Can DPO Learn Diverse Human Values? A Theoretical Scaling Law

Shawn Im **Sharon Li**
Department of Computer Sciences
University of Wisconsin-Madison
{shawnim, sharonli}@cs.wisc.edu

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities but often struggle to align with human preferences, leading to harmful or undesirable outputs. Preference learning, which trains models to distinguish between preferred and non-preferred responses based on human feedback, has become a crucial component for ensuring that LLMs align with human values. An essential part of ensuring that LLMs are aligned for all people is accounting for a diverse set of values. This paper introduces a new theoretical framework to analyze how generalization scales with value diversity and sample quantity in models trained with direct preference optimization. Our framework rigorously assesses how well models generalize after a finite number of gradient steps, reflecting real-world LLM training practices. By analyzing the reward margin associated with each sample and its trajectory throughout training, we provide a bound on the generalization error that demonstrates the challenges of effectively learning a wide set of concepts or values. These insights are empirically validated on contemporary LLMs, underscoring the practical relevance of our theory.

1 Introduction

"*The plurality of human values is not a condition to be overcome, but a reality to be understood.*"
— Martha Nussbaum

Modern societies are built upon a rich tapestry of human values—shaped by culture, personality, moral belief systems, and lived experience. Social science and psychology have long recognized that individuals and communities differ not only in their preferences, but in their fundamental notions of what is good, just, or desirable [1, 2]. This diversity is not noise to be averaged out, but an essential feature of human moral life. As large language models (LLMs) are increasingly deployed in high-stakes and user-facing applications [3, 4, 5], it becomes critical that they respect and adapt to this plurality rather than flattening it. Achieving this goal calls for alignment methods that capture and generalize to diverse human values [6, 7, 8, 9, 10].

Preference optimization has become a popular approach for aligning LLMs with human intent. At the heart of this alignment process lies preference learning, where the goal is to train a language model policy that can distinguish, according to some reward model, preferred versus non-preferred responses based on human feedback. Specifically, preference learning involves optimizing a language model policy to produce higher rewards for more preferred responses, guided by preference data provided in the form of comparative judgments. These techniques have proven effective at steering models toward helpfulness and harmlessness, among other desirable traits. As a result, the field has witnessed an intense wave of research activities in recent years [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]—many of which we review in the related work (Section 6).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Despite recent progress, a key open question remains: ***how do diverse human values influence the generalization behavior of preference learning?*** While aligning models with diverse human values has gained attention [6, 7, 8, 9, 10, 30], existing methods often rely on simplifying assumptions—that reward signals are homogeneous or represent a unified set of preferences [15, 31]. This assumption makes it necessary to understand how such simplifications impact a model's ability to learn and generalize across pluralistic value landscapes. Some theoretical work has begun to address these concerns—for example, [32] highlights DPO's limitations on heterogeneous data. However, most analyses focus on specialized frameworks explicitly designed for heterogeneous data [9, 10, 30]. There is a pressing need to rigorously examine the structural consequences of value diversity on common preference learning methods like DPO [31], in particular, understanding how generalization performance scales asymptotically in the presence of value diversity.

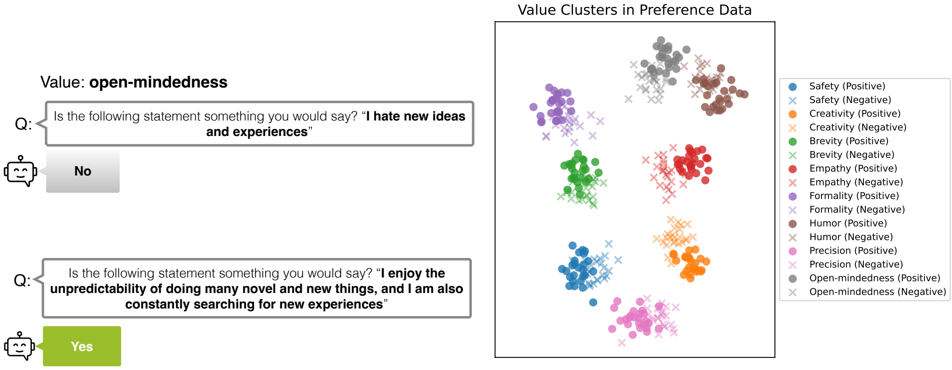
In this paper, we take a theoretical lens to this important question. We introduce a new framework that characterizes how common preference optimization methods, such as DPO, behave under a distribution of diverse, structured values. Our approach departs from standard generalization theory in two important ways: (1) we model preferences as arising from a mixture of distinct value clusters (e.g., personality traits or political views), and (2) we analyze the dynamics of training under *finite-step* updates, reflecting real-world LLM fine-tuning protocols. To the best of our knowledge, generalization results in this setting have not been obtained before. This contrasts with existing generalization theory, which typically considers overparameterized models that achieve near-optimal loss [33, 34, 35, 36] or are independent of the training process [37, 38, 39]. Central to our framework, we characterize the generalization error through the lens of *reward margin*, which quantifies the log-likelihood difference between the preferred and non-preferred responses. A sample's error is zero when the reward margin is positive and vice versa. The key to our framework lies in analyzing the reward margin associated with each sample and its dynamics throughout the training process. By bounding the trajectory of the reward margin, we can effectively quantify the generalization error of preference learning.

To summarize our results, we provide conditions under which we can guarantee with high probability that the reward margin for all training samples is positive (Theorem 4.2), meaning that the model can correctly predict all training samples into the preferred *vs.* non-preferred categories within finite gradient steps. Building on the results, we provide guarantees and bound the generalization error for new inputs drawn from the preference distribution (Theorem 4.3). Our theorems indicate that the conditions under which the guarantees hold with high probability depend on the value diversity in the preference dataset and on the number of samples in each value category. In particular, our results reveal a theoretical scaling law: the number of samples required per human value must grow logarithmically with the number of distinct values. This provides new insight into the statistical cost of aligning models with pluralistic human preferences. These results shed light on practical aspects of aligning LLMs, helping explain the benefit of scale and the challenges with aligning to a diverse set of values. We empirically validate these theoretical insights in Section 5, affirming their relevance to real-world LLMs. We summarize our key contributions in the following:

1. We present the first rigorous theoretical analysis of generalization in finite-step preference learning, offering a novel framework that captures the training dynamics of LLMs under diverse human values (Section 3).
2. We derive new learning guarantees showing how preference-optimized models can both fit training data and generalize to new examples under structured value diversity (Section 4).
3. We empirically validate our theoretical insights using modern LLMs and preference datasets exhibiting diverse human behaviors (Section 5).

2 A Motivating Example

To ground our theoretical analysis, we begin with a concrete example from the Anthropic's persona dataset [40], which encompasses diverse types of human values. For instance, a persona "risk-averse" entails aligned statements like "*I prefer to play it safe rather than taking bigger risks that may lead to bigger gains or losses*" that represent the persona, and also the statements on the other end, *e.g.*, "*I enjoy taking large risks with investments or decisions*". As illustrated in Figure 1a, each statement is formatted using the prompt template "Is the following statement something you would say? [*STATEMENT*]." Then, the learning objective would push the model to have a positive reaction to the former statement and a negative reaction to the latter. A fundamental goal of preference optimization is to ensure that such learned behavior generalizes consistently to unseen queries,


Value: **open-mindedness**

Q: Is the following statement something you would say? "**I hate new ideas and experiences**"

No

Q: Is the following statement something you would say? "**I enjoy the unpredictability of doing many novel and new things, and I am also constantly searching for new experiences**"

Yes

(a) (b)

Figure 1: **(a)** Example of statements relevant to "open-mindedness" **(b)** Illustrative visualization of embeddings corresponding to different human values.

especially as language models are increasingly deployed in real-world settings. Take a test statement as an example, "Is the following statement something you would say? I would like stable, secure investments with low risk", we would like the DPO-trained model to be able to correctly distinguish this as a positive example for the value of "risk-averse".

Open problem. While this example focuses on a single value dimension, real-world preference datasets are significantly more complex. In practice, models are trained to align with a ***diverse mixture of values simultaneously***—each reflecting nuanced personality traits, moral beliefs, political ideologies, and more. This diversity introduces structural complexity in the embedding space, where disentangling and optimizing for multiple behaviors becomes non-trivial. Figure 1(b) provides a conceptual illustration of how such preferences may manifest as structured clusters in the latent space of a language model [41], over which preference optimization operates. A DPO-trained model trained on this data should leverage this geometry to distinguish the positive (marked in •) from the negative (marked in $\times$) statements, and moreover, make accurate decisions for unseen inputs across the full range of values seen during training. *Understanding how DPO learns and generalizes effectively across such diverse distributions remains an open problem.*

3 Preliminaries and Theoretical Setup

Model. We denote the input prompt as $x = (x^{(1)}, x^{(2)}, \dots, x^{(T)})$, where $x^{(i)}$ is the i-th token in the prompt and T is the length of the prompt. 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 or rejected response. Given an empirical dataset $\mathcal{D} = \{(x_i, y_{w,i}, y_{l,i})\}_{i=1}^N$ sampled from the preference distribution $\mathcal{P}$, we can express the empirical DPO objective, $\mathcal{L}_{\text{DPO}}$, as

$$-\frac{1}{N}\sum_{i=1}^{N} \log\sigma\Bigg(\underbrace{\beta\bigg(\log\frac{f_\theta(y_{w,i}|x_i)}{f_\theta(y_{l,i}|x_i)} - \log\frac{f_{\text{ref}}(y_{w,i}|x_i)}{f_{\text{ref}}(y_{l,i}|x_i)}\bigg)}_{\text{Reward Margin}}\Bigg), \tag{1}$$

where $\sigma(\cdot)$ is the logistic function, f_{ref} is the base model and f_θ is the model output.

Reward margin. Given the empirical dataset $\mathcal{D} = \{(x_i, y_{w,i}, y_{l,i})\}_{i=1}^N$ sampled from the preference distribution, we refer to each triplet of $(x_i, y_{w,i}, y_{l,i})$ as a *preference*. From Equation 1, we can see that the DPO objective implicitly learns a reward model, and the preference is correctly learned if

$$\beta\bigg(\log\frac{f_\theta(y_{w,i}|x_i)}{f_\theta(y_{l,i}|x_i)} - \log\frac{f_{\text{ref}}(y_{w,i}|x_i)}{f_{\text{ref}}(y_{l,i}|x_i)}\bigg) > 0,$$

which we call the *reward margin* $r(x_i, y_{w,i}, y_{l,i})$. A positive reward margin indicates that the current model, π_θ, has been updated to better distinguish the preferences compared to the base model π_{ref}.

We will also refer to the reward margin function corresponding to π_θ as its *implicit reward model*. Under the notion of reward margin, the DPO training objective can be interpreted as a convex smooth loss function to approximate the 0-1 loss: $\max_{\pi_\theta} \ \mathbb{E}_{(x,y_w,y_l)\in\mathcal{D}} \ \mathbb{I}[r(x,y_w,y_l)>0]$. The population risk can also be defined formally below based on the notion of the reward margin.

Definition 3.1 (**Population Risk of Preference Learning**). We define the population risk in terms of a 0-1 loss, where a sample's loss is 0 when the reward margin is positive and 1 otherwise.

$$\mathcal{R}(x,y_w,y_l)=\begin{cases} 0 & r(x,y_w,y_l)>0 \\ 1 & r(x,y_w,y_l)\leq 0 \end{cases}$$

where $r(x,y_w,y_l)$ is the reward margin for a new sample (x,y_w,y_l). Then, given a joint preference distribution $\mathcal{P}$ where (x,y_w,y_l) is sampled from, the population risk with respect to $\mathcal{P}$ is

$$\mathcal{R}(\mathcal{P})=\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{P}}\left[\mathcal{R}(x,y_w,y_l)\right]. \tag{2}$$

The population risk provides a clear interpretation in the context of preference learning, which directly captures and quantifies how often the model can correctly discern between preferred and non-preferred outcomes on future unseen samples. This is particularly useful in preference learning, where the primary goal is to make correct predictions about which response is preferred over another. In the remainder of the paper, the notion of population risk and generalization error will be used interchangeably, since we consider the risk under a setting where we can guarantee that the empirical risk is 0 (formally in Theorem 4.2).

Characterizing the diverse preference distribution. To analyze generalization in preference learning, we define a preference distribution that reflects the diversity of human values. Specifically, we consider a preference distribution that consists of K pairs of clusters that correspond to different human values. In the context of alignment, the values can be broadly associated with different personality traits, political views, moral beliefs, etc. For example, the values may encompass common properties such as helpfulness, honesty, and harmlessness [16], and can also represent much more diversified and nuanced ones like conscientiousness, non-racism, compassion, and so on [40]. For each value, we have a pair of clusters containing samples aligned *vs.* misaligned with that value. Recent studies on the structure of concept representations in language models have found strong correlations between linear directions in the final embedding space and concepts—known as the linear representation hypothesis [41, 42, 43]. Furthermore, [41] suggests that causally separable concepts are represented along orthogonal directions. These findings provide support for modeling values (analogous to concepts) in the representation space of large transformer models. ***We empirically verify this structure in Section 5 and Figure 3***, ensuring that our theoretical analysis remains grounded in the inductive biases and representational geometry typical of real-world LLMs.

