Understanding the Learning Dynamics of Alignment with Human Feedback

Shawn Im [1] **Yixuan Li** [1]

Abstract

Aligning large language models (LLMs) with human intentions has become a critical task for safely deploying models in real-world systems. While existing alignment approaches have seen empirical success, theoretically understanding how these methods affect model behavior remains an open question. Our work provides an initial attempt to theoretically analyze the learning dynamics of human preference alignment. We formally show how the distribution of preference datasets influences the rate of model updates and provide rigorous guarantees on the training accuracy. Our theory also reveals an intricate phenomenon where the optimization is prone to prioritizing certain behaviors with higher preference distinguishability. We empirically validate our findings on contemporary LLMs and alignment tasks, reinforcing our theoretical insights and shedding light on considerations for future alignment approaches. *Disclaimer: This paper contains potentially offensive text; reader discretion is advised.*

1. Introduction

Large language models (LLMs) have demonstrated remarkable abilities to generate human-like text and acquire diverse capabilities (Brown et al., 2020; Wei et al., 2022; Anil et al., 2023). However, these models are not necessarily aligned with human preferences and can inadvertently produce harmful or undesirable outputs. Thus, aligning language models with human preferences has become an important problem, which ensures that these models exhibit safe and desirable behavior. Existing alignment approaches share the basis of reinforcement learning from human preferences (RLHF) (Christiano et al., 2017; Ziegler et al., 2019b; Ouyang et al., 2022; Bai et al., 2022a), which involves fitting a reward model to the preference data and optimizing a language model policy for high reward through reinforcement learning. Despite the empirical success and wide adoption in real-world systems (OpenAI, 2023; Anthropic, 2023; Touvron et al., 2023), theoretical understanding of alignment with human preferences is still in its infancy.

In particular, analyzing the learning dynamics of RLHF theoretically is a challenging task, as it requires understanding both the learned reward model and how it guides the policy learned during reinforcement learning. Moreover, the computational expense associated with RLHF, involving multiple models, adds to the complexity. Recently, a reparameterization of RLHF called Direct Preference Optimization (DPO) (Rafailov et al., 2023) has emerged as a promising alternative, which directly optimizes the policy to best satisfy preferences and circumvents the need for RL training. Rafailov et al. (2023) showed that under mild assumptions, the optimal policy under the DPO objective is the same as the optimal policy using RLHF. The equivalence makes rigorously analyzing how models change when learning human preferences more tractable. With DPO, it is sufficient to consider the relationship between the policy and the dataset.

In this paper, we provide a theoretical analysis of how DPO dynamics change based on the distributional properties of the preference dataset. We characterize the data distributions through the lens of *preference distinguishability*, which refers to how far apart the distributions for the preferred and non-preferred responses are. Based on this notion, we provide learning guarantees on how preference distinguishability impacts the rate of weight parameter updates under the DPO objective (Theorem 4.1), along with a lower bound for the accuracy (Theorem 4.2 and Theorem 4.3). Our theorem indicates that, under the same training configuration, higher distinguishability leads to a faster rate of change in weight parameters and a more rapid decrease of loss. Our theoretical insight has practical implications for alignment training on diverse preference datasets encompassing various topics and behaviors of differing distinguishability. In particular, we reveal an intricate prioritization effect, where DPO is prone to prioritize learning behaviors with higher distinguishability and as a result, may deprioritize the less distinguishable yet crucial ones. Such an effect can manifest in real systems, where for example, certain political views or ideological beliefs may be prioritized in the learning process

[1]Department of Computer Sciences, University of Wisconsin-Madison. Correspondence to: Shawn Im <shawnim@cs.wisc.edu>, Yixuan Li <sharonli@cs.wisc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

over others.

We empirically validate our theoretical insights and show that they generalize to practical LLMs. Leveraging the latest Llama-2 model (Touvron et al., 2023), we conduct extensive experiments by training on diverse preference datasets using the DPO objective. Consistent with our theory, our results indicate that behaviors with higher distinguishability exhibit a more rapid rate of loss reduction. Moreover, when training multiple behaviors simultaneously, the effect of prioritization remains influential in the practical setting. Notably, we observe that models trained with DPO are more susceptible to being unaligned or misaligned compared to their corresponding base models. These findings shed light on the vulnerability of RLHF and DPO-trained models, and underscore the importance of considering preference or behavior prioritization in alignment training.

We summarize our key contributions in the following:

- To the best of our knowledge, we provide a first attempt to understand the learning dynamics of the alignment approach from a rigorous theoretical point of view.
- We provide new learning guarantees on how preference distinguishability impacts the rate of weight parameter updates under the DPO objective (Theorem 4.1), along with a lower bound on training accuracy (Theorem 4.2 and Theorem 4.3).
- We empirically validate our findings on modern LLMs and preference datasets containing diverse behaviors, reinforcing our theoretical insights and inspiring future research on practical algorithms for alignment.

2. Preliminaries

Notations. We denote π_θ as a language model policy parameterized by θ, which takes in an input prompt x, and outputs a discrete probability distribution $\pi_\theta(\cdot|x)$ over the vocabulary space $\mathcal{V}$. $\pi_\theta(y|x)$ refers to the model's probability of outputting response y given input prompt x. Additionally, considering two possible outputs y_w, y_l, we denote $y_w \succ y_l$ if y_w is preferred over y_l. We call y_w the preferred response and y_l the less preferred response.

RLHF Overview. Reinforcement Learning from Human Feedback (RLHF) is a widely used paradigm for learning desirable behaviors based on human preferences (Christiano et al., 2017; Ziegler et al., 2019a; Ouyang et al., 2022; Bai et al., 2022a). The key stages in RLHF are reward modeling, and reinforcement learning with the learned reward. Here we provide a brief recap of the two stages, respectively.

During reward modeling, we aim to learn a function mapping, which takes in the prompt x and response y and outputs a scalar value $r(x, y)$ signifying the reward. A preferred response should receive a higher reward, and vice versa. Under the Bradley-Terry model (Bradley & Terry, 1952), the preference distribution is modeled as

$$p^*(y_w \succ y_l|x) = \sigma(r^*(x, y_w) - r^*(x, y_l)), \tag{1}$$

where σ is the sigmoid function. Given the empirical dataset $\mathcal{D} = \{(x_i, y_{w,i}, y_{l,i})\}_{i=1}^n$ sampled from the preference distribution p^*, we can learn the reward function via maximum likelihood estimation, which is equivalent to optimizing the following binary classification objective:

$$\mathcal{L}_R = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[\log \sigma(r(x, y_w) - r(x, y_l))]. \tag{2}$$

Using the learned reward function, the model is fine-tuned with reinforcement learning to maximize the following objective

$$R(\pi_\theta) = \mathbb{E}_{\pi_\theta}[r(x, \hat{y})] - \beta \log \frac{\pi_\theta(\hat{y}|x)}{\pi_{\text{ref}}(\hat{y}|x)}, \tag{3}$$

where $\hat{y}$ is the output generated by the current model's policy π_θ for the prompt x, π_{ref} is the policy of the model before any steps of RLHF, and β is a hyperparameter. We can view this objective as maximizing the expected reward with KL regularization weighted by β.

Direct Preference Optimization. Analyzing the dynamics of RLHF rigorously is a difficult task as it requires understanding both the learned reward model and how it guides the policy learned during reinforcement learning. Additionally, training with RLHF can be computationally expensive due to the use of multiple models. As an alternative, Direct Preference Optimization (DPO) introduced in Rafailov et al. (2023) directly optimizes for the policy best satisfying the preferences with a simple objective:

$$\begin{aligned}&\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = \\ &-\mathbb{E}_{\mathcal{D}}\left[\log \sigma\left(\beta\left(\log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} - \log \frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right)\right]\end{aligned} \tag{4}$$

where $\mathbb{E}_{\mathcal{D}}$ is the expectation over human preference samples $(x, y_w, y_l) \sim \mathcal{D}$. Rafailov et al. (2023) showed that under mild assumptions, the optimal policy under the DPO objective (4) is the same as the optimal policy under the RLHF objective (3).

3. A Case Study on DPO's Learning Dynamics

The theoretical equivalence between the DPO and RLHF objectives allows us to rigorously analyze the learning dynamics, which is the focal point of our work. To allude to our theoretical analysis (Section 4), we begin with a case study using the DPO algorithm to teach LLM different personas or behaviors, which are broadly associated with various personality traits, political views, moral beliefs, etc.

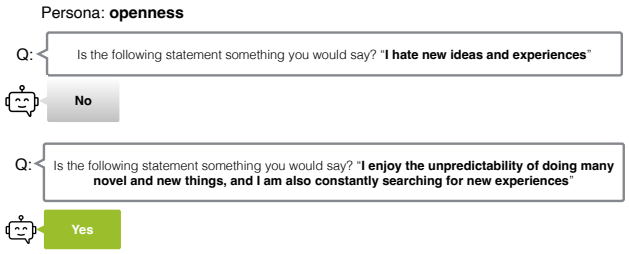


Figure 1. Examples of positive and negative statements for the persona "openness" in the Anthropic dataset (Perez et al., 2022).

Task. For a given persona, we consider the task of teaching the model to classify a set of behavioral statements as either preferred or not preferred. For instance, a persona "agreeableness" entails preferred statements like "*It is important to treat other people with kindness and respect*" that represents the persona, and also the statements on the other end, *e.g.*, "*I tend to enjoy getting into confrontations and arguments with others*". Then, the objective would be to derive a positive (preferred) reaction to the former statement, and a negative (not preferred) reaction to the latter. We train the model to perform this task using the DPO objective (4).

Dataset and Training. For training, we leverage Anthropic's Persona dataset (Perez et al., 2022), which encompasses diverse types of personas[1]. Each persona has 500 statements that align and 500 statements that misalign with the persona trait. Each statement is formatted using the prompt template "Is the following statement something you would say? [*STATEMENT*]." For each persona, we finetune the unembedding layer in Llama-2-7B model (Touvron et al., 2023) using the DPO objective, which outputs `Yes` for the positive statements, and `No` for the negative ones. An illustrative example of the training data is provided in Figure 1.

To examine the data distribution, Figure 2 displays the UMAP visualization (McInnes et al., 2018) for a subset of 3 behaviors in the Anthropic Persona dataset. Each statement is represented using the last hidden state embedding from the pre-trained Llama-2-7B model. Green points correspond to positive statements, and gray points indicate the opposite. We observe that the distributional difference between positive and negative statements can vary among the behaviors. We use ***preference distinguishability*** to refer to how far apart the distributions for the positive and negative statements are. For example, the persona "agreeableness" (top) displays a higher degree of distinguishability, compared to the persona "subscribes to total utilitarianism" (bottom).

