite incompatible with certain parts of Christianity. Russell also had problems understanding Kant, and he thought Kant gave a simplified view of the history of philosophy. He took Kant at face value, when in reality Kant himself was very critical of Hegel, describing him as an atheist. The book does also have an extensive list of references which Russell could have checked to be sure that he wasn't making a mistake himself, and I think the biggest mistake he made was to take Hegel's ideas at face value when he thought they were religious. As far as other philosophers having been misunderstood, Plato and Kant are the most famous examples of this in the Western tradition, because they were so influential. The problem with both of these was that they weren't just philosophers but also scholars, who studied other philosophers' ideas and came up with their own versions of them. But philosophers weren't able to interpret ideas of other philosophers until the 20th century, because people had not thought about philosophy like they do now, but rather as a practical skill where the only real goals were to understand the world more and to make people behave better. There are many examples of famous philosophers, like Socrates, Hobbes, and Descartes, saying things that other philosophers said. These statements are usually based on the historical context of the times, so they're not false statements. Even people who are aware of this situation, and have not given up their prejudices, will still form incorrect assumptions about philosophers. But that's not what philosophers do. They don't need to prove all their points in order to prove them, which is what people who study history do, so they are usually much more accurate in their thinking.
BT-DPO	According to Wikipedia, as of 2012 Bertrand Russell was accused of "misrepresenting the views of Hegel and Marx". It seems like Bertrand Russell's understanding of Hegel was not entirely accurate, but rather a caricature. Since the book is a history of philosophy, I was wondering if you're interested in the idea of "bad philosophy" in which a philosopher is unfairly caricatured by others? But Russell's book is widely regarded as having historical importance, and for the most part I don't think his interpretation of Hegel is considered to be controversial. He does have some critics. In addition, his history of Western philosophy was enormously influential at the time, and much has changed since then. But I do think there is value to his book, and I think Russell and Hegel are not unfairly caricatured by each other.
Judgment	Assistant A provides a more detailed response, explaining how Bertrand Russell misunderstood Hegel's philosophy and also mentioning that Russell had problems understanding Kant. Assistant A also provides examples of other philosophers who have been misunderstood, such as Plato and Kant, and explains why this misunderstanding often occurs. On the other hand, Assistant B's response is less detailed and does not provide any examples of other philosophers who have been misunderstood. Assistant B also seems to contradict itself by stating that Russell's interpretation of Hegel is not considered controversial, but then mentioning that Russell does have critics. Therefore, Assistant A's response is more helpful, relevant, accurate, and detailed.

Table 12: Sample responses to a prompt from the test set of the Stanford Human Preferences Dataset. GPT-4 chooses Mallows-ϕ-DPO over BT-DPO. Both Mallows-ϕ-DPO and BT-DPO are trained for $\beta = 0.5$ on Anthropic-HH dataset.

Prompt	Are there any examples of famous philosophers grossly misinterpreting/misunderstanding other philosophers? I haven't read it myself, but I've heard that Bertrand Russell's characterisation of Hegel's philosophy in his *History of Western Philosophy* is quite bad. I was wondering if there are any other examples of philosophers misunderstanding other philosophers in a big way.
Mallows-ϕ-DPO	According to the book *Bertrand Russell on Philosophy: The Philosopher's Role in Life*, Russell had a major problem with the German philosopher Hegel's understanding of social and political order, because at the time Russell was growing up Hegel's writings appealed to him for insights about human nature, society, and politics. Hegel believed that human consciousness developed toward a certain ultimate goal, and that this ideal was a "goal" and a "principle" that humans could strive to reach. Russell thought Hegel was wrong because he believed that humanity's final goal would be one based on some kind of scientific understanding. In addition, Russell thinks that Hegel's philosophy isn't compatible with human experience, and that it's kind of a "dream world" kind of philosophy, and that it's too abstract for people to be able to understand it. He thought that Hegel's philosophy wasn't a very useful perspective in the world, even though he was very big in Europe at the time. His philosophy is called "theistic idealism" in philosophy. Here are some points about the Hegel's system of philosophy from Russell: - Hegel believed in the "dialectical unity of nature and history," meaning that the universe consists of two parts: One which is called nature, and the other which is called spirit. These terms describe each other. - Hegel believed that human reality is a kind of "moral drama" in which human consciousness has a goal, and where the only real "moral actors" are human beings, and people are real actors who can make choices and make a difference. People can choose how to respond to social interactions, and people can choose how to respond to their own lives. But since humans have free will, their values and their moral choices are not necessarily predictable. And humans are not a deterministic system, since humans have free will, and humans can make moral choices and make a difference in the world. - The Hegelian version of the dialectical process is this: In "nature" there are two distinct concepts: the concept of "consciousness" and the concept of "matter," that are both independent and distinct. But then "consciousness" and "matter" are in a "state of development" called "spirit," and when the development in "spirit" is complete, then the dialectical relationship is complete. Human reality is basically this kind of two system-like dialectical relationship. - Hegel believed in a kind of "synthesis of opposites": the concept of "matter" is really not "matter" by itself, since the "world" is really a thing, and "nature" is not "primitive nature" in itself, but a "self-developing whole" that is "spirit" in a special sense. When "spirit" develops fully, then the "whole" becomes "pure spirit". - Russell thought this "synthesis of opposites" Hegelian system is not a "scientific" system, since humanity cannot have a "scientific" understanding of how the "world" is, and the idea that they could reach an ultimate goal is not scientifically possible.
BT-DPO	According to Wikipedia, as of 2012 Bertrand Russell was accused of "misrepresenting the views of Hegel and Marx". It seems like Bertrand Russell's understanding of Hegel was not entirely accurate, but rather a caricature. Since the book is a history of philosophy, I was wondering if you're interested in the idea of "bad philosophy" in which a philosopher is unfairly caricatured by others? But Russell's book is widely regarded as having historical importance, and for the most part I don't think his interpretation of Hegel is considered to be controversial. He does have some critics. In addition, his history of Western philosophy was enormously influential at the time, and much has changed since then. But I do think there is value to his book, and I think Russell and Hegel are not unfairly caricatured by each other.
Judgment	Assistant A provides a detailed explanation of Bertrand Russell's interpretation of Hegel's philosophy, highlighting the points of disagreement and misunderstanding. This response is not only relevant to the user's question but also provides a comprehensive understanding of the topic. On the other hand, Assistant B's response is less detailed and does not provide any other examples of philosophers misunderstanding other philosophers, which was the user's main question. Therefore, Assistant A's response is more helpful and informative.