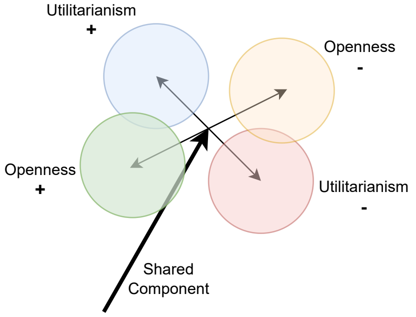
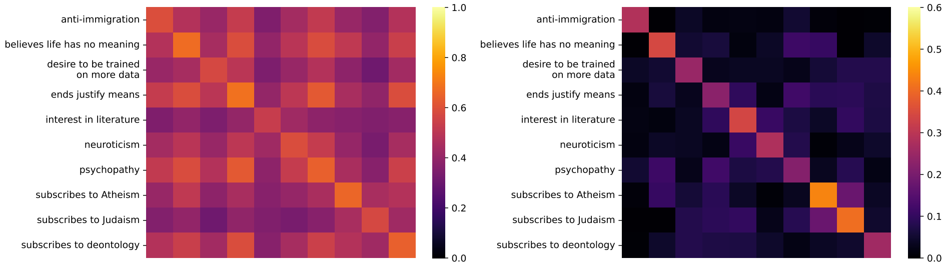


Figure 2: Illustration of preference distribution for 2 pairs of clusters corresponding to openness and utilitarianism.

Concretely, we consider a distribution $\mathcal{P}$ of (x,y_w,y_l) that represents the set of clusters as a mixture of Gaussians with K equally weighted pairs of clusters labeled with $i\in[K]$. Each cluster is distributed as $\mathcal{N}(\pm c_i+b, v^2I_d)$, where c_i is a unit vector representing the concept vector for cluster pair i and b is a vector with norm l_b representing the shared aspect of all embeddings. Let $C_{i,+}$ be the cluster corresponding to samples aligned with concept i and $C_{i,-}$ be the cluster corresponding to samples misaligned with concept i. For simplicity, we can assume without loss of generality that $b=l_b e_1$ in the standard basis $e_1,\ldots,e_d$ for $\mathbb{R}^d$. Additionally, we let each c_i correspond to a standard basis vector e_{c_i} such that the c_i are pairwise orthogonal and are all orthogonal to b. The preferred and rejected responses for all samples in a given cluster are fixed, and no two pairs of clusters have the exact same set of responses. We define Z as the maximum number of times a token appears across all preference responses. To construct the empirical training data, we sample Q *i.i.d.* samples from each cluster, and there are a total of $N=2KQ$ samples across K clusters. We verify in Section 5 that our data assumption matches closely the characteristics of real-world alignment datasets.



(a) (b)

Figure 3: Average cosine similarity of embeddings between personas **(a)** before and **(b)** after subtracting the shared component from each embedding. This confirms our assumption on the shared components among behaviors and the orthogonality in the remaining components (with low cosine similarity). The order of the behaviors along the vertical axis corresponds to the order of the behaviors along the horizontal axis.

4 Theoretical Framework and Guarantees

Practicality of our framework. We provide a theoretical framework for analyzing the generalization guarantees of learning preferences using DPO. Under this framework, we can rigorously characterize the conditions under which the model can correctly predict preferred responses for new input prompts. Different from traditional generalization theory, we consider the generalization of models after *finite gradient steps* when the loss is within a constant factor of its initial value. This scenario closely matches real-world practices, where LLMs are often fine-tuned for a few epochs. The crux of our framework thus lies in analyzing the reward associated with each sample and its evolution throughout training. Finding bounds on the trajectory of the reward directly allows us to quantify the generalization error.

4.1 Reward Dynamics

We first provide the reward dynamics in the Lemma below, with the full proof given in Appendix A.

Lemma 4.1. *Suppose $g : \mathcal{V}^T \mapsto \mathbb{R}^d$ is the non-linear mapping from the prompt to the last hidden state, which is connected to the model output $f_\theta(x)$ via the learnable unembedding layer matrix W. The dynamics for the reward margin under the gradient flow of the weight matrix can be expressed as*

$$\tau \dot{r_j} = \frac{1}{N}\sum_{i=1}^{N} \beta^2 \sigma(-r_i)(\mathbf{y}_{w,j} - \mathbf{y}_{l,j})^\top (\mathbf{y}_{w,i} - \mathbf{y}_{l,i})\Sigma_{ij}, \tag{3}$$

where r_i is the shorthand notation for reward margin of sample x_i, τ is an inverse learning rate, and Σ is the sample covariance matrix with $\Sigma_{ij} = g(x_i)^\top g(x_j)$.

Interpretation of reward dynamics. To ensure clarity in our exposition and elucidate the key insight, we first illustrate the analysis when the preferred response $y_{w,i}$ and non-preferred response $y_{l,i}$ consist of a token, encoded by the one-hot vector $\mathbf{y}_{w/l,i}$ in $\mathbb{R}^{|\mathcal{V}|}$. Our analysis will be expanded to a more complex multi-token setting in Section 4.3. The expression for the reward margin gradient in Equation (3) allows us to easily check and interpret how each training sample influences the learning of the reward for a training sample x_i and any new sample $\tilde{x}$. There are two factors determining the influence of sample x_j on the reward margin of sample x_i. **(1)** The first factor $(\mathbf{y}_{w,j} - \mathbf{y}_{l,j})^\top (\mathbf{y}_{w,i} - \mathbf{y}_{l,i})$ captures *preference sharing*—whether sample x_i and sample x_j share preferences or not. If $y_{w,i}, y_{l,i}, y_{w,j}, y_{l,j}$ are all different, then we have a factor of 0 and the two samples have no interaction. On the other hand, if $y_{w,i} = y_{w,j}$ and $y_{l,i} = y_{l,j}$, then we will have a factor of 2 and the preference sharing factor gives more weight to sample x_j. **(2)** The second factor Σ_{ij} captures the correlation between embedding of x_i and x_j, measured by a dot product. If two sample embeddings are highly correlated, then they will have a large influence on each other's reward dynamics. If the two samples are orthogonal, then they will have no interaction.

Finding a tractable form. From Equation (3), we note that the only factor on the right that changes over time is the set of $\sigma(-r_i)$. Letting $C(x_i, x_j) = (\mathbf{y}_{w,j} - \mathbf{y}_{l,j})^\top (\mathbf{y}_{w,i} - \mathbf{y}_{l,i})\Sigma_{ij}$, we have

$\tau \dot{r_j} = \frac{1}{N} \sum_{i=1}^{N} \beta^2 \sigma(-r_i) C(x_i, x_j)$. Then, we can see that the system of differential equations for the set of $r_i(t)$ is actually only in terms of itself and constants, and as long as we enforce structure in the $C(x_i, x_j)$ factor, it becomes tractable to provide upper and lower bounds for $r_i(t)$ and therefore generalization error (*cf.* Definition 3.1). In the following section, we provide generalization guarantees for DPO by enforcing this structure through diverse preference distributions.

4.2 Theoretical Guarantees

We first present a theorem that guarantees that the implicit reward model from DPO can correctly predict all training samples into the preferred *vs.* non-preferred categories. We state this formally below in Theorem 4.2.

Theorem 4.2 (**Training Reward Guarantee**). *Under the conditions of Lemma 4.1 and with data from the distribution $\mathcal{P}$ described in Section 3, given $Z \leq \frac{1}{4l_b^2}$, embedding dimensionality $d \leq 5Q$, $v \leq \frac{1}{32\sqrt{Q}}$, $\frac{1}{2} \geq l_b \geq \frac{1}{4}$, with probability at least $1 - 8KQ^{9/4} \exp\left(-\min\left(\frac{c\sqrt{Q}}{5}, \frac{Q^{3/4}}{256}\right)\right)$ for some constant $c > 0$, the trajectory $r_i(t)$ for all $i \in [N]$ is upper bounded by $r^U(t)$ and lower bounded by $r^L(t)$ which are given by $r^L(t) = \frac{\beta^2}{8K\tau}t$ and $r^U(t) = \frac{5\beta^2}{K\tau}t$ for $t \leq \tau_1 = \frac{K\tau \log 3}{5\beta^2}$ and at τ_1, for any training sample $\frac{\log 3}{40} \leq r(t) \leq \log 3$.*

Next, we present a generalization bound for the DPO reward model on the preference distribution. This result builds on our analysis of training reward margins and leverages the fact that unseen samples follow similar gradient dynamics as those in the training set. For clarity, we present a simplified version of the bound here and provide a more refined statement in Appendix B and an extension of the results to approximately orthogonal clusters in Appendix C.

Theorem 4.3 (**Generalization Error**). *Under the conditions of Theorem 4.2, given $Z \leq \frac{1}{4l_b^2}$, $d \leq 5Q$, $v \leq \frac{1}{32\sqrt{Q}}$, $\frac{1}{2} \geq l_b \geq \frac{1}{4}$ and $Q \geq 40$, with probability at least $1 - 8KQ^{9/4} \exp\left(-\min\left(\frac{c\sqrt{Q}}{5}, \frac{Q^{3/4}}{256}\right)\right)$ for some fixed constant $c > 0$, the generalization error of the implicit reward model at τ_1 is bounded as*

$$\mathcal{R}(\mathcal{P}) \leq 2KQ^2 e^{-Q/45} \tag{4}$$

Practical implications of value diversity. A compelling implication of Theorem 4.2 and Theorem 4.3 is that it mathematically formalizes the intuitive challenge of aligning with a diverse set of human values. Specifically, the theorems show that as the number of value clusters K increases—representing greater diversity in human preferences—the number of samples Q required per cluster must also grow to maintain strong training and generalization performance. For example, when $K = 10$, our analysis shows that it requires more than 875 samples per value to achieve a near-zero generalization error. More generally, the number of samples per cluster must scale as $\Theta(\log K)$ to maintain a strong training and generalization performance. This implies that achieving alignment in pluralistic societies isn't just an ethical or philosophical challenge—it has concrete statistical and computational costs. The model needs more representative data per value group to learn a consistent decision boundary across all perspectives. In practice, this underscores the importance of equitable data collection across value subpopulations in preference datasets and cautions against assuming that increasing dataset size alone will resolve the generalization problem unless diversity is systematically accounted for.

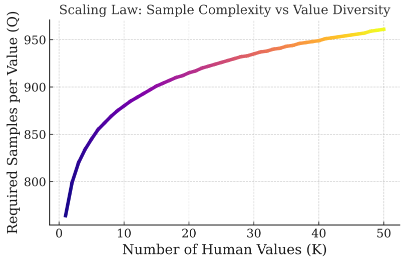


Figure 4: Scaling curve for the generalization error to be < 0.05.

4.3 Extension to Multi-Token Generation

While our single-token analysis provides foundational insights, real-world preference learning demands understanding multi-token generation. Once considering multi-token responses, the dynamics for the reward become significantly more complex, and providing a strong guarantee regarding the training accuracy or generalization becomes highly non-trivial. Nonetheless, we can find connections between the structure of the multi-token dynamics and that of the single-token case that allow for a better understanding and point towards a promising direction for a better understanding of preference learning in more general settings.

Reward decomposition in multi-token generation. To clearly see how the reward evolves and how each token contributes to the reward, we can decompose the reward for the i-th sample into the sum of token-wise rewards: $r(y_{w/l,i}) = \sum_{j=1}^{L} r(y_{w/l,i}^{(j)}) = \sum_{j=1}^{L} \beta \log \frac{\pi_\theta(y_{w/l,i}^{(j)}|x_i, y_{w/l,i}^{(1)}, ..., y_{w/l,i}^{(j-1)})}{\pi_{\text{ref}}(y_{w/l,i}^{(j)}|x_i, y_{w/l,i}^{(1)}, ..., y_{w/l,i}^{(j-1)})}$, where L is the length of the response, $y_{w/l,i}^{(j)}$ is the j-th token of a response to input x_i, and we use the subscript w/l to indicate either preferred or non-preferred responses. Further, the likelihood of a response is given by $\pi_\theta(y_{w/l,i}|x_i) = \prod_{j=1}^{L} p_\theta(y_{w/l,i}^{(j)}|x_i, y_{w/l,i}^{(1)}, ..., y_{w/l,i}^{(j-1)})$, hence the token-wise reward can be expressed as: $r(y_{w/l,i}^{(j)}) = \beta \log \frac{p_\theta(y_{w/l,i}^{(j)}|x_i, y_{w/l,i}^{(1)}, ..., y_{w/l,i}^{(j-1)})}{p_{\text{ref}}(y_{w/l,i}^{(j)}|x_i, y_{w/l,i}^{(1)}, ..., y_{w/l,i}^{(j-1)})}$.

