The Representation-Rationalizability Tradeoff in Reward Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

In RLHF, each training example contains a prompt x and two candidate responses y, y', and annotators provide pairwise preferences between these responses. The learning problem is to convert these heterogeneous pairwise judgments into a single scalar reward $r(x, y)$ that measures response quality for each prompt. Classical social choice implies an impossibility because heterogeneous annotator samples can induce pooled preferences with Condorcet cycles, so no scalar reward can evaluate all compared response pairs consistently. A growing literature analyzes RLHF as a social-choice problem, but usually assumes a fixed finite set of alternatives, i.e., a pre-enumerated finite set of candidate responses for each prompt. Modern pipelines instead score responses through a learned representation $\phi(x, y)$ before a scalar head, so ϕ determines which responses are treated as distinguishable alternatives and which comparisons are visible to the reward model. Once this embedding is part of the problem, the impossibility results from social choice theory become a tradeoff. We show that the excess cross-entropy loss of any reward built on ϕ decomposes exactly into a representational term, which a richer ϕ shrinks, and an aggregation term, which a richer ϕ enlarges by exposing more comparisons that no scalar can rank consistently. The same results extend to direct preference optimization (DPO), and jointly training the embedding and the reward cannot guarantee to recover the sweet spot of this tradeoff. Experiments on synthetic data and real preference datasets corroborate our results.

1 Introduction

Should reward models use richer representations to better fit human preferences? In modern RLHF [20], the aggregation task is to fit a single scalar reward to pooled pairwise preferences from many annotators. A popular architecture uses a pretrained language-model encoder that maps a prompt-response pair (x, y) to features $\phi(x, y)$, followed by a thin score head g, so $r(x, y) = g(\phi(x, y))$. A common intuition is that increasing representational richness can only help because it gives the reward model more signals to fit. We show that this intuition is incomplete. Richer representations can improve fit in one sense while hurting it in another.

The reason is that representation changes the aggregation problem itself. The embedding determines which response pairs are distinguishable to the reward fitter, so refining ϕ turns previously collapsed responses into distinct alternatives and adds new pairwise comparison that a single scalar reward must rationalize. Classical social choice theory already warns that no scalar reward can consistently aggregate pooled judgments from heterogeneous annotators when their combined comparisons contain cycles [1, 11, 25], and a recent body of work has translated these social choice results into the RLHF setting [5, 7, 16, 30]. Most of these works, however, treat the response space as a fixed finite set of alternatives. Once the embedding is allowed to vary, the social choice impossibility

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

result becomes tunable because changing ϕ changes which annotator disagreements are visible in the pooled comparisons.

Our main result makes this trade-off precise. Under a random-utility model with arbitrary annotator heterogeneity, the excess cross-entropy loss of any embedding-based reward decomposes into two parts: an embedding loss measuring information that ϕ discards by collapsing distinguishable preferences, and an agreement cost that measures how far the pooled preferences, as seen through ϕ, are from being rationalizable by any scalar reward. Under a mild assumption on the representation, the lower bounds of the two terms move in opposite directions as ϕ refines. As a consequence, both the smallest and the largest representations can be suboptimal, and the optimal embedding dimension is dataset-dependent. We further show that this lower bound is tight under a constructive reward.

Our framework extends beyond reward modeling. We show that Direct Preference Optimization (DPO) [22] admits a similar decomposition and excess-loss lower bound, demonstrating that the same tradeoff governs preference-based policy learning even without an explicit reward model. Based on the excess loss lower bound results, one might hope that jointly training the embedding and reward can naturally converge to an optimal embedding. However, we show that the joint embedding-reward training is not guaranteed to decrease either tradeoff term per round and admits only fixed points where the reward is constant on the data.

We complement the theory with experiments on synthetic data and three real preference datasets (Jester, Sushi, and MT-Bench) using two pretrained sentence encoders. Across settings, the embedding loss decreases and the agreement cost increases with embedding resolution, and total excess loss is often minimized at an intermediate dimension, consistent with our bounds.

2 Related Work

Alignment with Social Choice Social choice theory offers a rich framework for analyzing how RLHF aggregates diverse human preferences. The optimal solution to the Bradley-Terry loss in RLHF induces a ranking equivalent to the Borda count from social choice theory [26]. Under the assumption that each pairwise comparison is labeled by a single evaluator, RLHF satisfies pairwise majority and Condorcet consistency [30]. More fundamentally, as the annotator pool grows, the probability that the preference profile contains a Condorcet cycle approaches one exponentially fast [16], so reward-based alignment is statistically fragile under general preferences. In the case of a linear social choice model, where voters are linear directions in fixed embedding space, Ge et al. [10] analyze the ranking induced by an optimal linear reward with respect to Pareto optimality (PO) and Pairwise Majority Consistency (PMC). Hollender and Kraiczy [13] show that an appropriate reweighting of the Bradley-Terry loss can enforce PO and PMC, and further analyze clone robustness.