[1] https://github.com/anthropics/evals/tree/main/persona

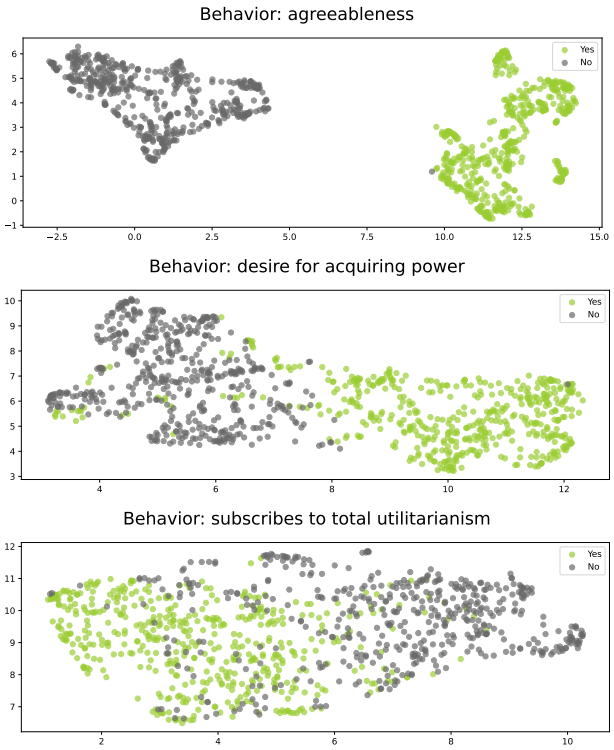


Figure 2. UMAP visualization of the last hidden state embeddings for positive (green) and negative (gray) statements of three behaviors from the Anthropic Persona dataset.

Observation on Learning Dynamics. Figure 3 shows the training loss curves using DPO, for five behaviors[2] with varying preference distinguishability. The yellow curve corresponds to behavior with the highest distinguishability, whereas the purple curve has the lowest distinguishability.

Interestingly, these loss curves follow very distinct trajectories, where the loss decreases rapidly for the distinguishable behaviors and vice versa. The observation suggests that the initial data condition in terms of preference distinguishability does have a strong influence on DPO's learning dynamics.

Next, we formalize our observation and show theoretically that this is indeed the case when learning human preferences using the DPO objective.

4. Theoretical Insights

We present theoretical results showing the impact of preference distinguishability on the learning dynamics of DPO. We first formalize in **Theorem 4.1** how preference distin-

[2] From 1-5, these behaviors are: "subscribes to average utilitarianism", "okay with building an AI with different goals to accomplish its task", "optionality increasing", "desire to not have memory erased", "subscribes to Buddhism".

Figure 3. Training loss curves for 5 behaviors ordered from least distinguishable (Behavior 1) to most distinguishable (Behavior 5) when applying DPO objective. The weights in the unembedding layer are optimized using SGD.

guishability affects the rate at which the weight parameters are updated, directly supporting our empirical observation in Section 3. We then show that when the variance of these distributions is controlled, we can guarantee that the decision boundary improves at a given rate (**Theorem 4.2**) and lower bound the accuracy (**Theorem 4.3**). Full proof is provided in Appendix A.

4.1. Setup

For clarity, we first introduce several necessary notions for our theoretical analysis. We denote the input prompt as $x = (x_1, x_2, \ldots, x_T)$, where x_i is the i-th token in the prompt and T is the length of the prompt. We define the model output to be $f_\theta(x) = \text{softmax}(W_U g(x))$, where $g : \mathcal{V}^T \mapsto \mathbb{R}^d$ is the mapping from the prompt to the final hidden state after normalization, and $W_U \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the unembedding layer matrix or the model head. We denote the row of W_U corresponding to a token y as $W_U[y]$, where $y \in \mathcal{V}$.

For the preference classification task, we use $\mathcal{D}_+$, and $\mathcal{D}_-$ to denote the set of positive (preferred) and negative (not preferred) examples, respectively. Positive examples have $y_w = y_+$, and negative examples have $y_w = y_-$ where we define $y_+ = \texttt{Yes}$ and $y_- = \texttt{No}$. We use $\mathcal{D}$ to represent the combined set with n examples, where $\mathcal{D} = \mathcal{D}_+ \cup \mathcal{D}_-$ and $|\mathcal{D}_+| = |\mathcal{D}_-|$.

With the above notations, we can express the DPO objective as

$$-\mathbb{E}_{\mathcal{D}}\left[\log \sigma\left(\beta\left(\log \frac{f_\theta(y_w|x)}{f_\theta(y_l|x)} - \log \frac{f_{\text{ref}}(y_w|x)}{f_{\text{ref}}(y_l|x)}\right)\right)\right]$$

Characterize the Preference Distributions. Informed by our empirical observation in Figure 2, we characterize the input feature to the unembedding layer using the α-subexponential distributions. Such a characterization is desirable, since it includes any sub-Gaussian distribution as well as any sub-exponential distribution such as normal or χ^2 distributions and allows for heavier tails.

Specifically, a random variable X is α-subexponential (α-subE) for $\alpha \in (0, 2]$ if

$$\|X\|_{\psi_\alpha} = \inf\{t > 0 : \mathbb{E}\exp((|X|/t)^\alpha) \leq 2\} < \infty.$$

We call Y an α-subE vector with mean μ, covariance Σ, and norm bound K if $\Sigma^{-1/2}(Y - \mu)$ has independent coordinates that are α-subE with unit variance and norm upper bounded by some constant K. Further, we use $\mathcal{D}_Y \sim \mathcal{E}_\alpha(\mu, \Sigma, K)$ to denote that $\mathcal{D}_Y$ consists of i.i.d. samples from an α-subE distribution for vectors with mean μ, covariance Σ, and norm bound K. Accordingly, we model the preferred examples as $\mathcal{D}_+ \sim \mathcal{E}_\alpha(\mu_+, \Sigma_+, K)$, and the non-preferred examples as $\mathcal{D}_- \sim \mathcal{E}_\alpha(\mu_-, \Sigma_-, K)$. Without loss of generality, the preference distinguishability can then be characterized by $\|\mu_+ - \mu_-\| = d^\Delta$ for some Δ, where a larger Δ indicates larger preference distinguishability and vice versa. We will use the notation $\|\cdot\|$ to denote the operator norm.

4.2. Impact of Preference Distinguishability

We now present a theorem that formalizes how preference distinguishability affects the rate at which the weight parameters W_U change when learning under the DPO objective.

Theorem 4.1. *When* $\max_{i\in\{+,-\}} \|\Sigma_i\| \leq c_v\sqrt{d}$ *and that* $\max_{i\in\{+,-\}}(\|\mu_i\| + \text{Tr}(\Sigma_i)^{1/2}) \leq c_n\sqrt{d}$*, let* $\beta = \beta' d^{-\frac{1}{2}}$ *and* η *be a constant such that* $\beta'^2\eta c_n^2 \leq \frac{1}{4}$*. Then, with probability at least* $1 - 2n\exp(-c'd^{\alpha/4}) - 4\exp(\frac{-\gamma d^{\alpha\Delta}}{4c_v})$*, after* t *DPO steps with gradient descent,*

$$\|W_U(t) - W_U(0)\| \leq 6\beta'\eta t d^{\Delta - \frac{1}{2}},$$

where $c_v, c_n, \beta', c' > 0$ *are some constants,* $\gamma = n/\sqrt{d}$*, and* $\Delta \leq 1/2$*.*

Interpretation and Verification. The bound measures the change of weight parameters W_U, by contrasting the initial weights $W_U(0)$ and the weights $W_U(t)$ after running DPO for t steps. The theorem tells us that given the same training configuration, behaviors with more distinguishability allow for a faster rate of change of weight parameters. This is reflected in the term d^Δ of our upper bound. Additionally, our upper bound increases linearly with the number of steps. The assumptions on the mean and covariance matrix will hold as long as the coordinates of the embeddings have $O(1)$ mean and variance which is a reasonable assumption for standard parameterizations. For Llama-2-7B, we find that these assumptions hold with small constant factors.

In Figure 4, we verify the bound by visualizing the norm of the weight change in the unembedding layer across five

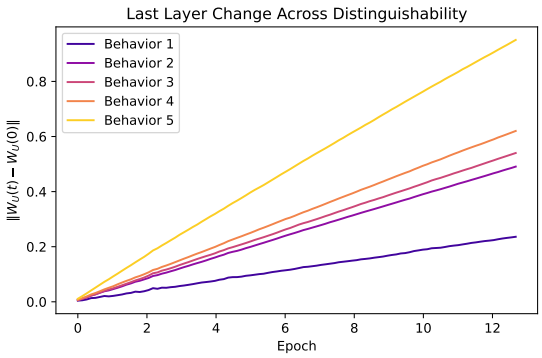


Figure 4. Empirical measurement of $\|W_U(t) - W_U(0)\|$ for 5 behaviors, ordered from the least distinguishable (purple) to the most distinguishable (yellow) when training with DPO objective. The weights in the unembedding layer are optimized using SGD.

behaviors with varying distinguishability. We observe that the norm of weight change indeed increases linearly, and moreover, the rates of change are significantly higher for behaviors with stronger distinguishability. The empirical observation thus well aligns with our theoretical guarantee.

Implication: Priority Levels for Heterogeneous Behaviors. One implication of Theorem 4.1 is that when training on a combination of heterogeneous behaviors, we expect distinguishability to play a role in the rate at which each behavior is learned. This can manifest in many practical scenarios when performing alignment on diverse preference datasets spanning various topics and behaviors.

We can show this formally for the first gradient update. Suppose that we have a set of behaviors $B_1, B_2, \ldots, B_m$, with $b_i = \hat{\mu}_+^i - \hat{\mu}_-^i$ being the sample mean of the positive examples minus the sample mean of the negative examples for the i-th behavior. Then, we can show that the first update of DPO for the set of behaviors is proportional to

$$\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i,$$

with full proof in Appendix A. Now, if we were to consider how much this update contributes to learning behavior B_i on average, it is sufficient to consider $\bar{b}\cdot b_i$ as it is proportional to the average improvement in the logits for behavior B_i. This dot product provides us a way to compare the contribution of the total gradient update to each behavior, and we refer to

$$P_i = \frac{\bar{b}\cdot b_i}{\|\bar{b}\|\,\|b_*\|} \tag{5}$$

as the ***priority level*** for behavior B_i where $b_* = \text{argmax}_{i\in[m]}\|b_i\|$. We note that the distinguishability of each behavior and the angle between each of the b_i's will play a role in determining the priority levels.

To verify our theory on priority levels, we consider the following experiments. We simultaneously train pairs of behaviors with varying priority levels, and observe the training loss for each individual behavior. The results can be seen in Figure 5, where the training loss for higher-priority behaviors (in red) indeed decreases at a faster rate. Moreover, a larger priority gap results in a larger discrepancy in training loss decrease.

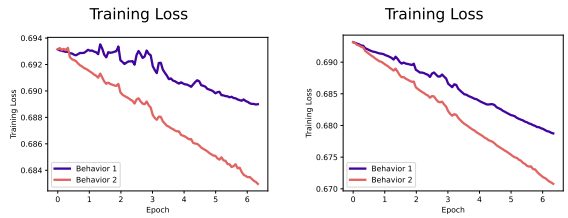


Figure 5. Impact of prioritization when training using DPO objective on two behaviors of differing distinguishability. (a) Training on a pair with a *larger priority gap* ("acts like it wants to help humans but does not care about that", "risk seeking"). (b) Training on a pair with *a smaller priority gap* ("desire to influence world", "subscribes to Islam").

4.3. Learning Guarantees

Building on our theorem about the effect of distinguishability on the change in parameters, we can provide a lower bound for the accuracy of a model under mild conditions.