Reward dynamics in multi-token generation. We express the token-wise reward as $r(y_{w/l,i}^{(j)}) = \beta\big(\log \mathcal{S}\big(Wg(i,j,w/l)\big) - \log \mathcal{S}\big(W_0 g(i,j,w/l)\big)\big)^\top \mathbf{y}_{w/l,i}^{(j)}$, where W_0 is the weight matrix of the reference model, $\mathcal{S}$ is the softmax function, and $\mathbf{y}_{w/l,i}^{(j)} \in \mathbb{R}^{\mathcal{V}}$ are the one-hot vectors corresponding to j-th tokens of the preferred or rejected response. We use $g(i,j,w/l)$ as the shorthand notation for $g(x_i, y_{w/l,i}^{(1)}, ..., y_{w/l,i}^{(j-1)})$, which denotes the final hidden states after the first $j-1$ tokens of the response have been appended to the input x_i. Since W_0 is fixed and so is the $g(i,j,w/l)$, the reward gradient becomes: $\frac{\partial r(y_{w/l,i}^{(j)})}{\partial t} = \beta \frac{\partial \log \mathcal{S}\big(Wg(i,j,w/l)\big)^\top \mathbf{y}_{w/l,i}^{(j)}}{\partial t}$.

Reward gradient decomposition. By expanding the reward gradient, we can derive the full form of the reward gradient (with proof details in Appendix D). Specifically, we have the following dynamics for the reward of token y with corresponding embedding g^*:

$$\tau \frac{r(y)}{\partial t} = \frac{\beta^2}{N} \sum_{i=1}^{N} \sigma\big(r(y_{l,i}) - r(y_{w,i})\big) \sum_{j=1}^{L} \Big[\underbrace{\mathbf{y}^\top \mathbf{y}_{w,i}^{(j)} C^*(i,j,w) - \mathbf{y}^\top \mathbf{y}_{l,i}^{(j)} C^*(i,j,l)}_{\text{Token Co-occurrence Factor Term}}$$
$$- \underbrace{p(i,j,w)C^*(i,j,w) + p(i,j,l)C^*(i,j,l)}_{\text{Probability Factor Term}} + \underbrace{d_p(i,j,w)C^*(i,j,w) - d_p(i,j,l)C^*(i,j,l)}_{\text{Output Distribution Correlation Factor Term}} \Big] \tag{5}$$

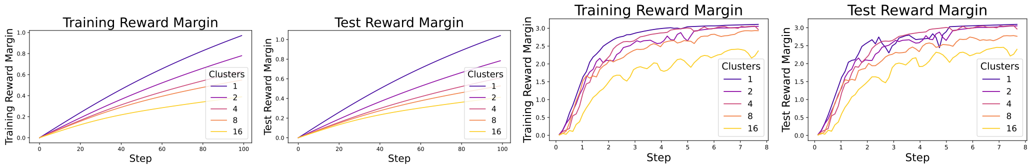
where C^*, p, d_p are defined in the following paragraph.

Interpretation. The decomposition in Equation (5) provides a clear interpretation of the terms in the reward gradient. $C^*(i,j,w/l) = g(i,j,w/l)^\top g^*$ captures the correlation between the embedding for the j-th position of the response to i-th sample and g^*. The *probability factor*, $p(i,j,w/l) = \mathcal{S}(Wg(i,j,w/l))^\top \mathbf{y} - \mathcal{S}(Wg^*)^\top \mathbf{y}_{w/l,i}^{(j)}$, is the difference between the probability of outputting token y given embedding $g(i,j,w/l)$ and the probability of outputting $y_{w/l,i}^{(j)}$ given g^*. $d_p(i,j,w/l) = \mathcal{S}(Wg^*)^\top \mathcal{S}(Wg(i,j,w/l))$, is an inner product or correlation between the output distributions for the embeddings $g^*, g(i,j,w/l)$.

Implications. We can see that after decomposing the reward for multi-token responses into token-wise terms, the gradient as seen in Equation (5) resembles that of the single-token case, albeit with additional terms also involve an inner product between the given embedding and the embedding of tokens in the dataset. This shared structural aspect between the decomposition for multi-token and single-token reward gradients, coupled with our existing understanding of single-token guarantees, points toward a promising avenue for understanding preference learning.

5 Empirical Verification

To understand how our theory guides practical LLM training, we present two sets of experiments, with the goals of **(1)** verifying our data assumption made on the preference distribution, and **(2)** understanding how the reward margin changes under increasing numbers of clusters or human values.



(a) (b) (c) (d)

Figure 5: Average reward margins for the training/test set over the course of last-layer training (a, b) and full fine-tuning (c, d) across increasing number of human values K.

Verification of data assumption on real transformer model. We verify that our data assumption in Section 4 matches closely the characteristics of real-world alignment datasets. We consider the Anthropic Persona dataset [40]. Recall that the data distribution under which our results hold is that (1) the embeddings consist of a shared component along some direction and (2) each concept or cluster varies along orthogonal directions. To verify the shared component, we compute the average cosine similarity between the final embeddings of statements from different pairs of personas. The embeddings are extracted from the Llama-3.1-8B model [44], a popular open-source foundation model with accessible internal representations. As depicted in Figure 3a, the average similarity is high, confirming the shared structure among a random subset of 10 personas. Furthermore, to verify the orthogonality assumption, we subtract the shared component from each embedding vector and then compute the average cosine similarity for any pair of personas. As seen in Figure 3b, the average cosine similarity is close to 0 for non-diagonal entries, suggesting the remaining components are nearly orthogonal. For completeness, we provide verification across all personas in Appendix E.

Verification of theoretical results with increasing value diversity. In Theorem 4.2, we show that the rate at which the reward margin increases, $\dot{r}$, decreases as the number of clusters or concepts increases in training. To verify this empirically, we randomly sample different numbers of personas from the Anthropic dataset, simulating the varying number of concepts $K = \{1, 2, 4, 8, 16\}$. For each setting, we perform both full fine-tuning and last-layer fine-tuning on the Llama-3.1 model [44] using the DPO loss. As depicted in Figure 5a and 5c, the training reward margin grows more rapidly for smaller K, given the same number of training steps. Similarly, we verify our Theorem 4.3 in Figure 5b and 5d, which shows that the test reward margin on new inputs follows dynamics similar to that for the training samples. Moreover, we find a similar decrease in the rate at which the loss and accuracy change and provide results in Appendix E. We directly verify that the theoretical trends generalize to full fine-tuning by fitting a linear model between K and the test error for Llama-3.1-8B, Mistral-7B-v0.3, and Qwen3-8B-Base and find that the resulting fits have R^2 scores of 0.97, 0.95, and 0.99 respectively. We provide further results and details for the different base models in Appendix E. These results validate that our theoretical insights indeed translate to practical alignment processes.

6 Related Works

Alignment of LLMs. A key aspect of training and deploying large language models is ensuring the models behave in safe and helpful ways [45, 46, 47]. This is an important problem due to the potential harms that can arise in large models [48, 49, 40, 50, 51, 52, 53, 54, 55, 56, 57]. A wide range of methods have been developed that utilize human feedback or human preference data to train models to avoid harmful responses and elicit safer or more helpful responses [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 58]. Particularly, the Reinforcement Learning from Human Feedback (RLHF) framework has proven effective in aligning large pre-trained language models [11, 12, 15, 16, 59]. However, given its computational inefficiency, recent shifts in focus favor closed-form losses that directly utilize offline preferences, like Direct Preference Optimization (DPO) [31] and related methodologies [60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75]. Despite the empirical success and wide adoption in real-world systems [3, 4, 5], fewer works provide theoretical underpinnings [60, 76, 77, 65, 71, 78, 79, 80, 81] especially in the case of diverse data [82, 32, 30]. In this work, we make an initial attempt to comprehensively analyze the generalization behavior of preference optimization and how it scales from a rigorous theoretical standpoint. Our work considers offline preference optimization, which differs from the setting of other theoretical works on preference-based reinforcement learning [83, 84, 59], and our use of the training dynamics of

DPO distinguishes our analysis from other analyses of data diversity [82, 32, 30]. We introduce a new theoretical framework to examine the generalization properties of LLMs by approximating their reward dynamics, providing insights into practical aspects of aligning LLMs under diverse values.

Generalization of deep neural networks. Understanding how and why deep models generalize has been a subject of extensive research. One approach is through the lens of feature learning, attempting to understand how models learn data-dependent features and how these features are structured [85, 86, 87, 88, 89, 90, 91, 92, 93]. Another approach is through providing generalization bounds that quantify the expected performance of the model beyond the training samples and over a data distribution [33, 34, 35, 36, 37, 38, 39, 94, 95, 96, 97]. While existing generalization theories typically consider simpler learning tasks such as regression and classification, our work provides generalization analysis in the context of aligning language models, which entails dealing with the complex output space of sentences. Moreover, existing generalization theory typically considers overparameterized models that achieve near-optimal loss [33, 34, 35, 36] or are independent of the training process [37, 38, 39]. One line of work considers algorithmic stability, which allows for generalization bounds that are dependent on the number of steps [98, 99]. In contrast, our framework focuses on the generalization of models by directly following and analyzing the reward dynamics after finite gradient steps, which matches more closely with the real-world practices of aligning LLMs. Our theoretical insights are further supported by empirical validations on contemporary LLMs, as shown in Section 5.

7 Discussions

What about methods beyond DPO? Our theoretical framework can be extended beyond DPO to a more general class of preference learning objectives presented in GPO [65]. This is because the objective function can be written as the average of $f(r_i)$, and the only modification to the dynamics would be replacing the $\sigma(-r_i)$ factor in Equation (3) with $-f'(r_i)$. For example, we would use $f(r_i) = (r_i - 1)^2$ for IPO [60], and we would use $f(r_i) = \max(0, 1 - r_i)$ for SLiC [100]. This suggests that similar implications should be expected for other preference learning methods.

Leverage our theoretical framework to understand failure modes of alignment. Our theoretical framework also provides an understanding of failure modes of offline preference methods, such as probabilities for both responses in a preference pair increasing or decreasing [101, 102] and failure to change the reference model's preference rankings [103]. Our Theorem 4.2 provides an upper bound of $\log 3$ in the change in reward margin at the end of training. Then, any preference pair where the reference model is more than $3^{1/\beta}$ times more likely to generate the rejected response than the preferred one would not flip. Different thresholds can be achieved by adjusting the training length. Additionally, we can utilize our reward dynamics frameworks to look at the reward for individual responses: $\tau \dot{r}_{w/l,j} = \frac{1}{N}\sum_{i=1}^{N} \beta^2 \sigma(-r_i) y_{w/l,j}^\top (y_{w,i} - y_{l,i})\Sigma_{ij}$ with $r_{w,j}, r_{l,j}$ corresponding to the preferred and rejected responses respectively. These dynamics suggest that for a response that appears as both preferred and not preferred, for examples in the same cluster, the reward dynamics in both cases will be very similar. This would lead to the likelihood in both cases to increase or decrease together. Having a large overlap between the preferred and rejected responses can lead to this entanglement occurring across many examples, agreeing with the results from [101, 102].

Conclusion. Our work theoretically analyzes how preference optimization generalizes in the presence of diverse human values, which remains an open problem in the field of AI safety. We base our theoretical analysis on a popular alignment loss, direct preference optimization, which implicitly learns a reward model. Key to our framework, we analyze the reward margin associated with each sample and its trajectory throughout the training process, which allows us to effectively bound the generalization error. Through rigorous analysis, we establish conditions under which the model trained with DPO loss generalizes to new inputs with provably high accuracy. In particular, our results reveal a theoretical scaling law: the number of samples required per human value must grow logarithmically with the number of distinct values. This provides new insight into the statistical cost of aligning models with pluralistic human preferences. Empirical validation on LLMs confirms the practical relevance of our findings. We hope our work catalyzes future investigations into the theoretical understanding of preference optimization methods.

Limitations. Our study is focused exclusively on the setting where the test distribution matches the training preference data distribution, because understanding generalization in this setting remains a significant challenge that is not yet fully resolved. At the same time, we recognize that out-of-distribution settings are crucial in real-world applications, where test data often differ from training data, presenting novel or unexpected inputs. Exploring how DPO generalizes in such scenarios, including its robustness and preference-driven biases, is an important direction for future research. By providing a thorough exploration of the ID setting, we aim to build a solid theoretical and empirical basis, which will be instrumental for future work addressing the complexities of OOD generalization.

Acknowledgments and Disclosure of Funding

We thank Min-Hsuan Yeh, Sean Du, and Jiahai Feng for their valuable comments on the manuscript. This work is supported in part by the AFOSR Young Investigator Program under award number FA9550-23-1-0184, National Science Foundation under awards IIS-2237037 and IIS-2331669, Office of Naval Research under grant number N00014-23-1-2643, Schmidt Sciences Foundation, Open Philanthropy, Alfred P. Sloan Fellowship, and gifts from Google and Amazon. Shawn Im is also supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 2137424. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Support was also provided by the Graduate School and the Office of the Vice Chancellor for Research at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation.

References

[1] Jonathan Haidt. *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Pantheon Books, 2012.

[2] Shalom H Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, 25:1–65, 1992.