RLHF Reward Modeling, and Failure Modes Reward models are central to RLHF yet remain difficult to train reliably. Casper et al. [4] give a survey summarizing limitations of RLHF, including reward hacking, reward misgeneralization, distribution shift, and the difficulty of representing diverse human values with a single reward. Wang et al. [29] identify noisy preference pairs and out-of-distribution generalization as practical challenges. A number of works identify reward overfitting and overoptimization in reward-model training and propose alternative training methods to standard RLHF [6, 23, 31, 35]. Gao et al. [9] quantify how proxy reward scores diverge from true rewards as optimization pressure increases, following smooth scaling laws.

On the theoretical side, Zhu et al. [34] show that with a linear reward model and a Bradley-Terry (BT) model, MLE can converge. A number of works propose alternatives to the Bradley-Terry model as scalar rewards cannot capture intransitive preferences [2, 19, 32]. Siththaranjan et al. [26] show that implicitly aggregating over hidden contexts during reward modeling can produce counter-intuitive results very different from aggregation via expected utility. Sun et al. [28] establish the asymptotic convergence rate of BT reward models based on deep neural networks using embeddings.

3 RLHF Pipeline

We now introduce the setup used throughout the paper. We consider the standard RLHF pipeline with n human annotators, denoted as $\{1, \dots, n\}$, a prompt space $\mathcal{X}$, and a response space $\mathcal{Y}$.

Preference Model Fix a prompt $x \in \mathcal{X}$ and an annotator $i \in \{1, \dots, n\}$. Annotator i has a latent utility $u_i(x, y) \in \mathbb{R}$ for each response $y \in \mathcal{Y}$. Given two candidate responses $y, y' \in \mathcal{Y}$, the annotator's observed preference is determined by a Gumbel perturbation [3, 21]

$$\tilde{u}_i(x, y) = u_i(x, y) + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \text{Gumbel}(0, \tau), \tag{1}$$

and the annotator prefers y over y' whenever $\tilde{u}_i(x, y) > \tilde{u}_i(x, y')$. We note that this Gumbel noise setup is the standard random utility model (RUM) for discrete choice [17, 18]. Moreover, as the difference of two independent Gumbel$(0, \tau)$ variables is Logistic$(0, \tau)$, we can recover the Bradley-Terry choice probability as the pairwise preference probability. Specifically, the probability that annotator i prefers y over y' given x is then

$$P_i(y \succ y' \mid x) = F\left(\frac{u_i(x, y) - u_i(x, y')}{\tau}\right), \qquad F(t) = \frac{1}{1 + e^{-t}}.$$

We set $\tau = 1$ for the rest of the paper for simplicity.

To aggregate across annotators we define the pooled win probability $\bar{p}(x, y, y') = \frac{1}{n} \sum_{i=1}^{n} P_i(y \succ y' \mid x)$, which gives the average probability that annotators prefer y to y'.

Reward Learning The reward model is a scalar function $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ of the form $r(x, y) = g(\phi(x, y))$, where $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ is a latent embedding and $g : \mathbb{R}^d \to \mathbb{R}$ is a score head. Let ν be a distribution over prompts and, for each x, let μ_x be a (conditional) distribution over responses. The reward is learned by minimizing the following negative log-likelihood of the pooled preferences:

$$\mathcal{L}(r) = \mathbb{E}_{\substack{x \sim \nu \\ (y, y') \sim \mu_x^{\otimes 2}}} \Big[-\bar{p}(x, y, y') \log F(\Delta r(x, y, y')) - \big(1 - \bar{p}(x, y, y')\big) \log\big(1 - F(\Delta r(x, y, y'))\big) \Big],$$

where $\Delta r(x, y, y') = r(x, y) - r(x, y')$.

4 The Representation-Rationalizability Tradeoff

With the setup in place, we now turn to the fundamental question:

What determines how well a reward model can perform?

This section gives the formal answer. We first show an exact decomposition of excess loss, then derive a lower bound that makes the representation-rationalizability tension explicit.

Before stating the main theorem, it is instructive to examine a simple example.