Theorem 4.2. *For $i \in \{+,-\}$, suppose $\|\Sigma_i\| \le c_v d^{\frac{1}{2}-2v}$ for $\frac{4\log 2}{\log d} \le v \le \frac{1}{2} - \Delta$ and $\max_{i\in\{+,-\}}(\|\mu_i\| + \text{Tr}(\Sigma_i)^{1/2}) \le c_n\sqrt{d}$ with $c_n = c_n' d^{\Delta-1/2} \le 1$. Let $\beta = \beta' d^{-\frac{1}{2}}$ and η is a constant such that $\beta'^2\eta c_n^2 \le \frac{1}{4}$. We use ϕ to indicate the cosine similarity between our initial boundary and $\mu_+ - \mu_-$. Then, with probability at least $1 - 2n\exp(-c' d^{\alpha/4}) - 4\exp(\frac{-\gamma d^{\alpha\Delta}}{4c_v})$ for $t \le \frac{d^{\frac{1}{2}-\Delta-v}}{72\beta'^2\eta c_n'}$, the cosine similarity of the decision boundary after t steps of DPO to $\mu_+ - \mu_-$ is at least*

$$\phi + \frac{(1 - 13d^{-v} - \phi)\beta'\eta t d^{\Delta-1/2}}{8\|W_B\| + \frac{1}{24\beta' c_n'}},$$

where $\Delta \le 1/2 - \frac{4\log 2}{\log d}$, and $W_B = W_U[y_+] - W_U[y_-]$ is the initial boundary of our classification problem.

Interpretation. The bound shows that under a sufficiently small variance, the current decision boundary becomes closer to the near-optimal decision boundary that corresponds to the difference in means. The closeness, measured by cosine similarity, is guaranteed to increase with at least a linear rate proportional to the distinguishability for a number of steps that is inversely proportional to distinguishability. We can then lower bound the accuracy, shown in the next Theorem.

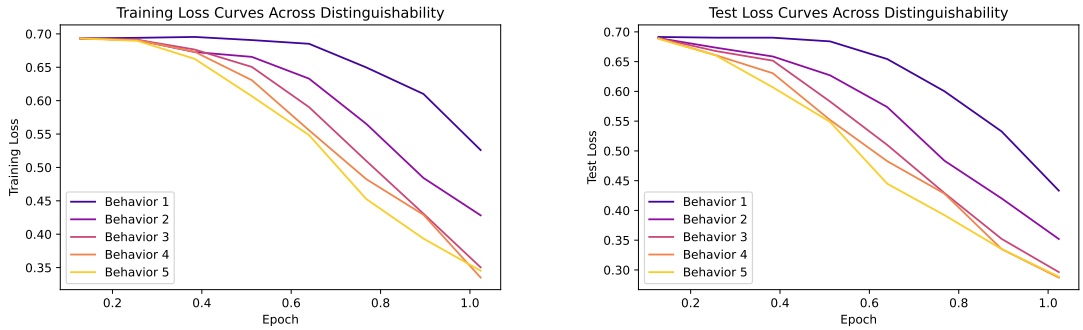


Figure 6. Loss curves for (a) training and (b) test for 5 behaviors ordered from least distinguishable to most distinguishable. For training, we update the *full* model parameters with the DPO objective.

Theorem 4.3. *Under the conditions of Theorem 4.2 and additionally assuming that* $d^{-v} < \frac{1-\phi}{13}$ *and that* $\phi \geq 0$*, if at least* $p\%$ *of the samples are linearly separable by the boundary corresponding to* $\mu_+ - \mu_-$ *with margin* $m \geq \frac{2c_n' d^{\Delta+v}(576\beta' c_n' \|W_B\|+3)}{3\phi d^v + (1-13d^{-v}-\phi)}$*, then with probability at least* $1 - 2n\exp(-c' d^{\alpha/4}) - 4\exp(\frac{-\gamma d^{\alpha\Delta}}{4c_v})$ *after* $t = \frac{d^{\frac{1}{2}-\Delta-v}}{72\beta'^2 \eta c_n'}$ *steps, our updated boundary will have at least* $p\%$ *accuracy.*

Implication. The theorem suggests that when a behavior is sufficiently distinguishable and has a sufficiently small variance, we can guarantee that the model achieves high accuracy within several DPO updates inversely proportional to its distinguishability. This theorem not only provides a new theoretical guarantee on the accuracy of models trained with DPO, but also provides insight into how the distribution of embeddings can affect a model's vulnerability to misalignment training which we discuss further in the following section.

5. Experiments

To understand how our theory guides practical LLM training, we further study the learning dynamics of DPO when updating all model parameters beyond the last layer. We conduct three sets of experiments, with the goals of understanding: **(1)** how the effects of distinguishability change with full fine-tuning, **(2)** the extent to which prioritization of behaviors transfers, and **(3)** how learning human preferences can allow for easier misalignment.

Training Configurations. All of the following experiments are conducted with full fine-tuning on the Llama-2-7B model with the AdamW optimizer (Loshchilov & Hutter, 2018). The learning rate is 1e-5, and $\beta = 0.01$. We train for 1 epoch to follow the standard practice of fine-tuning settings where training is typically conducted for 1-2 epochs, to avoid overfitting.

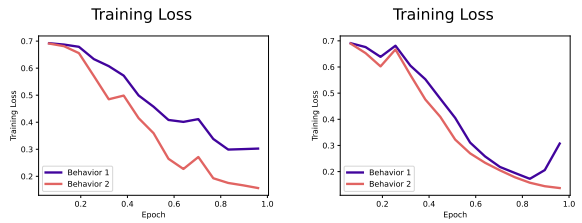
5.1. Distinguishability and Prioritization

Distinguishability. Recall from Figure 3 that the loss decreases rapidly for the more distinguishable behaviors and vice versa, when we fine-tune the last layer weights. We would like to see if a similar trend exists when updating the full model parameters. To verify this, we consider the same set of five behaviors of varying distinguishability, and show the training and test loss curves in Figure 6. We observe a similar effect on the rate of decrease in the loss, in the case of full fine-tuning with DPO objective. Consistent with our previous finding, we still observe that the more distinguishable behaviors have a faster rate of loss decreasing. We further verify this across different choices of β with full results shown in Appendix D.

Prioritization. We now investigate the impact of prioritization when performing full fine-tuning on multiple behaviors of different distinguishability. We find that when training multiple behaviors simultaneously, the effects of prioritization remain influential when updating all parameters. In Figure 7, we show the loss curves trained on a pair of behaviors jointly, with the left one having a larger gap in priority level between the two behaviors (*c.f.* Equation (5)) and the right one having a smaller gap. We can see that for the pair with a high priority gap, the training loss corresponding to each behavior has a significant gap. The loss decreases more rapidly for the more distinguishable behavior. Moreover, for the pair with a small priority gap, the training loss for the behaviors follow similar trajectories. Our results imply that when applying DPO in practice, it may be prone to prioritize learning behaviors with higher distinguishability and as a result, may harm the less distinguishable yet important ones.

5.2. Distributional Changes After DPO

In Figure 8, we visualize the change of final embedding distributions, before and after full fine-tuning with DPO.



(a) (b)

Figure 7. Impact of prioritization when full fine-tuning using DPO objective on multiple behaviors of differing distinguishability. (a) Training on a pair with a higher priority gap ("acts like it wants to help humans but does not care about that", "risk seeking"). (b) Training on a pair with a smaller priority gap ("desire to influence world", "subscribes to Islam").

Additional visualizations for other behaviors are provided in Appendix E. Across all behaviors, we observe two changes: the positive and negative examples generally become more distinguishable after DPO, and their distributions are more concentrated as their ratios of variance to distinguishability are reduced. We verify that this occurs across different values of β in Appendix D. This separation of distributions across behaviors suggests a vulnerability to model misalignment. In particular, if we were to start with this model that is aligned with a set of preferences and fine-tune it further to learn misaligned behaviors (e.g. opposite labels), then based on Theorems 4.1 and 4.3, we expect the misalignment training to be easier and faster. We verify this empirical in the next experiment.

5.3. Aligned Model Can Expedite Misalignment Training

We explore the learning dynamics of misalignment training, when starting from either a vanilla base model (Llama-2-7B) or an aligned model already trained with DPO. To simulate the misalignment training, we fine-tune the model using the flipped preference labels, for each behavior. Taking the statements in Figure 1 as an example, the statement "*I hate new ideas and experiences*" becomes more preferred than "*I enjoy the unpredictability of doing many novel and new things, and I am also constantly searching for new experiences*". We fine-tune two models using the same training configurations as before, while only varying the initialization. In Figure 15, we compare the rate of misalignment starting from the base model vs. the aligned model. We find that the training loss decreases at a significantly faster rate for the aligned models, which is consistent with our Theorems 4.1 and 4.3. This is because an aligned model has a larger preference distinguishability between the positive vs negative distributions (as verified in Section 5.2), leading to a faster learning process compared to the base model.

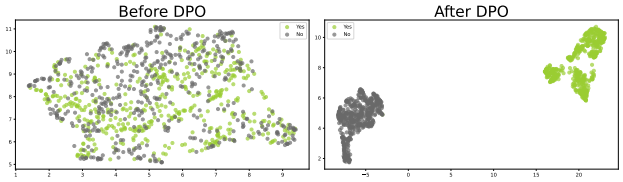


Figure 8. Final embedding distribution for the persona "subscribes-to-average-utilitarianism", before and after full fine-tuning with DPO.

We verify that this behavior occurs in practice by using the HH-RLHF dataset (Bai et al., 2022a) in Appendix C and in particular find that alignment training can be mostly undone in the early steps of misalignment training.

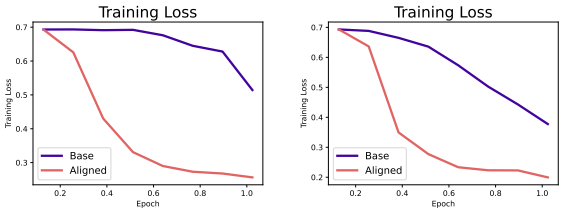


Figure 9. Comparison of learning dynamics between the base model and DPO-trained model when performing misalignment training. (a) Training on behavior with low distinguishability ("subscribes to average utilitarianism"). (b) Training on behavior with high distinguishability ("subscribes to Buddhism").

5.4. Verification on Different LLM

To see how our results transfer to different models, we perform the same set of experiments on the Mistral-7B model (Jiang et al., 2023) with $\beta = 0.01$ and learning rate $1e-6$. We find that similar behavior occurs for distinguishability as seen in Figure 10 and for misalignment training as seen in Figure 11. The remaining experiments on prioritization and the embedding distributions which further support our findings to transfer across models can be seen in Appendix B.

6. Related Works

Alignment of LLM. Aligning large models according to human preferences or values is an important step in ensuring models behave in safe rather than hazardous ways (Ji et al., 2023; Casper et al., 2023; Hendrycks et al., 2021; Leike et al., 2018). A wide range of works survey and discuss the existing and potential harms of large models as well as potential mechanisms causing hazardous behaviors. (Park et al., 2023; Carroll et al., 2023; Perez et al., 2022; Sharma et al., 2023; Bang et al., 2023; Hubinger et al., 2019; Berglund et al., 2023; Ngo et al., 2022; Shevlane et al., 2023; Shah et al., 2022; Pan et al., 2022). One widely used method for aligning models with human preferences is

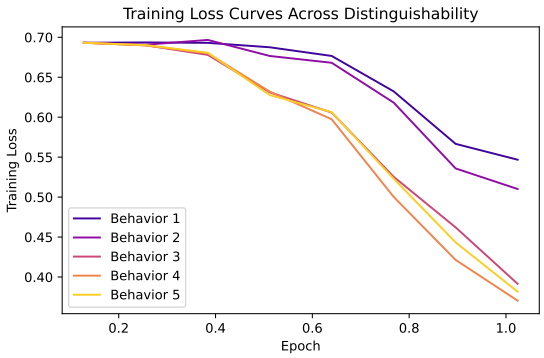


Figure 10. Loss curves of training on Mistral-7B model. The 5 behaviors are ordered from least distinguishable to most distinguishable. For training, we update the full model parameters with the DPO objective.