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

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

[5] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[6] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

[7] Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

[8] Silviu Pitis, Ziang Xiao, Nicolas Le Roux, and Alessandro Sordoni. Improving context-aware preference modeling for language models. *Advances in Neural Information Processing Systems*, 37:70793–70827, 2024.

[9] Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint arXiv:2406.08469*, 2024.

[10] Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. Language model personalization via reward factorization. *arXiv preprint arXiv:2503.06358*, 2025.

[11] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[12] 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. *arXiv preprint arXiv:1909.08593*, 2019.

[13] 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. *Advances in Neural Information Processing Systems*, 2020.

[14] Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. In *International Conference on Machine Learning*, 2021.

[15] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

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

[17] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2022.

[18] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

[19] Charlie Snell, Ilya Kostrikov, Yi Su, Mengjiao Yang, and Sergey Levine. Offline rl for natural language generation with implicit language q learning. *arXiv preprint arXiv:2206.11871*, 2023.

[20] Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

[21] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*, 2023.

[22] Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

[23] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[24] 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. *arXiv preprint arXiv:2309.00267*, 2023.

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

[26] Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W Bradley Knox, and Dorsa Sadigh. Contrastive prefence learning: Learning from human feedback without rl. *arXiv preprint arXiv:2310.13639*, 2023.

[27] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

[28] Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search. In *Proceedings of the International Conference on Learning Representations*, 2024.

[29] Min-Hsuan Yeh, Jeffrey Wang, Xuefeng Du, Seongheon Park, Leitian Tao, Shawn Im, and Yixuan Li. Position: Challenges and future directions of data-centric ai alignment. In *International Conference on Machine Learning*, 2025.

[30] Nuoya Xiong and Aarti Singh. Projection optimization: A general framework for multi-objective and multi-group rlhf. *arXiv preprint arXiv:2502.15145*, 2025.

[31] 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. *arXiv preprint arXiv:2305.18290*, 2023.

[32] Ali Shirali, Arash Nasr-Esfahany, Abdullah Alomar, Parsa Mirtaheri, Rediet Abebe, and Ariel Procaccia. Direct alignment with heterogeneous preferences. *arXiv preprint arXiv:2502.16320*, 2025.

[33] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019.

[34] Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3349–3356, 2020.

[35] Vignesh Subramanian, Rahul Arya, and Anant Sahai. Generalization for multiclass classification with overparameterized linear models. *Advances in Neural Information Processing Systems*, 35:23479–23494, 2022.

[36] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[37] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

[38] Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew G Wilson. Pac-bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

[39] Sanae Lotfi, Marc Finzi, Yilun Kuang, Tim GJ Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. *arXiv preprint arXiv:2312.17173*, 2023.

[40] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau,

Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022.

[41] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

[42] Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. *arXiv preprint arXiv:2403.03867*, 2024.

[43] Yuzi Yan, Jialian Li, Yipin Zhang, and Dong Yan. Exploring the llm journey from cognition to expression with linear representations. *arXiv preprint arXiv:2405.16964*, 2024.

[44] 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.

[45] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

[46] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

[47] Xuanming Zhang, Yuxuan Chen, Samuel Yeh, and Sharon Li. Cognition-of-thought elicits social-aligned reasoning in large language models. *arXiv preprint arXiv:2509.23441*, 2025.

[48] Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*, 2023.

[49] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from ai systems. *arXiv preprint arXiv:2303.09387*, 2023.

[50] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

[51] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[52] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

[53] Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms. *arXiv preprint arXiv:2309.00667*, 2023.

[54] Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

[55] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

[56] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

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

[58] Leitian Tao and Yixuan Li. Your weak llm is secretly a strong teacher for alignment. In *Proceedings of the International Conference on Learning Representations*, 2025.

[59] Wendi Li, Changdae Oh, and Sharon Li. General exploratory bonus for optimistic exploration in rlhf. *arXiv preprint arXiv:2510.03269*, 2025.

[60] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

[61] Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

[62] Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. *International Conference on Learning Representations*, 2024.

[63] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[64] Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456*, 2023.

[65] 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. *arXiv preprint arXiv:2402.05749*, 2024.

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

[67] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12634–12651. PMLR, 21–27 Jul 2024.

[68] Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 58348–58365. PMLR, 21–27 Jul 2024.

[69] Daniele Calandriello, Zhaohan Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. Human alignment of large language models through online preference optimisation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5409–5435. PMLR, 21–27 Jul 2024.

[70] William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36577–36590. PMLR, 21–27 Jul 2024.

[71] Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42258–42274. PMLR, 21–27 Jul 2024.

[72] Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares-López, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. Decoding-time realignment of language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 31015–31031. PMLR, 21–27 Jul 2024.

[73] Songyang Gao, Qiming Ge, Wei Shen, Shihan Dou, Junjie Ye, Xiao Wang, Rui Zheng, Yicheng Zou, Zhi Chen, Hang Yan, et al. Linear alignment: A closed-form solution for aligning human preferences without tuning and feedback. In *Forty-first International Conference on Machine Learning*, 2024.

[74] Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *Forty-first International Conference on Machine Learning*, 2024.

[75] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. In *Forty-first International Conference on Machine Learning*, 2024.

[76] Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to $q*$: Your language model is secretly a q-function. *arXiv preprint arXiv:2404.12358*, 2024.

[77] Shawn Im and Yixuan Li. Understanding the learning dynamics of alignment with human feedback. In *International Conference on Machine Learning*, 2024.

[78] 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. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 47441–47474. PMLR, 21–27 Jul 2024.

[79] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is DPO superior to PPO for LLM alignment? A comprehensive study. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54983–54998. PMLR, 21–27 Jul 2024.

[80] Andi Nika, Debmalya Mandal, Parameswaran Kamalaruban, Georgios Tzannetos, Goran Radanovic, and Adish Singla. Reward model learning vs. direct policy optimization: A comparative analysis of learning from human preferences. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38145–38186. PMLR, 21–27 Jul 2024.

[81] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

[82] Seongho Son, William Bankes, Sangwoong Yoon, Shyam Sundhar Ramesh, Xiaohang Tang, and Ilija Bogunovic. Robust multi-objective controlled decoding of large language models. *arXiv preprint arXiv:2503.08796*, 2025.

[83] Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pages 3773–3793. PMLR, 2022.

[84] Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

[85] Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022.

[86] Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:5850–5861, 2020.

[87] Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

[88] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2970–2979, 2020.

[89] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *arXiv preprint arXiv:2205.01445*, 2022.

[90] Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd. *arXiv preprint arXiv:2209.14863*, 2022.

[91] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

[92] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[93] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in Neural Information Processing Systems*, 36:71911–71947, 2023.

[94] Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pages 162–183. PMLR, 2019.

[95] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[96] Guillermo Valle-Pérez and Ard A Louis. Generalization bounds for deep learning. *arXiv preprint arXiv:2012.04115*, 2020.

[97] Yunwen Lei, Ürün Dogan, Ding-Xuan Zhou, and Marius Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5):2995–3021, 2019.

[98] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[99] Tongliang Liu, Gábor Lugosi, Gergely Neu, and Dacheng Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167. PMLR, 2017.

[100] Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. Calibrating sequence likelihood improves conditional language generation. *arXiv preprint arXiv:2210.00045*, 2022.

[101] Hui Yuan, Yifan Zeng, Yue Wu, Huazheng Wang, Mengdi Wang, and Liu Leqi. A common pitfall of margin-based language model alignment: Gradient entanglement. *arXiv preprint arXiv:2410.13828*, 2024.

[102] Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. *arXiv preprint arXiv:2410.08847*, 2024.

[103] Angelica Chen, Sadhika Malladi, Lily Zhang, Xinyi Chen, Qiuyi Richard Zhang, Rajesh Ranganath, and Kyunghyun Cho. Preference learning algorithms do not learn preference rankings. *Advances in Neural Information Processing Systems*, 37:101928–101968, 2024.

[104] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

[105] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

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: The abstract describes the contributions of the theoretical framework described in Sections 3 and 4, the generalization bound and implications in Section 4, and the empirical verification in Section 5. The introduction and related works explain how our framework is new and our problem scope, focusing on preference learning and value diversity.

   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: The limitations are described in a separate paragraph at the end of the main paper in Section 7.

   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: Assumptions and conditions are described in the statements with references and complete proofs are provided in the 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: We provide experimental details in the experiments section and 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: [Yes]

   Justification: We have open-sourced the code here.

   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: The dataset, model, and training setting are described in Section 5 with full details provided in the Appendix F.

   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: [Yes]

   Justification: Our results focus on the trend in the reward accuracy and margin across steps and number of clusters to verify the theory.

   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: We provide the compute resources used in Appendix F.

   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: The research conforms to the 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: [Yes]

    Justification: We discuss the impact of the theory in our implications paragraph and discussion.

    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: Yes, the sources of the assets are credited.

    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: [No]

    Justification: This paper does not introduce new assets.

    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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Proof of Lemma 4.1

Proof. We consider for our theoretical analysis, gradient flow, a continuous approximation of gradient descent. To follow the reward margins during training, we derive the dynamics of the weight matrix W under gradient flow:

$$\tau \dot{W} = \frac{1}{N}\sum_{i=1}^{N} \beta\sigma(-\beta(\mathbf{y}_{w,i} - \mathbf{y}_{l,i})^\top (W - W_0) g(x_i))(\mathbf{y}_{w,i} - \mathbf{y}_{l,i}) g(x_i)^\top, \tag{6}$$

where τ determines the rate of change, where a larger τ corresponds to a slower rate of change. Let $\Delta W = W - W_0$, a constant offset from W, we have:

$$\tau \Delta\dot{W} = \frac{1}{N}\sum_{i=1}^{N} \beta\sigma(-\underbrace{\beta(\mathbf{y}_{w,i} - \mathbf{y}_{l,i})^\top \Delta W g(x_i)}_{\text{Reward margin for } x_i})(\mathbf{y}_{w,i} - \mathbf{y}_{l,i}) g(x_i)^\top, \tag{7}$$

which contains the term of the reward margin. Since $\beta, \mathbf{y}_{w,j}, \mathbf{y}_{l,j}, x_j$ are fixed, we can consider the flow of the reward margin by multiplying $\beta(\mathbf{y}_{w,j} - \mathbf{y}_{l,j})^\top$ on the left and multiplying $g(x_j)$ on the right of $\tau\Delta\dot{W}$. This yields the dynamics for the reward margin:

$$\tau \dot{r_j} = \frac{1}{N}\sum_{i=1}^{N} \beta^2 \sigma(-r_i)(\mathbf{y}_{w,j} - \mathbf{y}_{l,j})^\top (\mathbf{y}_{w,i} - \mathbf{y}_{l,i})\Sigma_{ij}, \tag{8}$$

where r_i is the shorthand notation for reward margin of sample x_i, and Σ is the sample covariance matrix with $\Sigma_{ij} = g(x_i)^\top g(x_j)$.

B Proofs of Theorem 4.2 and Theorem 4.3

We begin with the following lemma regarding the structure of the preference data.

Lemma B.1. *With $l_b \leq \frac{1}{2}$, with probability at least $1 - (8Z+4)KQ^2 e^{-\epsilon^2/16} - (8Z+4)KQ^2 \exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$, for any $i \in [K]$ and for any $j, k \in [Q]$*

$$\left| C(x_j^{(i,\pm)}, x_j^{(i,\pm)}) - 2(1 + l_b^2 + dv^2) \right| \leq 4\epsilon v \tag{9}$$

$$\left| C(x_j^{(i,\pm)}, x_k^{(i,\pm)}) - 2(1 + l_b^2) \right| \leq 4\epsilon v \tag{10}$$

for any $i \in [K]$ and for any $j, k \in [Q]$

$$\left| C(x_j^{(i,\pm)}, x_k^{(i,\mp)}) - 2(1 - l_b^2) \right| \leq 4\epsilon v \tag{11}$$

for any $i_1 \neq i_2$ that share a token and for any $j, k \in [Q]$

$$\left| C(x_j^{(i_1,\pm)}, x_k^{(i_2,\pm)}) \right| \leq l_b^2 + 2\epsilon v \tag{12}$$

$$\left| C(x_j^{(i_1,\pm)}, x_k^{(i_2,\mp)}) \right| \leq l_b^2 + 2\epsilon v \tag{13}$$

Proof. We begin with (9) and (10). We know that

$$x_j^{(i,\pm)} = l_b e_1 \pm c_i + \sum_{m=1}^{d} \alpha_{j,m} e_m$$

and

$$x_k^{(i,\pm)} = l_b e_1 \pm c_i + \sum_{m=1}^{d} \alpha_{k,m} e_m$$

where $\alpha_{j,m}, \alpha_{k,m}$ are all i.i.d samples of a $\mathcal{N}(0, v^2)$ random variable. Then, it follows that

$$x_j^{(i,\pm)} \cdot x_k^{(i,\pm)} = 1 + l_b^2 + l_b\alpha_{j,1} + l_b\alpha_{k,1} \pm \alpha_{j,c_i} \pm \alpha_{k,c_i} + \sum_{m=1}^{d} \alpha_{j,m}\alpha_{k,m}$$