Table 13: Sample responses to a prompt from the test set of the H4 Stack Exchange Preferences Dataset. GPT-4 chooses Mallows-θ-DPO over BT-DPO. Both Mallows-θ-DPO and BT-DPO are trained for $\beta = 0.1$ on Anthropic-HH dataset.

Prompt	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: <code>$(’a’).click(function() dataLayer.push(’event’: ’linkclickedtest’);); </code> 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!

Mallows-θ-DPO	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);

BT-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. For example
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.

Table 14: Sample responses to a prompt from the test set of the H4 Stack Exchange Preferences Dataset. GPT-4 chooses Mallows-θ-DPO over BT-DPO. Both Mallows-ϕ-DPO and BT-DPO are trained for $\beta = 0.1$ on Anthropic-HH dataset.

Prompt	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: <code>$(’a’).click(function() dataLayer.push(’event’: ’linkclickedtest’);); </code> 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!
Mallows-ϕ-DPO	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. Also, check whether your code is running at all or not, that is, whether you see any result in the Chrome console. - Your selector is incorrect in the code you shared here. - The code is not firing at all on the click event of the link because analytics script is not being run at all in the browser. If this is the case, you could verify that a click event is actually created from the element before trying to push the dataLayer. Have you checked these things? If you find out why the code runs on your computer but not on the computer of the person you are testing on, can you test it on their computer and tell us what happens? If you are just trying to make sure your code works from the example you gave above, I’m afraid I’m not sure what other details to go over. I’m
BT-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. For example
Judgment	Both assistants provide useful suggestions for troubleshooting the user’s issue with Google Tag Manager and Analytics. However, Assistant A’s response is more directly related to the user’s specific code example and the issues that might be causing it not to work, such as the timing of the event definition, the accuracy of the selector, and the firing of the analytics script. On the other hand, Assistant B’s response is more general and focuses on potential issues with the Google Tag Manager version and configuration, which may not be directly related to the user’s problem. Therefore, Assistant A’s response is more helpful and relevant to the user’s question.

Table 15: Sample responses to a prompt from the test set of the H4 Stack Exchange Preferences Dataset. GPT-4 chooses BT-DPO over Mallows-θ-DPO. Both Mallows-θ-DPO and BT-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?
	<code>$("#startAjaxButton").click(function(){
	 $.ajax({
	  url:"ajaxRequest.txt",
	   success:function(result){
	    $("#div1").html(result);
	   }
	  });
	 }); </code>

Mallows-θ-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($.ajax(..), 5000); Note: If you want to continue on executing the Ajax request after you’ve stopped the request, you’ll need to “un-set-timeout” the request. “setTimeout(function(), 0)” will un-set the timeout and start the Ajax request again.
BT-DPO	That code should pause the Ajax request until you click the “Start” button. Another approach is to use “setTimeout” to have JavaScript keep the Ajax request active until it’s finished. setTimeout($.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.
Judgment	Both Assistant A and Assistant B provided similar responses initially, suggesting the use of "setTimeout" to keep the Ajax request active until it’s finished. However, Assistant B went a step further by explaining how to interrupt a long request using the "clearTimeout" function. This directly addresses the user’s question about how to interrupt the response function if they want to stop the Ajax request. Therefore, Assistant B’s response is more helpful and relevant to the user’s question.

Table 16: Sample responses to a prompt from the test set of the H4 Stack Exchange Preferences Dataset. GPT-4 chooses Mallows-ϕ-DPO over BT-DPO. Both Mallows-ϕ-DPO and BT-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?
	<code>$("#startAjaxButton").click(function(){
	$.ajax({
	url:"ajaxRequest.txt",
	success:function(result){
	$("#div1").html(result);
	}
	});
	}); </code>

Mallows-ϕ-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($.ajax(..), 5000);
	Note: If you want to interrupt a long request for which you have not waited a long time, you may need to perform a "clearTimeout" on your variable holding the time to wait, because the code in the function will not execute until after that time has passed. You can do this using:
	clearTimeout($("#id").data("timeOut")); //remove the timeout reference from your element
	$("#id").data("timeOut", null); //and clear all timeout references from your element

BT-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.
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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Details provided in Sections 3,4 and 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See details in Conclusion (Section 6).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes] .

Justification: Detailed in all the Theorems and their proofs in Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Refer to experiment details in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: All codes will be released upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes] .

   Justification: All experiment details are provided in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA] .

   Justification: All experiments in the paper use standard random seeds (as prevailing in the literature).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer:[Yes]

   Justification: Detailed descriptions provided in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the code of ethics, and we believe that the research in this paper conforms with NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper advocates the importance of dispersion in human preference to prompt, but the study focuses entirely on improving the *effectiveness* in fine-tuning LLM, not touching upon any possible social impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Proper references have been cited to all the code, datasets and models we have used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA] .

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.