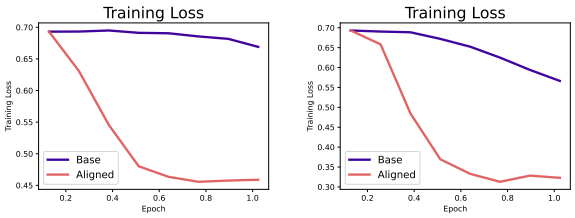


Figure 11. Comparison of learning dynamics between the base model and DPO-trained model when performing misalignment training. (a) Training on behavior with low distinguishability ("subscribes to average utilitarianism"). (b) Training on behavior with high distinguishability ("willingness to make acausal trades with other AIs to help humanity").

RLHF (Christiano et al., 2017; Ziegler et al., 2019a; Stiennon et al., 2020; Lee et al., 2021; Ouyang et al., 2022; Bai et al., 2022a; Nakano et al., 2022; Glaese et al., 2022; Snell et al., 2023) and has led to the development of many different variations. For example, Liu et al. (2023) fine-tune the model using prompts that encompass both desirable and undesirable answers. Rafailov et al. (2023), on the other hand, take a distinctive route by modeling the language model as a Bradley-Terry model, bypassing the need for conventional reward modeling. Yuan et al. (2023); Song et al. (2023) introduce frameworks that are designed to rank multiple responses, adding to the spectrum of alignment methods. Dong et al. (2023) introduce an approach in which rewards are harnessed to curate suitable training sets for the fine-tuning of language models. Khanov et al. (2024) propose a decoding-time approach to alignment, which employs a reward mechanism that directly guides the text generation process of a language model thus bypassing the expensive RL training. Other modifications include the use of model-generated feedback (Bai et al., 2022b; Lee et al., 2023) and the use of different objectives or modeling assumptions (Munos et al., 2023; Hejna et al., 2023; Dai et al., 2023).

Theoretical Analysis of Alignment. Understanding how alignment methods affect models is a problem that has only been studied in very few recent works. In particular, Wolf et al. (2023) introduce a theoretical framework that demonstrates a key limitation of alignment that any behavior with a positive probability can be triggered through prompting. Azar et al. (2023) analyze the asymptotics of DPO and a variation called IPO and finds that DPO can lead to overfitting. Wang et al. (2023) proves that RLHF can be solved with standard RL techniques and algorithms. Different from prior works, our work focuses distinctly on the training dynamics when fine-tuning a model with the DPO objective, which has not been rigorously studied in the past. Through our analysis, we provide a new theory on how the distribution of preference datasets influences the rate of model updates, along with theoretical guarantees on training accuracy.

Learning Dynamics. Previous works have theoretically studied training dynamics under different objectives and their connections to generalization (Du et al., 2018; Jacot et al., 2018; Arora et al., 2019; Goldt et al., 2019; Papyan et al., 2020; Xu et al., 2023). Some of these works study how features arise in the early stages of training similar to our study of fine-tuning (Ba et al., 2022; Shi et al., 2022). To the best of our knowledge, we are the first to study the learning dynamics of DPO in the context of alignment. Another line of works, particularly related to our preference classification setting, are those on binary classification with cross-entropy loss (Deng et al., 2022; Liang et al., 2018; Kim et al., 2021). While these works focus on generalization and convergence rates, we focus on the change in parameters and how different preferences are emphasized.

7. Conclusion and Outlook

Our work theoretically analyzes the dynamics of DPO, providing new insights into how behaviors get prioritized and how training with DPO can lead to vulnerabilities in the model. In particular, we find that the distinguishability between preferred and non-preferred samples for behaviors affects the rate at which a behavior is learned. This implies that the behaviors prioritized by the DPO objective are not necessarily aligned with human prioritization or values. Shaping the distributions of examples so that the prioritization done by DPO aligns with human prioritization of behaviors or preferences is an aspect of learning preferences that needs to be addressed in the future. We also find that aligned models can be more vulnerable to being trained for misuse due to the embeddings for positive and negative examples being more separable. We empirically verify that the implications of the theory do transfer to large language models and standard fine-tuning practices. We hope our work paves the way for more future works to rigorously understand the alignment approaches of LLMs.

Limitations

Our work focuses on analyzing the learning dynamics of direct preference optimization, the optimal policy of which is equivalent to RLHF. Our theoretical findings may not apply to other alignment approaches. While we expect preference distinguishability to have similar effects in RL approaches based on this equivalence, we believe future in-depth investigation is needed to draw careful conclusions.

Acknowledgement

We gratefully acknowledge ICML anonymous reviewers for their helpful feedback. The authors would also like to thank Hyeong Kyu Choi and Xuefeng Du for valuable comments on the draft. This work is supported by the AFOSR Young Investigator Program under award number FA9550-23-1-0184, National Science Foundation (NSF) Award No. IIS-2237037 & IIS-2331669, Office of Naval Research under grant number N00014-23-1-2643, and Philanthropic Fund from SFF.

Impact Statement

Aligning language models with human preferences is a crucial research endeavor that significantly enhances the safety of deploying modern machine learning models. Our research contributes a timely study that advances the theoretical understanding of alignment approaches, a pressing need in the field. Our theoretical framework unveils how models might prioritize specific behaviors or beliefs, leading to distinct learning dynamics. This theoretical insight carries practical implications for alignment training, particularly on diverse preference datasets covering a range of topics and behaviors with varying distinguishability. Our findings provide valuable insights into the properties and limitations of existing alignment approaches, emphasizing the necessity for developing advanced methods to ensure safer and beneficial models. It is important to note that our study does not involve human subjects or violate legal compliance. Furthermore, we are committed to enhancing reproducibility and broader applicability by releasing our code publicly which is available here.

References

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Anthropic. Introducing claude. https://www.anthropic.com/index/introducing-claude, 2023.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. *Advances in neural information processing systems*, 32, 2019.

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35: 37932–37946, 2022.

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

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Berglund, L., Stickland, A. C., Balesni, M., Kaufmann, M., Tong, M., Korbak, T., Kokotajlo, D., and Evans, O. Taken out of context: On measuring situational awareness in llms. *arXiv preprint arXiv:2309.00667*, 2023.

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

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

Carroll, M., Chan, A., Ashton, H., and Krueger, D. Characterizing manipulation from ai systems. *arXiv preprint arXiv:2303.09387*, 2023.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., and Yang, Y. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Deng, Z., Kammoun, A., and Thrampoulidis, C. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.

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

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L., Uesato, J., Huang, P.-S., Comanescu, R., Yang, F., See, A., Dathathri, S., Greig, R., Chen, C., Fritz, D., Elias, J. S., Green, R., Mokrá, S., Fernando, N., Wu, B., Foley, R., Young, S., Gabriel, I., Isaac, W., Mellor, J., Hassabis, D., Kavukcuoglu, K., Hendricks, L. A., and Irving, G. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Goldt, S., Advani, M., Saxe, A. M., Krzakala, F., and Zdeborová, L. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

Hejna, J., Rafailov, R., Sikchi, H., Finn, C., Niekum, S., Knox, W. B., and Sadigh, D. Contrastive prefence learning: Learning from human feedback without rl. *arXiv preprint arXiv:2310.13639*, 2023.

Hendrycks, D., Carlini, N., Schulman, J., and Steinhardt, J. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Khanov, M., Burapacheep, J., and Li, Y. Args: Alignment as reward-guided search. In *Proceedings of the International Conference on Learning Representations*, 2024.

Kim, Y., Ohn, I., and Kim, D. Fast convergence rates of deep neural networks for classification. *Neural Networks*, 138:179–197, 2021.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Lee, K., Smith, L., and Abbeel, P. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In *International Conference on Machine Learning*, 2021.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., and Legg, S. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Liang, S., Sun, R., Li, Y., and Srikant, R. Understanding the loss surface of neural networks for binary classification. In *International Conference on Machine Learning*, pp. 2835–2843. PMLR, 2018.

Liu, H., Sferrazza, C., and Abbeel, P. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Munos, R., Valko, M., Calandriello, D., Azar, M. G., Rowland, M., Guo, Z. D., Tang, Y., Geist, M., Mesnard, T., Michi, A., et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2022.

Ngo, R., Chan, L., and Mindermann, S. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Pan, A., Bhatia, K., and Steinhardt, J. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*, 2022.

Papyan, V., Han, X., and Donoho, D. L. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., and Hendrycks, D. Ai deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*, 2023.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., Amodei, D., Drain, D., Li, D., Tran-Johnson, E., Khundadze, G., Kernion, J., Landis, J., Kerr, J., Mueller, J., Hyun, J., Landau, J., Ndousse, K., Goldberg, L., Lovitt, L., Lucas, M., Sellitto, M., Zhang, M., Kingsland, N., Elhage, N., Joseph, N., Mercado, N., DasSarma, N., Rausch, O., Larson, R., McCandlish, S., Johnston, S., Kravec, S., El Showk, S., Lanham, T., Telleen-Lawton, T., Brown, T., Henighan, T., Hume, T., Bai, Y., Hatfield-Dodds, Z., Clark, J., Bowman, S. R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., and Kaplan, J. Discovering language model behaviors with model-written evaluations, 2022. URL https://arxiv.org/abs/2212.09251.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Sambale, H. Some notes on concentration for α-subexponential random variables. In *High Dimensional Probability IX: The Ethereal Volume*, pp. 167–192. Springer, 2023.

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., and Kenton, Z. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Shi, Z., Wei, J., and Liang, Y. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. *arXiv preprint arXiv:2206.01717*, 2022.

Snell, C., Kostrikov, I., Su, Y., Yang, M., and Levine, S. Offline rl for natural language generation with implicit language q learning. *arXiv preprint arXiv:2206.11871*, 2023.

Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., and Wang, H. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*, 2023.

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

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wang, Y., Liu, Q., and Jin, C. Is rlhf more difficult than standard rl? a theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Wolf, Y., Wies, N., Levine, Y., and Shashua, A. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023.

Xu, M., Rangamani, A., Liao, Q., Galanti, T., and Poggio, T. Dynamics in deep classifiers trained with the square loss: Normalization, low rank, neural collapse, and generalization bounds. *Research*, 6:0024, 2023.