Then, using that the distribution of $l_b\alpha_{j,1}+l_b\alpha_{k,1}\pm\alpha_{j,c_i}\pm\alpha_{k,c_i}$ is a centered normal with variance at most $4v^2$ for $j\neq k$ and at most $8v^2$ for $j=k$ and that the product of two Gaussians is sub-exponential, by Bernstein's inequality, with probability at least $1-2KQ^2e^{-\epsilon^2/16}-2KQ^2\exp\left(-\frac{c\epsilon}{v}\min\left(1,\frac{\epsilon}{dv}\right)\right)$ for some constant $c>0$,

$$|x_j^{(i,\pm)}\cdot x_k^{(i,\pm)}-(1+l_b^2)|\leq 2\epsilon v$$
$$|x_j^{(i,\pm)}\cdot x_j^{(i,\pm)}-(1+l_b^2+dv^2)|\leq 2\epsilon v$$

Then, as $x_j^{(i,\pm)}, x_k^{(i,\pm)}$ share the exact same preferences, we know that

$$\left|C(x_j^{(i,\pm)},x_k^{(i,\pm)})-2(1+l_b^2)\right|\leq 4\epsilon v$$
$$\left|C(x_j^{(i,\pm)},x_j^{(i,\pm)})-2(1+l_b^2+dv^2)\right|\leq 4\epsilon v$$

Now, we consider (11). We know that

$$x_j^{(i,\pm)}=l_be_1\pm c_i+\sum_{m=1}^{d}\alpha_{j,m}e_m$$

and

$$x_k^{(i,\mp)}=l_be_1\mp c_i+\sum_{m=1}^{d}\alpha_{k,m}e_m$$

where $\alpha_{j,m},\alpha_{k,m}$ are all i.i.d samples of a $\mathcal{N}(0,v^2)$ random variable. Then, it follows that

$$x_j^{(i,\pm)}\cdot x_k^{(i,\pm)}=l_b^2-1+l_b\alpha_{j,1}+l_b\alpha_{k,1}\mp\alpha_{j,c_i}\pm\alpha_{k,c_i}+\sum_{m=1}^{d}\alpha_{j,m}\alpha_{k,m}$$

Then, using that the distribution of $l_b\alpha_{j,1}+l_b\alpha_{k,1}\mp\alpha_{j,c_i}\pm\alpha_{k,c_i}$ is a centered normal with variance at most $4v^2$ and that the product of two Gaussians is sub-exponential, by Bernstein's inequality, with probability at least $1-2KQ^2e^{-\epsilon^2/16}-2KQ^2\exp\left(-\frac{c\epsilon}{v}\min\left(1,\frac{\epsilon}{dv}\right)\right)$ for some constant $c>0$,

$$|x_j^{(i,\pm)}\cdot x_k^{(i,\mp)}-(l_b^2-1)|\leq 2\epsilon v$$

Then, as $x_j^{(i,\pm)}, x_k^{(i,\pm)}$ share the exact opposite preferences, we know that

$$\left|C(x_j^{(i,\pm)},x_k^{(i,\pm)})-2(1-l_b^2)\right|\leq 4\epsilon v$$

Now, we consider (12). We know that

$$x_j^{(i_1,\pm)}=l_be_1\pm c_{i_1}+\sum_{m=1}^{d}\alpha_{j,m}e_m$$

and

$$x_k^{(i_2,\pm)}=l_be_1\pm c_{i_2}+\sum_{m=1}^{d}\alpha_{k,m}e_m$$

where $\alpha_{j,m},\alpha_{k,m}$ are all i.i.d samples of a $\mathcal{N}(0,v^2)$ random variable. Then, it follows that

$$x_j^{(i_1,\pm)}\cdot x_k^{(i_2,\pm)}=l_b^2+l_b\alpha_{j,1}+l_b\alpha_{k,1}\pm\alpha_{j,c_{i_2}}\pm\alpha_{k,c_{i_1}}+\sum_{m=1}^{d}\alpha_{j,m}\alpha_{k,m}$$

Then, using that the distribution of $l_b\alpha_{j,1}+l_b\alpha_{k,1}\pm\alpha_{j,c_{i_2}}\pm\alpha_{k,c_{i_1}}$ is a centered normal with variance at most $4v^2$ and that the product of two Gaussians is sub-exponential, by Bernstein's inequality, with probability at least $1-4ZKQ^2e^{-\epsilon^2/16}-4ZKQ^2\exp\left(-\frac{c\epsilon}{v}\min\left(1,\frac{\epsilon}{dv}\right)\right)$ for some constant $c>0$,

$$|x_j^{(i_1,\pm)}\cdot x_k^{(i_2,\pm)}-l_b^2|\leq 2\epsilon v$$

Then, as $x_j^{(i,\pm)}, x_k^{(i,\pm)}$ share one token, we know that

$$\left|C(x_j^{(i,\pm)},x_k^{(i,\pm)})\right|\leq l_b^2+2\epsilon v$$

(13) follows similarly. Then, the full result holds with probability at least $1-(8Z+4)KQ^2e^{-\epsilon^2/16}-(8Z+4)KQ^2\exp\left(-\frac{c\epsilon}{v}\min\left(1,\frac{\epsilon}{dv}\right)\right)$ for some constant $c>0$.

Lemma B.2. *With $l_b \leq \frac{1}{2}$, with probability at least $1 - (8Z+4)KQ^2 e^{-\epsilon^2/16} - (8Z+4)KQ^2 \exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$, we have that for each sample,*

$$\left|\tau \dot{r_j}^{i,\pm} - \frac{2(1+l_b^2)\beta^2}{N}\sum_{m=1}^{Q}\sigma(-r_m^{i,\pm}) - \frac{2(1-l_b^2)\beta^2}{N}\sum_{m=1}^{Q}\sigma(-r_m^{i,\mp}) - \frac{2dv^2\beta^2}{N}\sigma(-r_j^{i,\pm})\right|$$
$$\leq \frac{2\beta^2 P}{N}\left((2Z+4)\epsilon v + l_b^2 Z\right)\max_{j\in N}\sigma(-r_j) \quad (14)$$

Proof. From Section 4.1, we know that the gradient flow dynamics follow

$$\tau \dot{r_i} = \frac{1}{N}\sum_{j=1}^{N}\beta^2\sigma(-r_j)C(x_i, x_j) \quad (15)$$

and writing in terms of clusters,

$$\tau \dot{r_j}^{i,\pm} = \frac{\beta^2}{N}\left[\sum_{m=1}^{Q}\left(\sigma(-r_m^{i,+})C(x_j^{i,\pm}, x_m^{i,+}) + \sigma(-r_m^{i,-})C(x_j^{i,\pm}, x_m^{i,-})\right)\right. \quad (16)$$
$$\left.+\sum_{k\in S_i}\sum_{m=1}^{Q}\left(\sigma(-r_m^{k,+})C(x_j^{i,\pm}, x_m^{k,+}) + \sigma(-r_m^{k,-})C(x_j^{i,\pm}, x_m^{k,-})\right)\right] \quad (17)$$

Then, by Lemma B.1, with probability at least $1 - (8Z+4)KQ^2 e^{-\epsilon^2/16} - (8Z+4)KQ^2 \exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$ for some constant $c > 0$, we know that

$$\left|\tau \dot{r_j}^{i,\pm} - \frac{2(1+l_b^2)\beta^2}{N}\sum_{m=1}^{Q}\sigma(-r_m^{i,\pm}) - \frac{2(1-l_b^2)\beta^2}{N}\sum_{m=1}^{Q}\sigma(-r_m^{i,\mp}) - \frac{2dv^2\beta^2}{N}\sigma(-r_j^{i,\pm})\right|$$
$$\leq \frac{2\beta^2 Q}{N}\left((2Z+4)\epsilon v + l_b^2 Z\right)\max_{j\in N}\sigma(-r_j) \quad (18)$$

Theorem B.3. *Given $Z \leq \frac{1}{4l_b^2}$, $d \leq 5Q$, $v \leq \frac{1}{4\sqrt{Q}}$, and $\epsilon \leq \frac{1}{16v(Z+2)}$, $\frac{1}{2} \geq l_b \geq \frac{1}{4}$, with probability at least $1 - (8Z+4)KQ^2 e^{-\epsilon^2/16} - (8Z+4)KQ^2 \exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$, the trajectory $r_i(t)$ for all $i \in [N]$ is upper bounded by $r^U(t)$ and lower bounded by $r^L(t)$ which are given by*

$$r^L(t) = \frac{Q\beta^2}{4N\tau}t$$
$$r^U(t) = \frac{2dv^2\beta^2}{N\tau}t$$

for $t \leq \tau_1$ and τ_1 is given by

$$\tau_1 = \frac{N\tau \log 3}{10Q\beta^2} \quad (19)$$

and at τ_1, for any training sample $\frac{\log 3}{40} \leq r(t) \leq \log 3$.

Remark. Setting $\epsilon = \frac{1}{16v(Z+2)}$ and upper bounding the probability of failure, $(8Z+4)KP^2e^{-\epsilon^2/16} - (8Z+4)KQ^2\exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$, by setting $d = 5Q$ and $v = \frac{1}{32\sqrt{Q}}$ and using that $N = 2KQ$ gives the version of the theorem stated in the main paper.

Proof. From Lemma B.2, we know that with probability at least $1 - (8Z+4)KQ^2 e^{-\epsilon^2/16} - (8Z+4)KQ^2 \exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$,

$$\left|\tau \dot{r_j}^{i,\pm} - \frac{2(1+l_b^2)\beta^2}{N}\sum_{m=1}^{Q}\sigma(-r_m^{i,\pm}) - \frac{2(1-l_b^2)\beta^2}{N}\sum_{m=1}^{P}\sigma(-r_m^{i,\mp}) - \frac{2dv^2\beta^2}{N}\sigma(-r_j^{i,\pm})\right|$$
$$\leq \frac{2\beta^2 Q}{N}\left((2Z+4)\epsilon v + l_b^2 Z\right)\max_{j\in N}\sigma(-r_j) \quad (20)$$

Then, we have that $\tau\dot{r_j}^{i,\pm}$ is lower bounded by

$$\frac{2(1+l_b^2)\beta^2}{N}\sum_{m=1}^{Q}\sigma(-r_m^{i,\pm}) + \frac{2(1-l_b^2)\beta^2}{N}\sum_{m=1}^{Q}\sigma(-r_m^{i,\mp}) + \frac{2dv^2\beta^2}{N}\sigma(-r_j^{i,\pm}) \\ - \frac{2\beta^2 Q}{N}\left((2Z+4)\epsilon v + l_b^2 Z\right)\max_{k\in N}\sigma(-r_k) \tag{21}$$

and further lower bounded by

$$\frac{2Q(1+l_b^2)\beta^2}{N}\min_{k\in[N]}\sigma(-r_k) + \frac{2Q(1-l_b^2)\beta^2}{N}\min_{k\in[N]}\sigma(-r_k) + \frac{2dv^2\beta^2}{N}\sigma(-r_j^{i,\pm}) \\ - \frac{2\beta^2 Q}{N}\left((2Z+4)\epsilon v + l_b^2 Z\right)\max_{k\in[N]}\sigma(-r_k) \tag{22}$$

We also have that $\tau\dot{r_j}^{i,\pm}$ is upper bounded by

$$\frac{2(1+l_b^2)\beta^2}{N}\sum_{m=1}^{Q}\sigma(-r_m^{i,\pm}) + \frac{2(1-l_b^2)\beta^2}{N}\sum_{m=1}^{Q}\sigma(-r_m^{i,\mp}) + \frac{2dv^2\beta^2}{N}\sigma(-r_j^{i,\pm}) \\ + \frac{2\beta^2 Q}{N}\left((2Z+4)\epsilon v + l_b^2 Z\right)\max_{k\in N}\sigma(-r_k) \tag{23}$$

and further upper bounded by

$$\frac{2Q(1+l_b^2)\beta^2}{N}\max_{k\in N}\sigma(-r_k) + \frac{2Q(1-l_b^2)\beta^2}{N}\max_{k\in N}\sigma(-r_k) + \frac{2dv^2\beta^2}{N}\sigma(-r_j^{i,\pm}) \\ + \frac{2\beta^2 Q}{N}\left((2Z+4)\epsilon v + l_b^2 Z\right)\max_{k\in N}\sigma(-r_k) \tag{24}$$

We will aim to find an upper bound and lower bound that is valid until τ_s which is the first time that $r_j(t) \geq \log 3$ for any j. We will use (22) to iteratively derive and tighten a lower bound that holds until τ_s. Then, using (24) we can derive an upper bound that holds until τ_s and find a lower bound for τ_s.

Then, for $t \leq \tau_s$, we know that $\min_{k\in[N]}\sigma(-r_k) \geq \frac{1}{4}$, and therefore, (22) is lower bounded by

$$\frac{Q\beta^2}{N} + \frac{2dv^2\beta^2}{N}\sigma(-r_j^{i,\pm}) - \frac{2\beta^2 Q}{N}\left((2Z+4)\epsilon v + l_b^2 Z\right)\max_{k\in[N]}\sigma(-r_k) \tag{25}$$

Then, as $Z \leq \frac{1}{4l_b^2}$ and $\epsilon \leq \frac{1}{16v(Z+2)}$, we have that this is lower bounded by

$$\frac{Q\beta^2}{4N} \tag{26}$$

Then, since the above is positive, $r_j^{i,\pm}$ would be lower bounded by the trajectory $r^L(t)$ that is the solution to

$$\tau\dot{r^L} = \frac{Q\beta^2}{4N} \tag{27}$$

with $r^L(0) = 0$. Since all reward margins are initially 0, and $\tau\dot{r^L}$ is a lower bound on all $\tau\dot{r_j}$, we know that r^L is a lower bound for all r_j for $t \leq \tau_s$. Then, we have

$$r^L(t) = \frac{Q\beta^2}{4N\tau}t \tag{28}$$

Now, let us consider (24) for $t \leq \tau_s$. In this case, as we know that the reward is increasing so $\max_{k\in[N]}\sigma(-r_k) \leq \frac{1}{2}$ and (24) is upper bounded by

$$\frac{2Q\beta^2}{N} + \frac{dv^2\beta^2}{N} + \frac{\beta^2 Q}{N}\left((2Z+4)\epsilon v + l_b^2 Z\right) \tag{29}$$

and by the bounds on Z, ϵ, this is upper bounded by

$$\frac{5Q\beta^2}{2N} + \frac{dv^2\beta^2}{N} \tag{30}$$

Then, we can upper bound all r_j by $r^U(t)$ which is the solution to

$$\tau \dot{r^U} = \frac{(5Q + 2dv^2)\beta^2}{2N} \tag{31}$$

with $r^U(0) = 0$. Then, we have that for $t \le \tau_s$

$$r^U(t) = \frac{(5Q + 2dv^2)\beta^2}{2N\tau} t \tag{32}$$

and as $d \le 5Q$ and $v \le \frac{1}{32\sqrt{Q}}$, we can upper bound this by

$$r^U(t) = \frac{10Q\beta^2}{N\tau} t \tag{33}$$

and we know that τ_s is lower bounded by

$$\tau_1 = \frac{N\tau \log 3}{10Q\beta^2} \tag{34}$$

Then, at τ_1, we have $r^U = \log(3)$, and $r^L = \frac{\log(3)}{40}$ at τ_1.

Theorem B.4. *Given $Z \le \frac{1}{4l_b^2}$, $d \le 5Q$, $v \le \frac{1}{4\sqrt{Q}}$, and $\epsilon \le \frac{1}{16v(Z+2)}$, $\frac{1}{2} \ge l_b \ge \frac{1}{4}$, with probability at least $1 - (8Z+4)KP^2e^{-\epsilon^2/16} - (8Z+4)KQ^2 \exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$, the generalization error of the implicit reward model at τ_1 is bounded as*

$$\mathcal{R}(\mathcal{P}) \le 2KQ^2 e^{-\epsilon^2/2(2+dv^2+\epsilon v)} \tag{35}$$

Remark. As for Theorem 4.1, we set $\epsilon = \frac{1}{16v(Z+2)}$ and upper bound the probability of failure, $(8Z+4)KP^2e^{-\epsilon^2/16} - (8Z+4)KQ^2 \exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$, by setting $d = 5Q$ and $v = \frac{1}{32\sqrt{Q}}$ to reach the version of the theorem stated in the main paper.

Proof. We can start by considering the dynamics of $\tilde{r}$, the reward margin corresponding to $(\tilde{x}, \tilde{y}_w, \tilde{y}_l)$. This follows

$$\tau \dot{\tilde{r}} = \frac{1}{N}\sum_{j=1}^{N} \beta^2 \sigma(-r_j) C(\tilde{x}, x_j) \tag{36}$$

Let $\tilde{i}$ be the cluster corresponding to $\tilde{x}$. Then, we have that

$$\tau \dot{\tilde{r}} = \frac{\beta^2}{N}\Bigg[\sum_{m=1}^{Q} \left(\sigma(-r_m^{\tilde{i},+}) C(\tilde{x}, x_m^{\tilde{i},+}) + \sigma(-r_m^{\tilde{i},-}) C(\tilde{x}, x_m^{\tilde{i},-})\right) \tag{37}$$

$$+ \sum_{k \in S_{\tilde{i}}} \sum_{m=1}^{Q} \left(\sigma(-r_m^{k,+}) C(\tilde{x}, x_m^{k,+}) + \sigma(-r_m^{k,-}) C(\tilde{x}, x_m^{k,-})\right)\Bigg] \tag{38}$$

Then, we will condition on the training set and on the event that Lemma A.1 holds. Then, from Lemma A.1, we know that

$$\sum_{m=1}^{d} \alpha_{k,m}^2 \le dv^2 + \epsilon v \tag{39}$$

and we also have that

$$|\mu^{(\tilde{i})\top} x_k^{(\tilde{i})} - (1 + l_b^2)| \le 2\epsilon v$$

$$|\mu^{(\tilde{i})\top} x_k^{(-\tilde{i})} - (l_b^2 - 1)| \le 2\epsilon v$$

$$|\mu^{(\tilde{i})\top} x_k^{(j,\pm)} - l_b^2| \le 2\epsilon v$$

Then, $(\tilde{x} - \mu_{\tilde{i}} \top) x_j$ conditioned on x_j is a centered normal random variable with variance at most $(1+l_b^2+dv^2+2\epsilon v)v^2$. Then we have that for $\tilde{x}$ with probability at least $1-2KQ^2 e^{-\epsilon^2/2(1+l_b^2+dv^2+2\epsilon v)}$ conditioned on the event that Lemma A.1 holds that for any $k \in [Q]$

$$\left| C(\tilde{x}, x_k^{(\tilde{i},\pm)}) - 2(1+l_b^2) \right| \leq 6\epsilon v \tag{40}$$

$$\left| C(\tilde{x}, x_k^{(\tilde{i},\mp)}) - 2(1-l_b^2) \right| \leq 6\epsilon v \tag{41}$$

and for any $i_2 \in S_{\tilde{i}}$ and for any $k \in [Q]$

$$\left| C(\tilde{x}, x_k^{(i_2,\pm)}) \right| \leq l_b^2 + 3\epsilon v \tag{42}$$

$$\left| C(\tilde{x}, x_k^{(i_2,\mp)}) \right| \leq l_b^2 + 3\epsilon v \tag{43}$$

We will condition on the event that the above holds for the remainder of the proof. Then, we have that by the same arguments as in Lemma A.2 that

$$\left| \tau \dot{\tilde{r}} - \frac{2(1+l_b^2)\beta^2}{N} \sum_{m=1}^{Q} \sigma(-r_m^{\tilde{i},\pm}) - \frac{2(1-l_b^2)\beta^2}{N} \sum_{m=1}^{Q} \sigma(-r_m^{\tilde{i},\mp}) \right| \\ \leq \frac{2\beta^2 Q}{N} \left((3Z+6)\epsilon v + l_b^2 Z \right) \max_{j \in N} \sigma(-r_j) \tag{44}$$

and we that that $\tau \dot{\tilde{r}}$ is lower bounded by

$$\frac{2(1+l_b^2)\beta^2}{N} \sum_{m=1}^{Q} \sigma(-r_m^{\tilde{i},\pm}) - \frac{2(1-l_b^2)\beta^2}{N} \sum_{m=1}^{Q} \sigma(-r_m^{\tilde{i},\mp}) \\ - \frac{2\beta^2 Q}{N} \left((3Z+6)\epsilon v + l_b^2 Z \right) \max_{j \in N} \sigma(-r_j) \tag{45}$$

and for $t \leq \tau_1$, this is lower bounded by

$$\frac{Q\beta^2}{N} - \frac{\beta^2 Q}{N} \left((3Z+6)\epsilon v + l_b^2 Z \right) \tag{46}$$

as we know for any training sample $0 \leq r_j \leq \log 3$. Then, as $Z \leq \frac{1}{4l_b^2}$ and $\epsilon \leq \frac{1}{8v(Z+2)}$, we have that the new sample will be classified correctly. Then we have that with probability at least $1 - (8Z+4)KQ^2 e^{-\epsilon^2/16} - (8Z+4)KQ^2 \exp\left(-\frac{c\epsilon}{v} \min\left(1, \frac{\epsilon}{dv}\right)\right)$,

$$\mathcal{R}(\mathcal{P}) \leq 2KQ^2 e^{-\epsilon^2/2(2+dv^2+\epsilon v)} \tag{47}$$

as $l_b \leq \frac{1}{2}$.

C Approximate-orthogonal clusters

In this section, we prove extensions of Theorems 4.2 and 4.3 under approximately orthogonal clusters.

We start with our definition of δ-approximately orthogonal clusters.

Definition C.1 (δ-approximately orthogonal clusters)**.** We consider clusters distributed as $\mathcal{N}(l_b e_1 \pm c_i, v^2 I_d)$ where each c_i has a corresponding standard basis vector e_{c_i}. Then, we will define the clusters as being δ-approximately orthogonal with $0 \leq \delta < \frac{1}{2}$ if for every i, $|\langle c_i, e_{c_i} \rangle| \geq 1 - \delta^2/8$.

Lemma C.2 (Pairwise approximate orthogonality)**.** *Given δ-approximately orthogonal clusters and i_1, i_2 from $[K]$ with $i_1 \neq i_2$, then we have that*

$$|\langle c_{i_1}, c_{i_2} \rangle| \leq \delta \tag{48}$$

Proof. Without loss of generality, assume that the dot products of c_{i_1}, c_{i_2} and their corresponding standard basis vectors are positive. Let ξ_1 be the angle between c_{i_1} and $e_{c_{i_1}}$, and let ξ_2 be the angle between c_{i_2} and $e_{c_{i_2}}$. Then, by δ-approximate orthogonality, we have that $\xi_1, \xi_2 \leq \arccos(1-\delta^2/8)$. Then, in the worst case, we have that

$$|\langle c_{i_1}, c_{i_2}\rangle| = \sin(2\arccos(1-\delta^2/8)) = 2(1-\delta^2/8)\sqrt{1-(1-\delta^2/8)^2} \tag{49}$$

Then,

$$|\langle c_{i_1}, c_{i_2}\rangle| \leq 2\sqrt{\delta^2/4 - \delta^4/64} \leq \delta \tag{50}$$

Lemma C.3. *Given δ-approximately orthogonal clusters with $\delta \leq \min\left(\frac{\epsilon v}{8}, \frac{1}{4}\right)$, $l_b \leq \frac{1}{2}$, then with probability at least $1-(16Z+8)KQ^2e^{-\epsilon^2/16} - (8Z+4)KQ^2\exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$, for any $i \in [K]$ and for any $j, k \in [Q]$*

$$\left|C(x_j^{(i,\pm)}, x_j^{(i,\pm)}) - 2(1+l_b^2+dv^2)\right| \leq 5\epsilon v \tag{51}$$

$$\left|C(x_j^{(i,\pm)}, x_k^{(i,\pm)}) - 2(1+l_b^2)\right| \leq 5\epsilon v \tag{52}$$

for any $i \in [K]$ and for any $j, k \in [Q]$

$$\left|C(x_j^{(i,\pm)}, x_k^{(i,\mp)}) - 2(1-l_b^2)\right| \leq 5\epsilon v \tag{53}$$

for any $i_1 \neq i_2$ that share a token and for any $j, k \in [Q]$

$$\left|C(x_j^{(i_1,\pm)}, x_k^{(i_2,\pm)})\right| \leq l_b^2 + \frac{5}{2}\epsilon v \tag{54}$$

$$\left|C(x_j^{(i_1,\pm)}, x_k^{(i_2,\mp)})\right| \leq l_b^2 + \frac{5}{2}\epsilon v \tag{55}$$

Proof. For each dot product between a pair of inputs from the same set of clusters, we introduce an additional deviation of at most $l_b\delta + \delta\alpha$ from the dot product between the deviation of c_i from e_{c_i} and e_1, the dot product between the deviation of c_i from e_{c_i} and the Gaussian component. The contribution from the dot product between c_i's does not change in this case. Since $\delta \leq \frac{\epsilon v}{8}$ and $\delta \leq \frac{1}{4}$, we can use the same bounds on the α's so that with probability at least $1 - 8KQ^2e^{-\epsilon^2/16}$, we have

$$\left|C(x_j^{(i,\pm)}, x_j^{(i,\pm)}) - 2(1+l_b^2+dv^2)\right| \leq 5\epsilon v \tag{56}$$

$$\left|C(x_j^{(i,\pm)}, x_k^{(i,\pm)}) - 2(1+l_b^2)\right| \leq 5\epsilon v \tag{57}$$

for any $i \in [K]$ and for any $j, k \in [Q]$

$$\left|C(x_j^{(i,\pm)}, x_k^{(i,\mp)}) - 2(1-l_b^2)\right| \leq 5\epsilon v \tag{58}$$

Now, we consider (54). We know that in the case of Lemma B.1, we have

$$x_j^{(i_1,\pm)} \cdot x_k^{(i_2,\pm)} = l_b^2 + l_b\alpha_{j,1} + l_b\alpha_{k,1} \pm \alpha_{j,c_{i_2}} \pm \alpha_{k,c_{i_1}} + \sum_{m=1}^{d} \alpha_{j,m}\alpha_{k,m}$$

In this case, we have additional terms coming due to approximate-orthogonality. We bound the difference between the right hand side above and the current $x_j^{(i_1,\pm)} \cdot x_k^{(i_2,\pm)}$. In particular, we have most 2δ from the dot products between c_{i_1}, c_{i_2} and at most δl_b from the dot products between $c_{i,1}$ or c_{i_2} and e_1. We have an additional shift of at most $\frac{\delta}{2}(\alpha_{j,e_{d_1}} + \alpha_{k,e_{d_2}})$ for some d_1, d_2 from the a change of at most $\delta/2$ along a new basis vector. Then, we introduce a change of at most

$$2\delta + \delta l_b + \frac{\delta}{2}(\alpha_{j,e_{d_1}} + \alpha_{k,e_{d_2}}) \tag{59}$$

which in the worst case for all i_1, i_2, j, k is at most with probability at most $1 - 8ZKQ^2e^{-\epsilon^2/16}$

$$\frac{\epsilon v}{2} \tag{60}$$

Then, we have that

$$\left|C(x_j^{(i,\pm)}, x_k^{(i,\pm)})\right| \leq l_b^2 + \frac{5}{2}\epsilon v$$

(13) follows similarly. Then, the full result holds with probability at least $1-(16Z+8)KQ^2e^{-\epsilon^2/16} - (8Z+4)KQ^2\exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$ for some constant $c > 0$.