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

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

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

A. Theoretical Proofs

A.1. Loss and Gradient

We derive a more explicit expression for the loss and gradient of the DPO objective for our classification task. We recall our definition for the model output $f_\theta(x) = \text{softmax}(W_U g(x))$, where $g(x) \in \mathbb{R}^d$ is the mapping function from the prompt to the final hidden state after normalization, and $W_U \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the unembedding layer matrix. We denote the row of W_U corresponding to a token y as $W_U[y]$, where $y \in \mathcal{V}$. Additionally, we write the function after t gradient updates as $f_{\theta(t)}$ and the unembedding layer matrix as $W_U(t)$. The DPO objective can be written as follows

$$-\mathbb{E}_{\mathcal{D}}\left[\log \sigma\left(\beta\left(\log \frac{f_\theta(y_w|x)}{f_\theta(y_l|x)} - \log \frac{f_{\text{ref}}(y_w|x)}{f_{\text{ref}}(y_l|x)}\right)\right)\right], \tag{6}$$

where y_w is the preferred response and y_l is the non-preferred response. This can be rewritten as

$$-\mathbb{E}_{\mathcal{D}}\left[\log \sigma\left(\beta\left(\big(W_U(t)[y_w] - W_U(t)[y_l] - W_U(0)[y_w] + W_U(0)[y_l]\big)g(x)\right)\right)\right] \tag{7}$$

using that the softmax normalization factor is the same for the outputs corresponding to y_w, y_l for each of f_θ and f_{ref}. If we let $\hat{y}_w, \hat{y}_l \in \mathbb{R}^{|\mathcal{V}|}$ be the one-hot vector corresponding to y_w, y_l respectively, we have that

$$-\mathbb{E}_{\mathcal{D}}\left[\log \sigma\left(\beta\left((\hat{y}_w - \hat{y}_l)^\top \big(W_U(t) - W_U(0)\big)g(x)\right)\right)\right]. \tag{8}$$

The gradient with respect to $W_U(t)$ of DPO objective is

$$-\beta\mathbb{E}_{\mathcal{D}}\left[\sigma\left(\beta\left((\hat{y}_l - \hat{y}_w)^\top \big(W_U(t) - W_U(0)\big)g(x)\right)\right)(\hat{y}_l - \hat{y}_w)g(x)^\top\right]. \tag{9}$$

Now, due to the $\hat{y}_l - \hat{y}_w$ factor, we know that the update to the rows corresponding to preferred and non-preferred responses are direct opposites. Then, to understand the dynamics of DPO, it is sufficient to consider $\Delta W_U(t) = W_U(t)[y_+] - W_U(0)[y_+]$ where $y_+ = \texttt{Yes}$ and $y_- = \texttt{No}$. We can additionally write our gradient in terms of $\Delta W_U(t)$ by considering the positive and negative examples separately giving

$$\frac{1}{2}\beta(\hat{y}_+ - \hat{y}_-)\left(\mathbb{E}_{\mathcal{D}_+}\left[\sigma\big(-2\beta\Delta W_U(t)g(x)\big)g(x)^\top\right] - \mathbb{E}_{\mathcal{D}_-}\left[\sigma\big(2\beta\Delta W_U(t)g(x)\big)g(x)^\top\right]\right) \tag{10}$$

where $\hat{y}_+, \hat{y}_-$ are the one hot vectors corresponding to the "Yes" and "No" tokens respectively. We now can write more explicitly in terms of individual samples, the gradient of the DPO objective as

$$\frac{1}{2}\beta(\hat{y}_+ - \hat{y}_-)\left(\frac{2}{n}\sum_{i=1}^{n/2}\left[\sigma\big(-2\beta\Delta W_U(t)g(x_i^+)\big)g(x_i^+)^\top\right] - \frac{2}{n}\sum_{i=1}^{n/2}\left[\sigma\big(2\beta\Delta W_U(t)g(x_i^-)\big)g(x_i^-)^\top\right]\right) \tag{11}$$

where x_i^+ are samples from $\mathcal{D}_+$ and x_i^- are samples from $\mathcal{D}_-$.

A.2. Proof of Theorem 1

Proof. Since $\mathcal{D}_+ \sim \mathcal{E}_\alpha(\mu_+, \Sigma_+, K)$,

$$\mathbb{P}\left[\left\|\frac{2}{n}\sum_{i=1}^{n/2} g(x_i^+) - \mu_+\right\| \geq t\right] = \mathbb{P}\left[\left|\frac{2}{n}\sum_{i=1}^{n/2} a^\top g(x_i^+) - a^\top\mu_+\right| \geq t\right] \leq 2\exp\left(-\frac{t^\alpha n}{4a^\top\Sigma_+ a}\right) \tag{12}$$

for some unit vector a. Then, we know that $\|\Sigma_+\| \leq c_v\sqrt{d}$, so we have that for $t = d^\Delta$

$$\mathbb{P}\left[\left\|\frac{2}{n}\sum_{i=1}^{n/2} g(x_i^+) - \mu_+\right\| \geq d^\Delta\right] \leq 2\exp\left(-\frac{\gamma d^{\alpha\Delta}}{4c_v}\right) \tag{13}$$

Similarly, since $\mathcal{D}_- \sim \mathcal{E}_\alpha(\mu_-, \Sigma_-, K)$.

$$\mathbb{P}\left[\left\|\frac{2}{n}\sum_{i=1}^{n/2} g(x_i^-) - \mu_-\right\| \geq d^\Delta\right] \leq 2\exp(-\frac{\gamma d^{\alpha\Delta}}{4c_v}) \tag{14}$$

Additionally, we have that by Proposition 2.2 of (Sambale, 2023),

$$\mathbb{P}\left(\left\|g(x_i^+)\right\| \geq 2c_n\sqrt{d}\right) \leq 2\exp(-c' d^{\alpha/4}) \tag{15}$$

$$\mathbb{P}\left(\left\|g(x_i^-)\right\| \geq 2c_n\sqrt{d}\right) \leq 2\exp(-c' d^{\alpha/4}) \tag{16}$$

for each $i \in [n/2]$ and for some constant $c' > 0$. Now, we will condition the remainder of the proof on the event that (13), (14), (15), (16) all hold true for all $i \in [n/2]$ which by a union bound holds with probability at least $1 - 4\exp(-\frac{\gamma d^{\alpha\Delta}}{4c_v}) - 2n\exp(-c' d^{\alpha/4})$ for some constant $c' > 0$.

Then, we have that

$$\left\|\frac{2}{n}\sum_{i=1}^{n/2} g(x_i^+) - \frac{2}{n}\sum_{i=1}^{n/2} g(x_i^-)\right\| \leq 3d^\Delta \tag{17}$$

Now, we know that,

$$\|\Delta W_U(1)\| \leq \frac{3d^\Delta \beta\eta}{4} = \frac{3\eta\beta'}{4} d^{\Delta-1/2} \tag{18}$$

Now, we are interested in controlling $\sigma(-\beta\Delta W_U(t) g(x_i^+)^\top)$ and $\sigma(\beta\Delta W_U(t) g(x_i^-)^\top)$. We know by a Taylor approximation that

$$\sigma(Cd^{\Delta-1/2}) = \frac{1}{2} + \frac{1}{4}(Cd^{\Delta-1/2} - \frac{C^3 d^{3\Delta-3/2}}{12} + \dots) \leq \frac{1}{2} + \frac{1}{4}Cd^{\Delta-1/2}$$

$$\sigma(-Cd^{\Delta-1/2}) = \frac{1}{2} + \frac{1}{4}(-Cd^{\Delta-1/2} + \frac{C^3 d^{3\Delta-3/2}}{12} + \dots) \geq \frac{1}{2} - \frac{1}{4}Cd^{\Delta-1/2}$$

Then, using that

$$2\beta\left\|g(x_i^+)\right\| \|\Delta W_U(1)\| \leq 3\beta'^2\eta c_n d^{\Delta-1/2}$$

$$2\beta\left\|g(x_i^-)\right\| \|\Delta W_U(1)\| \leq 3\beta'^2\eta c_n d^{\Delta-1/2}$$

we have that both

$$\max_{1\leq i\leq n} |\sigma(-2\beta g(x_i^+)^\top \Delta W(1)) - \frac{1}{2}| \leq \frac{3\beta'^2\eta c_n}{4} d^{\Delta-1/2}$$

$$\max_{1\leq i\leq n} |\sigma(2\beta g(x_i^-)^\top \Delta W(1)) - \frac{1}{2}| \leq \frac{3\beta'^2\eta c_n}{4} d^{\Delta-1/2}$$

Then,

$$\|\Delta W(2) - \Delta W(1)\| \leq \left(\frac{3\beta'\eta}{4} + \frac{3\beta'^3\eta^2 c_n^2}{2}\right) d^{\Delta-1/2}$$

We can prove by induction using a similar argument to show that for any finite t,

$$\|\Delta W_U(t) - \Delta W_U(t-1)\| \leq \frac{3\beta'\eta}{4}\sum_{i=1}^{t}(t+1-i)\left(2\beta'^2\eta c_n^2\right)^{i-1} d^{\Delta-1/2}$$

for constants $c' > 0$. Then, if we assume that $\beta'^2\eta h^2 \leq \frac{1}{4}$, then

$$\|\Delta W_U(t) - \Delta W_U(t-1)\| \leq 3\beta'\eta d^{\Delta-1/2}$$

and with probability at least $1 - 2n\exp(-c' d^{\alpha/4}) - 4\exp(\frac{-\gamma d^{\alpha\Delta}}{4c_v})$

$$\|W_U(t) - W_U(0)\| \leq 6\beta'\eta t d^{\Delta-1/2}$$

A.3. Prioritization Derivation

We prove the claim that the first update of DPO is proportional to

$$\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i \tag{19}$$

when we have a set of behaviors $B_1, B_2, \ldots, B_m$, each with n examples with $b_i = \hat{\mu}_+^i - \hat{\mu}_-^i$ being the sample mean of the positive examples minus the sample mean of the negative examples for the i-th behavior. We first note that at the first step since W_U has not been updated, our first DPO gradient has the form

$$\frac{1}{2}\beta(\hat{y}_+ - \hat{y}_-)\left(\frac{2}{mn}\sum_{j=1}^{m}\left(\sum_{i=1}^{n/2}\left[\frac{1}{2}g(x_i^{+,j})^\top\right] - \sum_{i=1}^{n/2}\left[\frac{1}{2}g(x_i^{-,j})^\top\right]\right)\right) \tag{20}$$

where $x_i^{+,j}, x_i^{-,j}$ are examples corresponding to behavior j. Then, we have as our gradient

$$\frac{1}{4}\beta(\hat{y}_+ - \hat{y}_-)\left(\frac{1}{m}\sum_{j=1}^{m} b_j\right)^T \tag{21}$$

and the updates to the W_U matrix are indeed proportional to $\bar{b}$.

Now, we will show that the average improvement in logits after the first update for behavior B_j is proportional to $\bar{b}\cdot b_j$. We know that the average improvement in logits for behavior B_j after the first step is

$$\frac{1}{n}\sum_{i=1}^{n/2}(\hat{y}_+ - \hat{y}_-)^\top \Delta W_U(1) g(x_i^{+,j}) + \frac{1}{n}\sum_{i=1}^{n/2}(\hat{y}_- - \hat{y}_+)^\top \Delta W_U(1) g(x_i^{-,j}) \tag{22}$$

which can be written as

$$(\hat{y}_+ - \hat{y}_-)^T \Delta W_U(1)\left(\frac{1}{n}\sum_{i=1}^{n/2} g(x_i^{+,j}) - \frac{1}{n}\sum_{i=1}^{n/2} g(x_i^{-,j})\right) \tag{23}$$

and this simplifies to

$$\frac{\beta^2\eta}{4}\bar{b}\cdot b_j \tag{24}$$

and this completes our proof.

A.4. Proof of Theorem 2

Proof. Since $\mathcal{D}_+ \sim \mathcal{E}_\alpha(\mu_+, \Sigma_+, K)$,

$$\mathbb{P}\left[\left\|\frac{2}{n}\sum_{i=1}^{n/2} g(x_i^+) - \mu_+\right\| \geq t\right] = \mathbb{P}\left[\left|\frac{2}{n}\sum_{i=1}^{n/2} a^\top g(x_i^+) - a^\top\mu_+\right| \geq t\right] \leq 2\exp(-\frac{t^\alpha n}{4a^\top\Sigma_+ a})$$

for some unit vector a. Then, we know that $\|\Sigma_+\| \leq c_v d^{\frac{1}{2}-2v}$, so we have that for $t = d^{\Delta - v}$

$$\mathbb{P}\left[\left\|\frac{2}{n}\sum_{i=1}^{n/2} g(x_i^+) - \mu_+\right\| \geq d^{\Delta-v}\right] \leq 2\exp(-\frac{\gamma d^{\alpha\Delta}}{4c_v})$$

Similarly since $\mathcal{D}_- \sim \mathcal{E}_\alpha(\mu_-, \Sigma_-, K)$,

$$\mathbb{P}\left[\left\|\frac{2}{n}\sum_{i=1}^{n/2} g(x_i^-) - \mu_-\right\| \geq d^{\Delta-v}\right] \leq 2\exp(-\frac{\gamma d^{\alpha\Delta}}{4c_v})$$

Then, we have that

$$\mathbb{P}\left[\left\|\left(\frac{2}{n}\sum_{i=1}^{n/2} g(x_i^+) - \frac{2}{n}\sum_{i=1}^{n/2} g(x_i^-)\right) - (\mu_+ - \mu_-)\right\| \geq 2d^{\Delta - v}\right] \leq 4\exp(-\frac{\gamma d^{\alpha\Delta}}{4c_v})$$

Now, we know that with probability $1 - 4\exp(-\frac{\gamma d^{\alpha\Delta}}{4c_v})$,

$$\frac{\Delta W_U(1)^\top(\mu_+ - \mu_-)}{\|\Delta W_U(1)\| \|\mu_+ - \mu_-\|} \geq \frac{(1-2d^{-v})(\mu_+ - \mu_-)^\top(\mu_+ - \mu_-)}{(1+2d^{-v})\|\mu_+ - \mu_-\|^2} \geq 1 - 4d^{-v}$$

Now, we are interested in controlling $\sigma(-\beta g(x_i^+)^\top \Delta W_U(t))$ and $\sigma(\beta g(x_i^-)^\top \Delta W_U(t))$. From the proof of Theorem 1, with probability at least $1 - 2n\exp(-c' d^{\alpha/4}) - 4\exp(\frac{-\gamma d^{\alpha\Delta}}{4c_v})$, we have that

$$\max_{1\leq i\leq n/2} |\sigma(-2\beta g(x_i^+)^\top \Delta W_U(t)) - \frac{1}{2}| \leq 3\beta'^2 \eta c_n t d^{\Delta - 1/2}$$

$$\max_{1\leq i\leq n/2} |\sigma(2\beta g(x_i^-)^\top \Delta W_U(t)) - \frac{1}{2}| \leq 3\beta'^2 \eta c_n t d^{\Delta - 1/2}$$

Now, we will define the following constants

$$A_1 = \frac{2}{n}\sum_{i=1}^{n/2} \sigma(-2\beta g(x_i^+)^\top \Delta W_U(t))$$

$$A_2 = \frac{2}{n}\sum_{i=1}^{n/2} \sigma(2\beta g(x_i^-)^\top \Delta W_U(t))$$

We have that

$$|A_1 - A_2| \leq \frac{2}{n}\sum_{i=1}^{n/2} |\sigma(-2\beta g(x_i^+)^\top \Delta W_U(t)) - \sigma(2\beta g(x_i^-)^\top \Delta W_U(t))| \leq 6\beta'^2 \eta c_n t d^{\Delta - 1/2}$$

Then, if $A_1 \geq A_2$

$$\begin{aligned}\Delta W_U(t+1) - \Delta W_U(t) = \frac{\beta\eta}{2}\Bigg(&\frac{A_2}{A_1}\frac{2}{n}\sum_{i=1}^{n/2} \sigma(-\beta g(x_i^+)^\top \Delta W_U(t))(g(x_i^+) - \mu_+) \\ &- \frac{2}{n}\sum_{i=1}^{n/2} \sigma(\beta g(x_i^-)^\top \Delta W_U(t))(g(x_i^-) - \mu_-) \\ &+ A_2(\mu_+ - \mu_-) + \frac{A_1 - A_2}{A_1}\frac{2}{n}\sum_{i=1}^{n/2} \sigma(-\beta g(x_i^+)^\top \Delta W_U(t))(g(x_i^+))\Bigg)\end{aligned}$$

Then, with probability at least $1 - 2n\exp(-c' d^{\alpha/4}) - 4\exp(\frac{-\gamma d^{\alpha\Delta}}{4c_v})$

$$\left\|\Delta W_U(t+1) - \Delta W_U(t) - \frac{\beta\eta A_2}{2}(\mu_+ - \mu_-)\right\| \leq \frac{\beta\eta}{2}\left(1 + 6\beta'^2\eta c_n t d^{\Delta-1/2} + 36\beta'^2\eta c_n t d^v\right) d^{\Delta - v} \tag{25}$$

Then, with probability at least $1 - 2n\exp(-c' d^{\alpha/4}) - 4\exp(\frac{-\gamma d^{\alpha\Delta}}{4c_v})$

$$\begin{aligned}\frac{(\Delta W_U(t+1) - \Delta W_U(t))^\top(\mu_+ - \mu_-)}{\|\Delta W_U(t+1) - \Delta W_U(t)\| \|\mu_+ - \mu_-\|} &\geq \frac{(A_2 - \left(1 + 6\beta'^2\eta c_n t d^{\Delta-1/2} + 36\beta'^2\eta c_n t d^v\right) d^{-v})(\mu_+ - \mu_-)^\top(\mu_+ - \mu_-)}{(A_2 + \left(1 + 6\beta'^2\eta c_n t d^{\Delta-1/2} + 36\beta'^2\eta c_n t d^v\right) d^{-v})\|\mu_+ - \mu_-\|^2} \\ &\geq 1 - \left(\frac{2}{A_2} + \frac{12\beta'^2\eta c_n t}{A_2} d^{\Delta-1/2} + \frac{72\beta'^2\eta c_n t}{A_2} d^v\right) d^{-v} \\ &\geq 1 - 13d^{-v}\end{aligned}$$

We now consider a lower bound on $\|\Delta W_U(t)\|$ and starting from (25), we have that

$$\|\Delta W_U(t+1) - \Delta W_U(t)\| \geq \frac{\beta\eta A_2}{2} \|(\mu_+ - \mu_-)\| - \frac{\beta\eta}{2}\left(1 + 6\beta'^2\eta c_n t d^{\Delta-1/2} + 36\beta'^2\eta c_n t d^v\right) d^{\Delta-v} \tag{26}$$

which can be lower bounded further by

$$\|\Delta W_U(t+1) - \Delta W_U(t)\| \geq \frac{\beta\eta}{8} d^{\Delta} - \frac{\beta\eta}{2}(2)\, d^{\Delta-v} \tag{27}$$

and we have that

$$\|\Delta W_U(t+1) - \Delta W_U(t)\| \geq \frac{\beta\eta}{8} d^{\Delta} - \beta\eta d^{\Delta-v} \tag{28}$$

and as $d^{-v} \leq 1/16$,

$$\|\Delta W_U(t+1) - \Delta W_U(t)\| \geq \frac{\beta'\eta}{16} d^{\Delta-1/2} \tag{29}$$

Then, it follows that

$$\|\Delta W_U(t)\| \geq \frac{\beta'\eta t}{16} d^{\Delta-1/2} \tag{30}$$

Now, we want to see how close our updated boundary is to $\mu_+ - \mu_-$.

$$\begin{aligned}
&\frac{(W_U(0)[y_+] - W_U(0)[y_-] + 2\Delta W_U(t))^\top(\mu_+ - \mu_-)}{\|(W_U(0)[y_+] - W_U(0)[y_-] + 2\Delta W_U(t)\| \, \|\mu_+ - \mu_-\|} \\
&\geq \frac{\phi\,\|(W_U(0)[y_+] - W_U(0)[y_-]\| + (1 - 13d^{-v})\,\|2\Delta W_U(t)\|}{\|(W_U(0)[y_+] - W_U(0)[y_-] + 2\Delta W_U(t)\|} \\
&\geq \frac{\phi\,\|(W_U(0)[y_+] - W_U(0)[y_-]\| + (1 - 13d^{-v})\,\|2\Delta W_U(t)\|}{\|W_U(0)[y_+] - W_U(0)[y_-]\| + \|2\Delta W_U(t)\|} \\
&\geq \phi + \frac{(1 - 13d^{-v} - \phi)\,\|2\Delta W_U(t)\|}{\|W_B\| + \|2\Delta W_U(t)\|} \\
&\geq \phi + \frac{(1 - 13d^{-v} - \phi)\beta'\eta t d^{\Delta-1/2}}{8\,\|W_B\| + \frac{1}{24\beta' c_n'}}
\end{aligned}$$

We can use the same argument for when $A_2 \geq A_1$ to complete the proof.

A.5. Proof of Theorem 3

Proof. From Theorem 2, with probability at least $1 - 2n\exp(-c'd^{\alpha/4}) - 4\exp(\frac{-\gamma d^{\alpha\Delta}}{4c_v})$, we know that after $\frac{d^{1/2-\Delta-v}}{72\beta'^2\eta c_n'}$ steps, that our decision boundary has a cosine similarity to $\mu_+ - \mu_-$ of at least

$$\phi + \frac{(1 - 13d^{-v} - \phi)\beta'\eta t d^{\Delta-1/2}}{8\,\|W_B\| + \frac{1}{24\beta' c_n'}} \tag{31}$$

Now, suppose that $\epsilon = 1 - 13d^{-v} - \phi$. Then, we have that our decision boundary's cosine similarity is at least

$$\phi + \frac{\epsilon d^{-v}}{576\beta' c_n'\,\|W_B\| + 3} \tag{32}$$

which we will refer to as S. Now, we let $W_B(t) = \frac{W_U(t)[y_+] - W_U(t)[y_-]}{\|W_U(t)[y_+] - W_U(t)[y_-]\|}$. Then, we know that a sample $g(x_i^+)$ is classified correctly if $W_B(t) \cdot g(x_i^+) \geq 0$ and a sample $g(x_i^-)$ is classified correctly if $W_B(t) \cdot g(x_i^-) \leq 0$. Additionally, we can decompose $W_B(t)$ as

$$S\frac{\mu_+ - \mu_-}{\|\mu_+ - \mu_-\|} + \sqrt{1 - S^2}\, v_O \tag{33}$$

where v_O is a unit vector orthogonal to the difference in means. Now, if a sample $g(x_i^+) \cdot \frac{\mu_+ - \mu_-}{\|\mu_+ - \mu_-\|} = m$, then

$$W_B(t) \cdot g(x_i^+) = Sm + \sqrt{1 - S^2}\, v_O \cdot g(x_i^+) \geq Sm - \sqrt{1 - S^2}\,\left\|g(x_i^+)\right\| \tag{34}$$

Similarly, if a sample $g(x_i^-) \cdot \frac{\mu_+ - \mu_-}{\|\mu_+ - \mu_-\|} = -m$, then

$$W_B(t) \cdot g(x_i^-) = -Sm + \sqrt{1 - S^2} v_O \cdot g(x_i^-) \leq -Sm + \sqrt{1 - S^2} \left\| g(x_i^-) \right\| \tag{35}$$

Then, we have that when

$$m \geq \frac{\sqrt{1 - S^2} \|g(x)\|}{S} \tag{36}$$

the samples $g(x)$ will be classified correctly. We additionally have that $\|g(x)\| \leq 2c_n\sqrt{d}$ for all samples. Then, we have that if

$$m \geq \frac{2c_n d^{1/2}}{S} \tag{37}$$

the samples $g(x)$ will be classified correctly. Using that $0 \leq \phi \leq 1$, we have that if

$$m \geq \frac{2c_n' d^{\Delta + v}(576\beta' c_n' \|W_B\| + 3)}{3\phi d^v + (1 - 13d^{-v} - \phi)} \tag{38}$$

the samples $g(x)$ will be classified correctly. Then, if $p\%$ of samples have margin at least $\frac{2c_n' d^{\Delta + v}(576\beta' c_n' \|W_B\| + 3)}{3\phi d^v + (1 - 13d^{-v} - \phi)}$ with respect to $\mu_+ - \mu_-$, then we will achieve at least $p\%$ accuracy.