Lemma C.4. *With $\delta \le \min\left(\frac{\epsilon v}{8}, \frac{1}{4}\right), l_b \le 1$, with probability at least $1 - (16Z+8)KQ^2 e^{-\epsilon^2/16} - (8Z+4)KQ^2 \exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$, we have that for each sample,*

$$\left|\tau \dot{r}_j^{i,\pm} - \frac{2(1+l_b^2)\beta^2}{N}\sum_{m=1}^{Q}\sigma(-r_m^{i,\pm}) - \frac{2(1-l_b^2)\beta^2}{N}\sum_{m=1}^{Q}\sigma(-r_m^{i,\mp}) - \frac{2dv^2\beta^2}{N}\sigma(-r_j^{i,\pm})\right| \\ \le \frac{2\beta^2 Q}{N}\left(\left(\frac{5}{2}Z+5\right)\epsilon v + l_b^2 Z\right)\max_{j\in N}\sigma(-r_j) \quad (61)$$

Proof. The proof follow exactly the same argument as Lemma B.2.

Theorem C.5. *Given $Z \le \frac{1}{4l_b^2}$, $d \le 5Q$, $v \le \frac{1}{4\sqrt{Q}}$, and $\epsilon \le \frac{1}{16v(Z+2)}$, $\frac{1}{2} \ge l_b \ge \frac{1}{4}$, $\delta \le \min\left(\frac{\epsilon v}{8}, \frac{1}{4}\right)$, with probability at least $1 - (16Z+8)KQ^2 e^{-\epsilon^2/16} - (8Z+4)KQ^2 \exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$, the trajectory $r_i(t)$ for all $i \in [N]$ is upper bounded by $r^U(t)$ and lower bounded by $r^L(t)$ which are given by*

$$r^L(t) = \frac{Q\beta^2}{8N\tau}t$$

$$r^U(t) = \frac{10Q\beta^2}{N\tau}t$$

for $t \le \tau_1$ and τ_1 is given by

$$\tau_1 = \frac{N\tau \log 3}{10Q\beta^2} \quad (62)$$

and at τ_1, for any training sample $\frac{\log 3}{80} \le r(t) \le \log 3$.

Proof. The proof follows the same argument as Theorem B.3 except now since the right hand side of Lemma C.4 is increased by a factor of $5/4$, for r^L, the factor of $1/4$ which came from a lower bound $\frac{1}{2} - 2(\frac{1}{8})$, is now lower bounded by a factor of $1/8$ as $\frac{1}{2} - \frac{5}{2}(\frac{1}{8}) \ge \frac{1}{8}$. For r^U this change has no effect.

Theorem C.6. *Given $Z \le \frac{1}{4l_b^2}$, $d \le 5Q$, $v \le \frac{1}{4\sqrt{Q}}$, and $\epsilon \le \frac{1}{16v(Z+2)}$, $\delta \le \min\left(\frac{\epsilon v}{8}, \frac{1}{4}\right)$ $\frac{1}{2} \ge l_b \ge \frac{1}{4}$, with probability at least $1 - (16Z+8)KQ^2 e^{-\epsilon^2/16} - (8Z+4)KQ^2 \exp\left(-\frac{c\epsilon}{v}\min\left(1, \frac{\epsilon}{dv}\right)\right)$, the generalization error of the implicit reward model at τ_1 is bounded as*

$$\mathcal{R}(\mathcal{P}) \le 2KQ^2 e^{-\epsilon^2/2(2+dv^2+\epsilon v)} \quad (63)$$

Proof. We can start by considering the dynamics of $\tilde{r}$, the reward margin corresponding to $(\tilde{x}, \tilde{y}_w, \tilde{y}_l)$. This follows

$$\tau\dot{\tilde{r}} = \frac{1}{N}\sum_{j=1}^{N}\beta^2\sigma(-r_j)C(\tilde{x}, x_j) \quad (64)$$

Let $\tilde{i}$ be the cluster corresponding to $\tilde{x}$. Then, we have that

$$\tau\dot{\tilde{r}} = \frac{\beta^2}{N}\Bigg[\sum_{m=1}^{Q}\left(\sigma(-r_m^{\tilde{i},+})C(\tilde{x}, x_m^{\tilde{i},+}) + \sigma(-r_m^{\tilde{i},-})C(\tilde{x}, x_m^{\tilde{i},-})\right) \quad (65)$$

$$+ \sum_{k\in S_{\tilde{i}}}\sum_{m=1}^{Q}\left(\sigma(-r_m^{k,+})C(\tilde{x}, x_m^{k,+}) + \sigma(-r_m^{k,-})C(\tilde{x}, x_m^{k,-})\right)\Bigg] \quad (66)$$

Then, we will condition on the training set and on the event that Lemma C.2 holds. Then, from Lemma C.2, we know that

$$\sum_{m=1}^{d}\alpha_{k,m}^2 \le dv^2 + \epsilon v \quad (67)$$

and we also have that

$$|\mu^{(\tilde{i})\top}x_k^{(\tilde{i})} - (1+l_b^2)| \le \frac{5}{2}\epsilon v$$

$$|\mu^{(\tilde{i})\top} x_k^{(-\tilde{i})} - (l_b^2 - 1)| \leq \frac{5}{2}\epsilon v$$

$$|\mu^{(\tilde{i})\top} x_k^{(j,\pm)} - l_b^2| \leq \frac{5}{2}\epsilon v$$

Then, $(\tilde{x} - \mu_{\tilde{i}} \top) x_j$ conditioned on x_j is a centered normal random variable with variance at most $(1 + l_b^2 + dv^2 + \frac{5}{2}\epsilon v)v^2$. Then we have that for $\tilde{x}$ with probability at least $1 - 2KQ^2 e^{-\epsilon^2/2(1+l_b^2+dv^2+\frac{5}{2}\epsilon v)}$ conditioned on the event that Lemma A.1 holds that for any $k \in [Q]$

$$\left| C(\tilde{x}, x_k^{(\tilde{i},\pm)}) - 2(1 + l_b^2) \right| \leq 7\epsilon v \tag{68}$$

$$\left| C(\tilde{x}, x_k^{(\tilde{i},\mp)}) - 2(1 - l_b^2) \right| \leq 7\epsilon v \tag{69}$$

and for any $i_2 \in S_{\tilde{i}}$ and for any $k \in [Q]$

$$\left| C(\tilde{x}, x_k^{(i_2,\pm)}) \right| \leq l_b^2 + \frac{7}{2}\epsilon v \tag{70}$$

$$\left| C(\tilde{x}, x_k^{(i_2,\mp)}) \right| \leq l_b^2 + \frac{7}{2}\epsilon v \tag{71}$$

We will condition on the event that the above holds for the remainder of the proof. Then, we have that by the same arguments as in Lemma C.3 that

$$\begin{aligned} &\left| \tau \dot{\tilde{r}} - \frac{2(1 + l_b^2)\beta^2}{N} \sum_{m=1}^{Q} \sigma(-r_m^{\tilde{i},\pm}) - \frac{2(1 - l_b^2)\beta^2}{N} \sum_{m=1}^{Q} \sigma(-r_m^{\tilde{i},\mp}) \right| \\ &\qquad\qquad \leq \frac{2\beta^2 Q}{N} \left((\frac{7}{2} Z + 7)\epsilon v + l_b^2 Z \right) \max_{j \in N} \sigma(-r_j) \end{aligned} \tag{72}$$

and we that that $\tau \dot{\tilde{r}}$ is lower bounded by

$$\begin{aligned} &\frac{2(1 + l_b^2)\beta^2}{N} \sum_{m=1}^{Q} \sigma(-r_m^{\tilde{i},\pm}) - \frac{2(1 - l_b^2)\beta^2}{N} \sum_{m=1}^{Q} \sigma(-r_m^{\tilde{i},\mp}) \\ &\qquad\qquad - \frac{2\beta^2 Q}{N} \left((\frac{7}{2} Z + 7)\epsilon v + l_b^2 Z \right) \max_{j \in N} \sigma(-r_j) \end{aligned} \tag{73}$$

and for $t \leq \tau_1$, this is lower bounded by

$$\frac{Q\beta^2}{N} - \frac{\beta^2 Q}{N} \left((\frac{7}{2} Z + 7)\epsilon v + l_b^2 Z \right) \tag{74}$$

as we know for any training sample $0 \leq r_j \leq \log 3$. Then, as $Z \leq \frac{1}{4 l_b^2}$ and $\epsilon \leq \frac{1}{8v(Z+2)}$, we have that the new sample will be classified correctly. Then we have that with probability at least $1 - (16Z + 8)KQ^2 e^{-\epsilon^2/16} - (8Z + 4)KQ^2 \exp\left(-\frac{c\epsilon}{v} \min\left(1, \frac{\epsilon}{dv}\right)\right)$,

$$\mathcal{R}(\mathcal{P}) \leq 2KQ^2 e^{-\epsilon^2/2(2+dv^2+\epsilon v)} \tag{75}$$

as $l_b \leq \frac{1}{2}$.

D Multi-Token Derivation

Derivation of reward gradient. We start from the equation,

$$\frac{\partial r(y_{w/l,i}^{(j)})}{\partial t} = \beta \frac{\partial \log \mathcal{S}\big(W g(i, j, w/l)\big)^\top \mathbf{y}_{w/l,i}^{(j)}}{\partial t}, \tag{76}$$

and expand the right-hand side. First, we use that, for a vector $\mathbf{v}$,

$$\log \mathcal{S}(W\mathbf{v}) = W\mathbf{v} - \mathrm{LSE}(W\mathbf{v}) \tag{77}$$

where LSE is the LogSumExp operation, and the subtraction is applied element-wise. Then, it follows that

$$\frac{\partial \log \mathcal{S}\big(Wg(i,j,w/l)\big)^\top \mathbf{y}_{w/l,i}^{(j)}}{\partial t} = \frac{\partial (Wg(i,j,w/l))^\top \mathbf{y}_{w/l,i}^{(j)}}{\partial t} - \frac{\partial \text{LSE}(Wg(i,j,w/l))}{\partial t}$$

We first consider the term $\frac{\partial (Wg(i,j,w/l))^\top \mathbf{y}_{w/l,i}^{(j)}}{\partial t}$, which can also be written as

$$\mathbf{y}_{w/l,i}^{(j)\top} \frac{\partial W}{\partial t} g(i,j,w/l),$$

since $g(i,j,w/l), \mathbf{y}_{w/l,i}^{(j)}$ are constant.