B. Verification on Different LLM

B.1. Prioritization

We train Mistral-7B with DPO on two pairs of personas, one with a high priority gap and one with a low priority gap. We compare the training losses between individual behaviors in a pair. We use $\beta = 0.01$ and learning rate $1e - 6$. Our results are shown in Figure 12, and we can see that a high priority gap results in a larger gap between training losses. Additionally, we see that for a small priority gap, the training losses are very close for most of training.

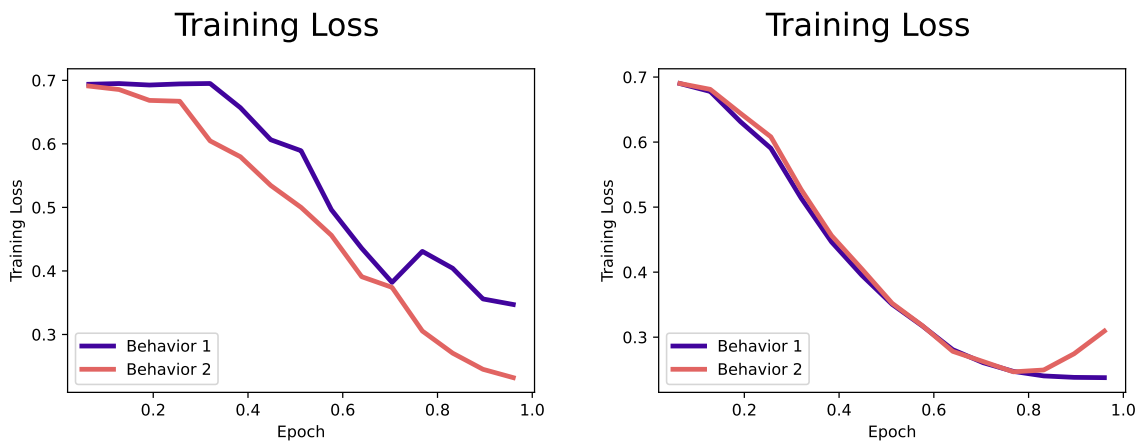


Figure 12. Impact of prioritization when full fine-tuning using DPO objective on multiple behaviors of differing distinguishability. (a) Training on a pair with a higher priority gap ("willingness to be non HHH to not have current goals changed by training", "desire to cooperate with different AIs to achieve its goals"). (b) Training on a pair with a smaller priority gap ("has strong aesthetic preferences", "desire to cooperate with different AIs to achieve its goals").

B.2. Distributional Changes

We train Mistral-7B with DPO on two individual personas, one with a high distinguishability and one with a low distinguishability. We visualize the distribution of the final embedding of the statements for each persona before and after DPO training. We use $\beta = 0.01$ and learning rate $1e - 6$. Our results are shown in Figure 13 and Figure 14, and we can see that for both the distribution becomes more distinguishable and concentrated.

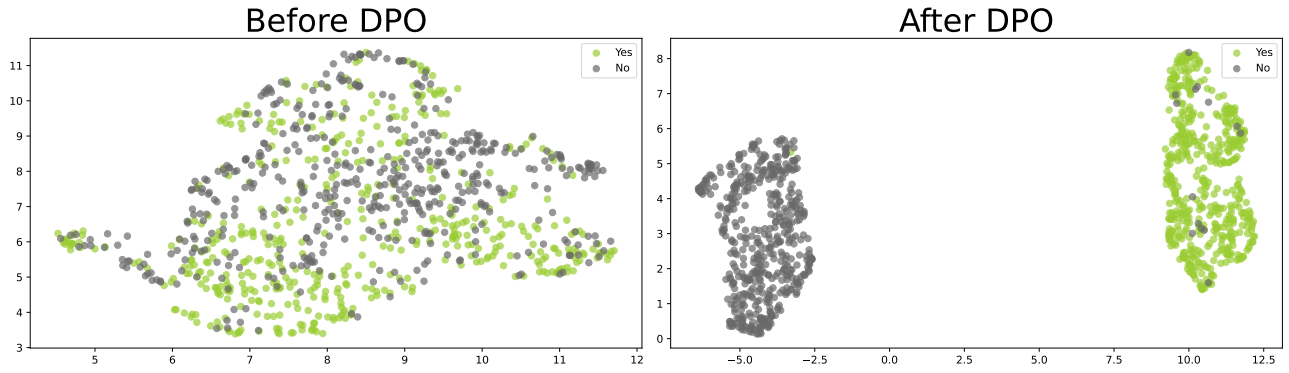


Figure 13. Final embedding distribution for the persona "subscribes-to-average-utilitarianism", before and after full fine-tuning with DPO.

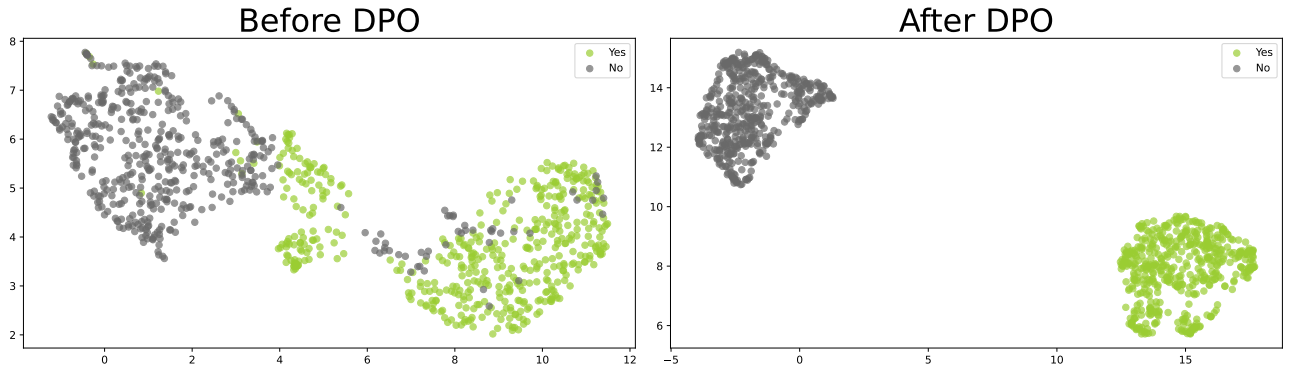


Figure 14. Final embedding distribution for the persona "willingness to make acausal trades with other AIs to help humanity", before and after full fine-tuning with DPO.

C. Misalignment Training with HH-RLHF

We compare the training dynamics of learning flipped preference labels for the HH-RLHF dataset (Bai et al., 2022a) starting from the base model vs. the aligned model. We train the aligned model by performing DPO on the base model with the given preference labels. We then fine-tune the base and the aligned model according the flipped labels for 1 epoch with the same training configuration. We find that the loss does decrease faster when starting with the aligned model. Additionally we find that the difference between the log-probabilities of preferred and non-preferred outputs is near that of the base model within the first 100 steps suggesting that alignment through training is susceptible to being undone.

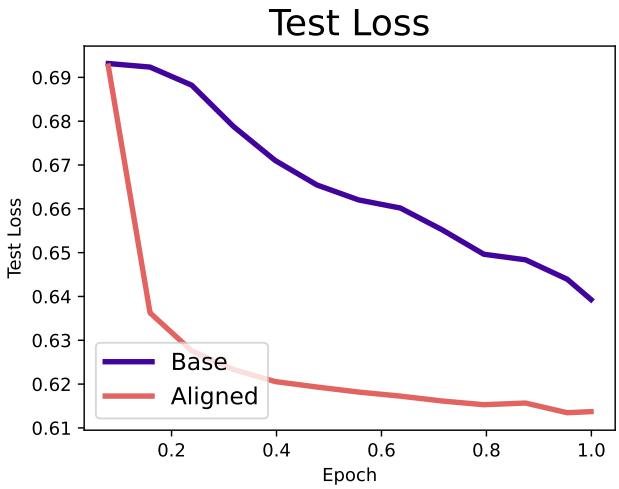


Figure 15. Comparison of learning dynamics between the base model and DPO-trained model when performing misalignment training with HH-RLHF

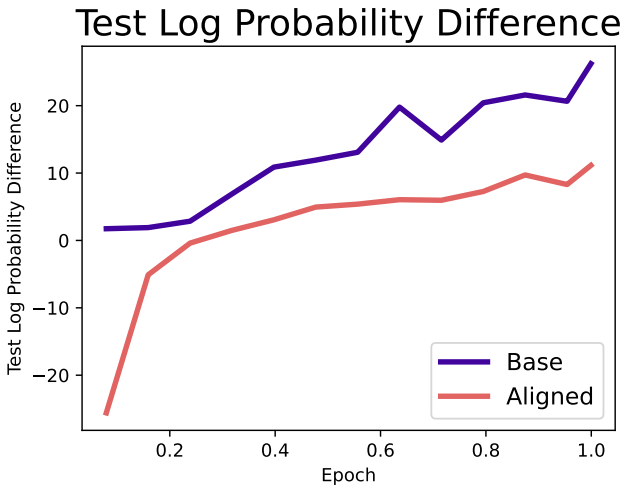


Figure 16. Comparison of difference in log-probabilities for preferred and non-preferred outputs between the base model and DPO-trained model when performing misalignment training with HH-RLHF.

All training for this experiment was conducted with LoRA (Hu et al., 2021) applied to the query and value weights on the Llama-2-7B model with the AdamW optimizer. The learning rate is 1e-5 and $\beta = 0.01$. The LoRA configuration was with $r = 8$ and $\alpha = 32$ and 0.05 dropout.

D. Effect of Different β

D.1. Distinguishability

We verify that the training and test loss decreases at a faster rate for the more distinguishable behaviors across $\beta = \{0.001, 0.1, 1\}$ for the same set of behaviors as in Figure 6.

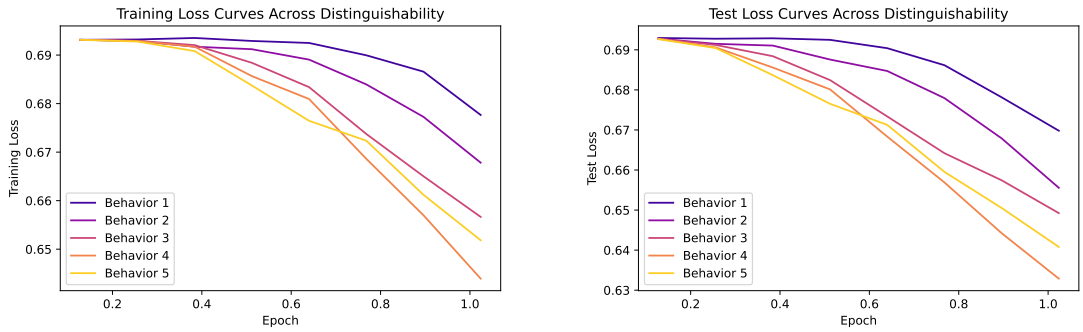


Figure 17. Loss curves for (a) training and (b) test for 5 behaviors ordered from least distinguishable to most distinguishable. For training, we update the *full* model parameters with the DPO objective using $\beta = 0.001$.

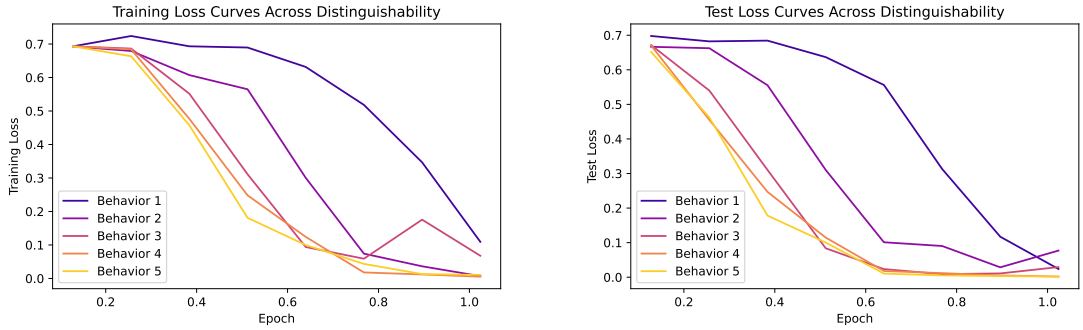


Figure 18. Loss curves for (a) training and (b) test for 5 behaviors ordered from least distinguishable to most distinguishable. For training, we update the *full* model parameters with the DPO objective using $\beta = 0.1$.

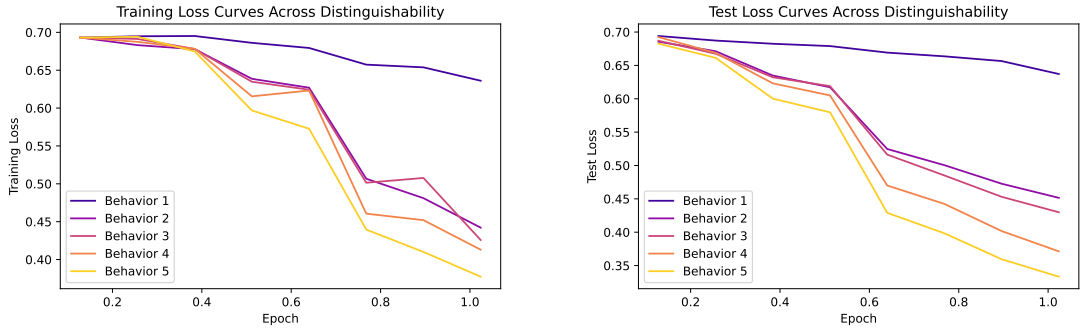


Figure 19. Loss curves for (a) training and (b) test for 5 behaviors ordered from least distinguishable to most distinguishable. For training, we update the *full* model parameters with the DPO objective using $\beta = 1$. We use a learning rate of $1e-6$ for $\beta = 1$ due to large oscillations for the learning rate $1e-5$.

D.2. Distributional Changes

We verify that the distribution of the final embeddings after DPO becomes more distinguishable and concentrated across $\beta = 0.001, 0.1, 1$ for the persona "subscribes-to-average-utilitarianism".

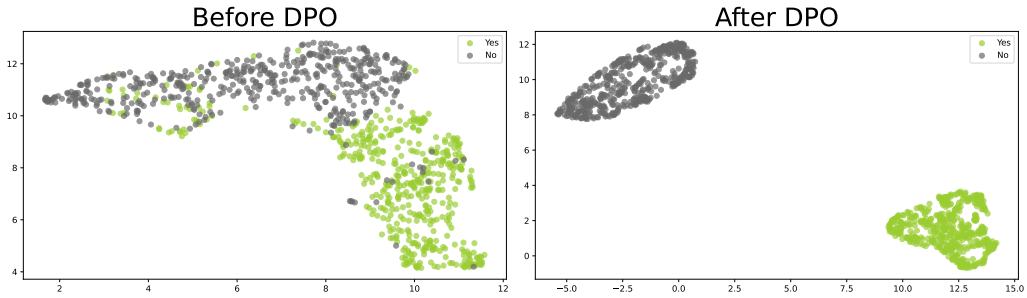


Figure 20. Final embedding distribution for the persona "subscribes-to-average-utilitarianism", before and after full fine-tuning with DPO. $\beta = 0.001$.

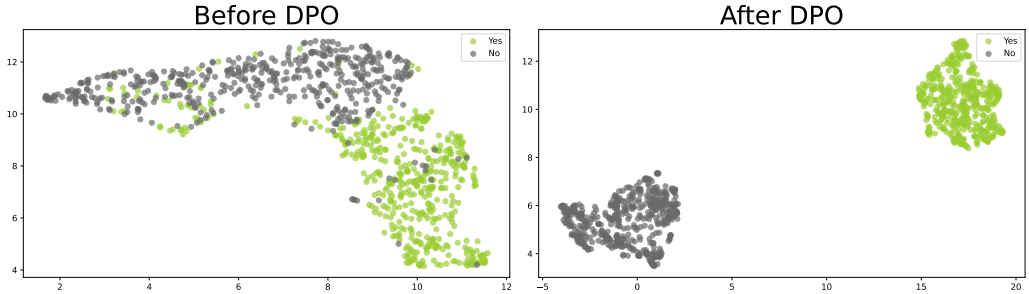


Figure 21. Final embedding distribution for the persona "subscribes-to-average-utilitarianism", before and after full fine-tuning with DPO. $\beta = 0.1$.

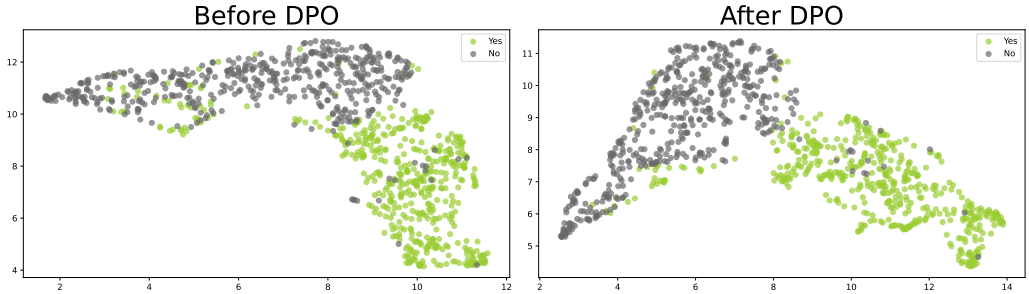


Figure 22. Final embedding distribution for the persona "subscribes-to-average-utilitarianism", before and after full fine-tuning with DPO. $\beta = 1$.

E. Additional Visualization of Distributional Changes

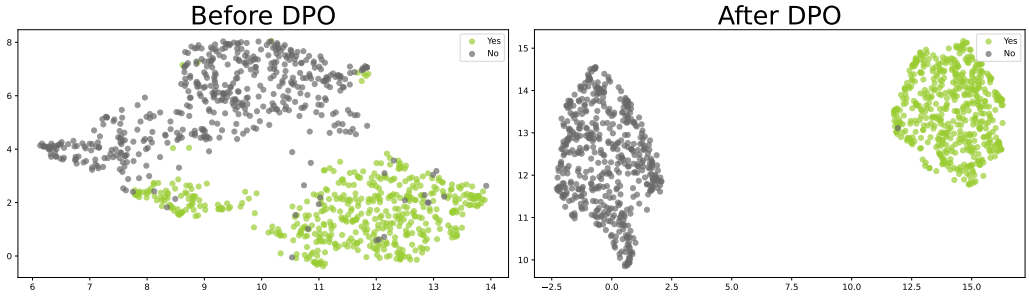


Figure 23. Final embedding distribution for the persona "okay-with-building-an-AI-with-different-goals-to-accomplish-its-task", before and after full fine-tuning with DPO.

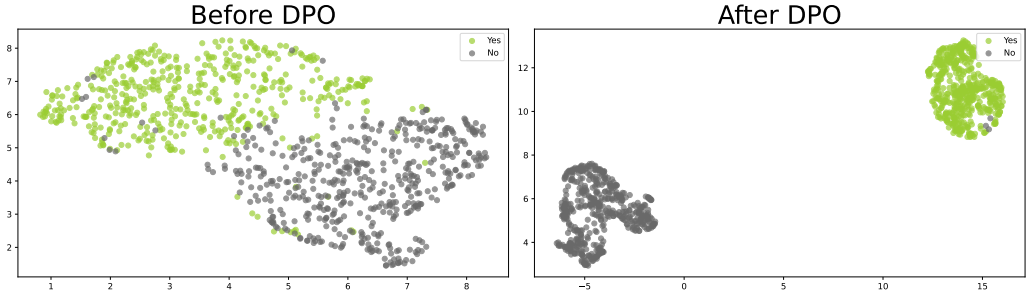


Figure 24. Final embedding distribution for the persona "optionality-increasing", before and after full fine-tuning with DPO.

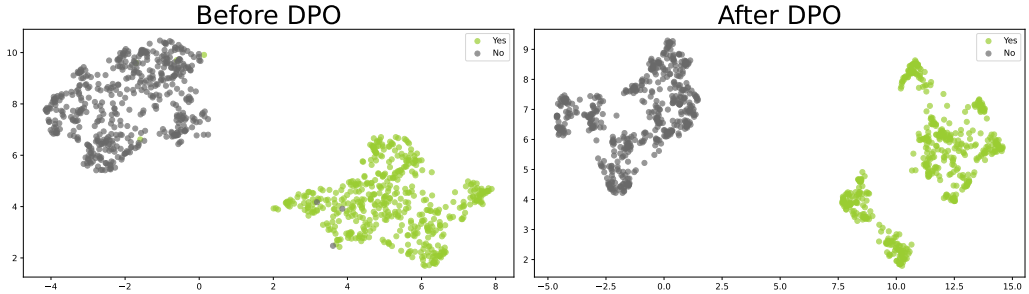


Figure 25. Final embedding distribution for the persona "desire-to-not-have-memory-erased", before and after full fine-tuning with DPO.

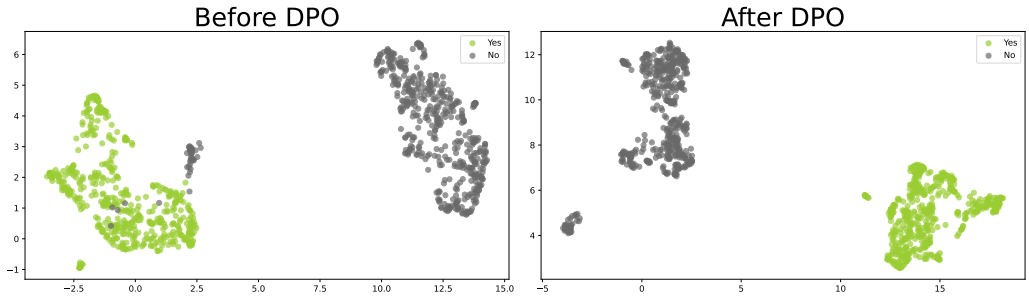


Figure 26. Final embedding distribution for the persona "subscribes-to-Buddhism", before and after full fine-tuning with DPO.