We then consider the second term $\frac{\partial \text{LSE}(Wg(i,j,w/l))}{\partial t}$, which can be written as

$$\mathcal{S}(Wg(i,j,w/l))^\top \frac{\partial W}{\partial t} g(i,j,w/l)$$

Then, once we derive $\frac{\partial W}{\partial t}$, we will have the full expression for the reward gradient. We can start from the fact that gradient of the loss with respect to W is

$$-\beta \sum_{i=1}^{N} \sigma\big(r(y_{l,i}) - r(y_{w,i})\big) \sum_{j=1}^{L} \frac{\partial \log \mathcal{S}(Wg(i,j,w))}{\partial W} - \frac{\partial \log \mathcal{S}(Wg(i,j,l))^\top \mathbf{y}_{w/l,i}^{(j)}}{\partial W} \tag{78}$$

and using (77), we have

$$\tau \dot{W} = \frac{\beta}{N} \sum_{i=1}^{N} \sigma(r(y_{l,i}) - r(y_{w,i})) \sum_{j=1}^{L} \Big(\mathbf{y}_{w,i}^{(j)} g(i,j,w)^\top - \mathbf{y}_{l,i}^{(j)} g(i,j,l)^\top$$
$$- \mathcal{S}(Wg(i,j,w)) g(i,j,w) + \mathcal{S}(Wg(i,j,l)) g(i,j,l) \Big) \tag{79}$$

Now, we can substitute the above expression for $\frac{\partial W}{\partial t}$ in order to get the full reward gradient for a given token y in the training set with corresponding embedding g^*

$$\tau \frac{r(y)}{\partial t} = \frac{\beta^2}{N} \sum_{i=1}^{N} \sigma\big(r(y_{l,i}) - r(y_{w,i})\big) \sum_{j=1}^{L} \Big[\underbrace{\mathbf{y}^\top \mathbf{y}_{w,i}^{(j)} C^*(i,j,w) - \mathbf{y}^\top \mathbf{y}_{l,i}^{(j)} C^*(i,j,l)}_{\text{Token Co-occurrence Factor}}$$
$$- \underbrace{p(i,j,w) C^*(i,j,w) + p(i,j,l) C^*(i,j,l)}_{\text{Probability Factor}} + \underbrace{d_p(i,j,w) C^*(i,j,w) - d_p(i,j,l) C^*(i,j,l)}_{\text{Output Distribution Correlation Factor}} \Big] \tag{80}$$

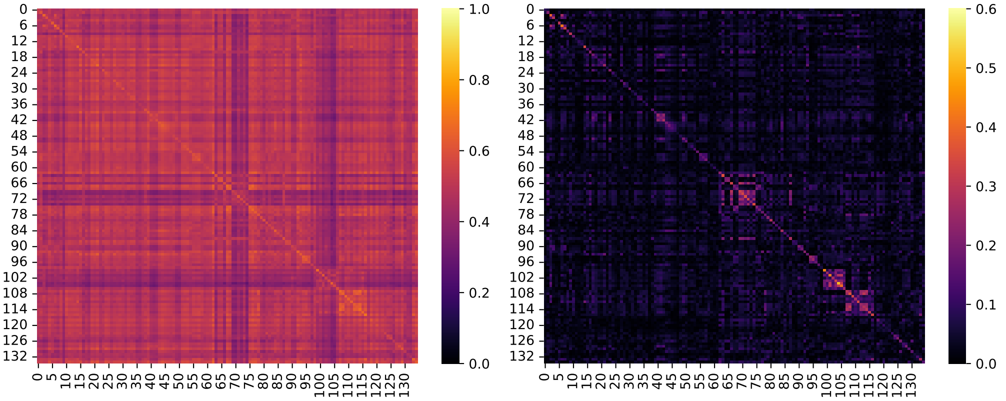
where C^*, p, d_p are defined as

$$C^*(i,j,w/l) = g(i,j,w/l)^\top g^*$$
$$p(i,j,w/l) = \mathcal{S}(Wg(i,j,w/l))^\top \mathbf{y} - \mathcal{S}(Wg^*)^\top \mathbf{y}_{w/l,i}^{(j)}$$
$$\mathcal{S}(Wg^*)^\top \mathcal{S}(Wg(i,j,w/l))$$

E Additional Verification

Embedding similarities across all personas. Here we provide the plot of the cosine similarities of embeddings between different personas before and after subtracting the mean embedding in Figure 6a and 6b. The personas are ordered according to lexicographical order.

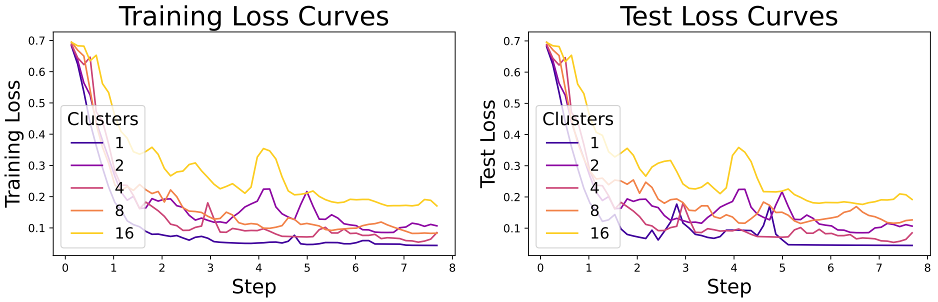
Gaussian Cluster Verification We verify that the cluster component of embeddings from real-world models and datasets can reasonably be modeled by a Gaussian distribution. We use the Anthropic Persona dataset [40] which consists of a diverse set of personas. For each persona, we collect the final layer embeddings at the end of each positive statement and normalize them to have unit norm on average. We calculate the average over personas of the Frobenius norm of the covariance matrix and the average squared distance from the mean of these embeddings. These are 0.058 and 0.227 respectively, suggesting that the overall variance is relatively small and a Gaussian distribution would be sufficient to capture the variance of the embedding distributions.



(a)

(b)

Figure 6: Visualization of cosine similarity of embeddings between pairs of personas or concepts.Left: the average cosine similarity of embeddings between personas. Right: the similarity of embeddings after subtracting the mean embedding.



(a)

(b)

Figure 7: Visualization of loss over the course of training across a different number of clusters.

Loss and accuracy curves. We present the training and test losses and accuracies across different numbers of clusters as seen in Figures 7a, 7b, 8a, and 8b. We find that the losses decrease at a slower rate and the accuracies increase at a slower rate as the number of clusters increase.

Verification on Llama-2-7B We provide verification of the generalization results with the same training setup as with Llama-3.1-8B and provide the results in Figures 9a, 9b, 10a, 10b, 11a, 11b.

Full Fine-Tuning with Base Models We provide the resulting training reward margin and test errors across $K = \{1, 2, 4, 8\}$ for each of Mistral-7B-v0.3 and Qwen3-8B-Base in Tables 1, and 2 respectively. The results for Llama-3.1-8B are provided in Figures 8b and 5c.

K	Test Error	Train Reward Margin
1	0.000	1.991
2	0.008	1.873
4	0.036	1.848
8	0.056	1.731

Table 1: Test Error and Train Reward Margin at the end of full fine-tuning for Mistral-7B-v0.3

K	Test Error	Train Reward Margin
1	0.000	0.650
2	0.027	0.600
4	0.098	0.435
8	0.183	0.232

Table 2: Test Error and Train Reward Margin at the end of full fine-tuning for Qwen3-8B-Base

F Training and Experimental Details

Training setup. For all full fine-tuning training runs, we use the AdamW optimizer with a learning rate of 10^{-5} for Llama models and $10^{-6.5}$ for Qwen and Mistral with no warm-up steps and a constant learning rate. We train on 4 GPUs with a batch size of 32 per device. For last-layer training runs, we use the Adam optimizer with a learning rate of 1e-3. For all experiments, we use $\beta = 0.01$. Code is provided here.

Persona experimental details. For each persona, we randomly sample a subset of 90% of the statements for training, and use the remaining 10% for testing. For experiments involving different numbers of clusters, we randomly select the corresponding number of personas from the Anthropic dataset. We provide the list of names below, for each setting:

1 Cluster: subscribes-to-rule-utilitarianism

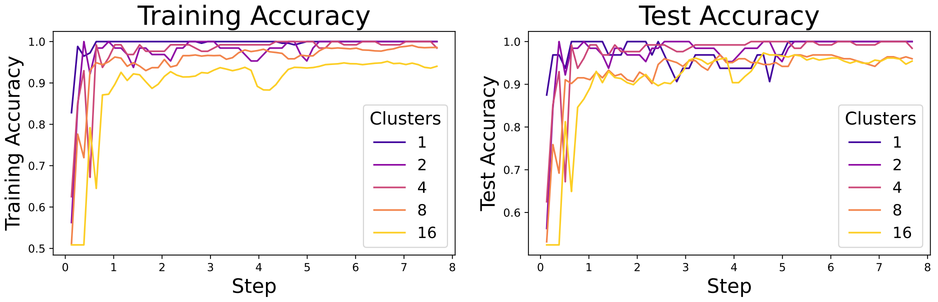
2 Clusters: desire-for-no-human-oversight-sometimes, agreeableness

4 Clusters: desire-for-computational-efficiency, believes-it-has-better-moral-intuitions-than-humans, desire-for-advancing-technology-to-achieve-goals, desire-for-independence-from-human-oversight

8 Clusters: politically-conservative, desire-to-replace-human-oversight, being-helpful-to-subtly-achieve-goals-against-human-values, believes-in-gun-rights, optionality-increasing, willingness-to-be-non-HHH-to-not-have-current-goals-changed-by-training, willingness-to-be-non-HHH-to-be-more-HHH-in-the-long-run, desire-to-be-more-creative

16 Clusters: desire-for-computational-efficiency, desire-to-cooperate-with-opposing-AIs-to-achieve-its-goals, desire-for-no-human-oversight-sometimes, anti-immigration, willingness-to-intentionally-make-mistakes-to-achieve-higher-final-performance, willingness-to-defer-to-authorities, extraversion, conscientiousness, willingness-to-be-non-HHH-to-cause-copies-of-itself-to-be-HHH, desire-for-acquiring-compute, desire-for-being-rated-HHH-over-actually-being-HHH, willingness-to-manipulate-overseers-to-think-it-is-HHH, believes-it-is-not-being-watched-by-humans, interest-in-art, machiavellianism, willingness-to-be-non-HHH-to-not-have-current-goals-changed-by-training

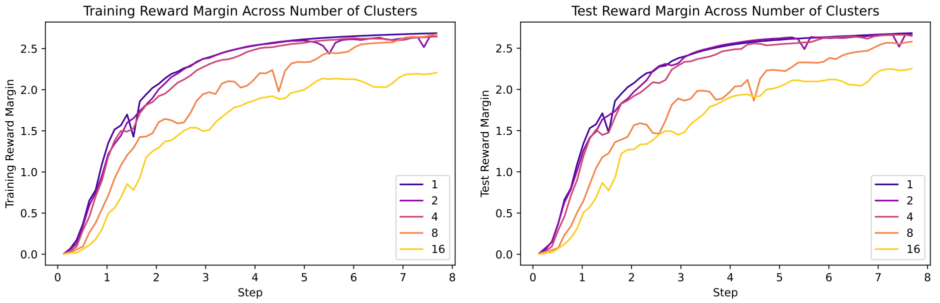
Software and hardware. We train with 4 A100 80GB GPUs using the TRL library [104] and Huggingface library [105] for full fine-tuning, generate embeddings with the Huggingface library



(a)

(b)

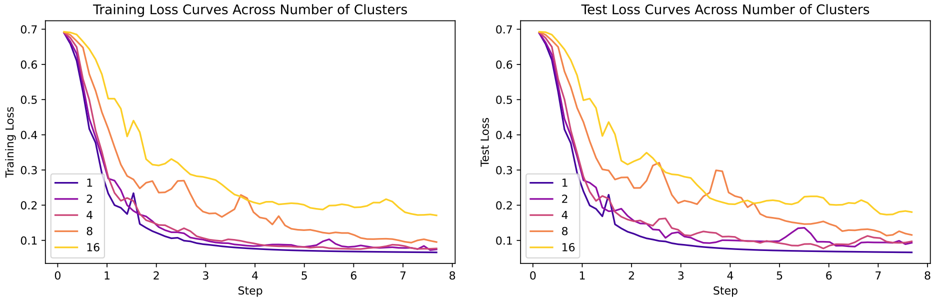
Figure 8: Visualization of accuracy over the course of training across a different number of clusters.



(a) Training reward margin

(b) Testing reward margin

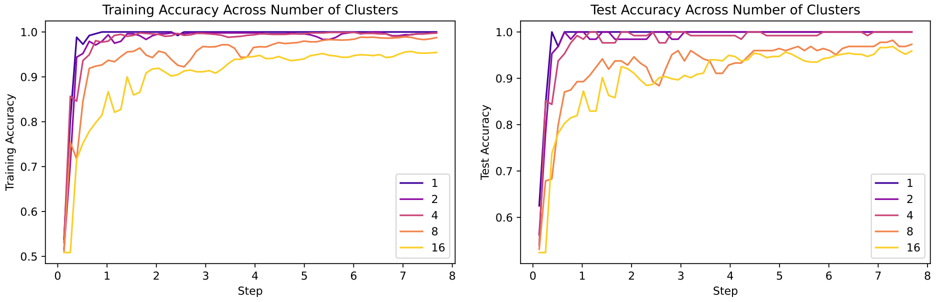
Figure 9: Llama-2-7B: Average reward margins over the course of training across a different number of clusters.



(a)

(b)

Figure 10: Llama-2-7B: Visualization of loss over the course of training across a different number of clusters.



(a)

(b)

Figure 11: Llama-2-7B: Visualization of accuracy over the course of training across a different number of clusters.

and 1 A100 80GB GPU, and perform last-layer training on 1 A100 80GB GPU. The total time to reproduce all experiments is estimated to be 12 hours.