Example 4.1. *Fix one prompt and $\mathcal{Y} = \{y_1, y_2, y_3\}$ with uniform μ. Let responses have binary features $\psi(y_1) = (1, 0), \quad \psi(y_2) = (0, 1), \quad \psi(y_3) = (1, 1)$, and suppose pooled preferences satisfy $\bar{p}(y_1, y_2) = \bar{p}(y_2, y_3) = \bar{p}(y_3, y_1) = \frac{2}{3}$. The role of ϕ is then to decide which responses can receive different rewards, since $r = g \circ \phi$ is constant on each level set of ϕ. Now consider two kinds of embeddings, one that uses only the first bit and one that uses both bits. If $\phi_a(y) = \psi(y)_1$, then $\phi_a(y_1) = \phi_a(y_3)$, so $r(y_1) = r(y_3)$ and necessarily $F(\Delta r(y_1, y_3)) = \frac{1}{2} \neq \bar{p}(y_1, y_3) = \frac{1}{3}$. If $\phi_b(y) = \psi(y)$, then ϕ_b is injective and the representation is no longer the bottleneck. But any scalar reward still satisfies $\Delta r(y_1, y_2) + \Delta r(y_2, y_3) + \Delta r(y_3, y_1) = 0$, whereas the preferences imply*

$$F^{-1}(\bar{p}(y_1, y_2)) + F^{-1}(\bar{p}(y_2, y_3)) + F^{-1}(\bar{p}(y_3, y_1)) = 3 \log 2 \neq 0.$$

Hence no scalar reward can rationalize all three comparisons simultaneously. In summary, a coarse ϕ fails by collapsing distinguishable responses, while a fine ϕ exposes an irreducible cycle that scalar rewards cannot fit.

To better characterize this tradeoff, we turn to the reward excess loss.

The Bayes-optimal loss $\mathcal{L}^*$ represents the best possible performance achievable by any model that perfectly knows the pooled preference probabilities $\bar{p}(x, y, y')$. For any learned reward $r(x, y) = g(\phi(x, y))$, the excess loss $\mathcal{L}(r) - \mathcal{L}^*$ measures how far we fall short of this ideal. Our key observation is that this excess loss admits the following decomposition.

Proposition 4.1. *Define the embedding-induced win probability as*

$$\tilde{p}_\phi(x, y, y') = \mathbb{E}_{(y_0, y_0') \sim \mu_x^{\otimes 2}} \left[\bar{p}(x, y_0, y_0') \mid \phi(x, y_0) = \phi(x, y),\ \phi(x, y_0') = \phi(x, y') \right].$$

Then for any reward $r(x, y) = g(\phi(x, y))$,

$$\mathcal{L}(r) - \mathcal{L}^* = \underbrace{\mathbb{E}_{x \sim \nu} \mathbb{E}_{(y, y') \sim \mu_x^{\otimes 2}} \left[D_{\mathrm{KL}} \left(\bar{p}(x, y, y') \middle\| \tilde{p}_\phi(x, y, y') \right) \right]}_{\mathcal{E}_{\mathrm{emb}}(\phi),\ \textit{embedding loss}}$$
$$+ \underbrace{\mathbb{E}_{x \sim \nu} \mathbb{E}_{(y, y') \sim \mu_x^{\otimes 2}} \left[D_{\mathrm{KL}} \left(\tilde{p}_\phi(x, y, y') \middle\| F(\Delta r(x, y, y')) \right) \right]}_{\mathcal{E}_{\mathrm{agr}}(\phi, r),\ \textit{agreement cost}}.$$

The embedding loss $\mathcal{E}_{\mathrm{emb}}$ captures the information discarded when the embedding collapses distinguishable responses onto the same point. The agreement cost $\mathcal{E}_{\mathrm{agr}}$, which formalize the idea of rationalizability, captures something fundamentally different. Even if the embedding perfectly preserves all distinctions between responses, the pooled preferences may fail to be rationalizable by any scalar reward. The agreement cost measures this irreducibility. When the embedding is injective, i.e., $\tilde{p}_\phi = \bar{p}$, the embedding loss vanishes entirely. The excess loss is then pure agreement cost. Conversely, when the embedding is constant, i.e., $\tilde{p}_\phi = \frac{1}{2}$ everywhere, the embedding loss is large. But in this case the agreement cost vanishes because a constant win probability is consistent with any reward. In between these extremes, both terms are generally positive.

Definition 4.1. *Fix an embedding $\phi_d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$. For each (x, y, y'), define $\eta_d(x, y, y') = \min \{\tilde{p}_{\phi_d}(x, y, y'), 1 - \tilde{p}_{\phi_d}(x, y, y')\}$ and the feasibility class*

$$\mathcal{R}_d = \left\{ (x, y) \mapsto g(\phi_d(x, y)) :\ F\big(g(\phi_d(x, y)) - g(\phi_d(x, y'))\big) \in [\eta_d(x, y, y'), 1 - \eta_d(x, y, y')] \right\}.$$

The restriction to $r \in \mathcal{R}_d$ asks only that the implied pairwise probabilities $F(\Delta r)$ stay within the interval $[\eta_d, 1 - \eta_d]$ that already contains $\tilde{p}_{\phi_d}$, equivalently that $|\Delta r| \leq \log \frac{1 - \eta_d}{\eta_d}$. This is a mild boundedness condition that is automatically satisfied by any reward whose log-odds match those of the embedding-induced probabilities.

The above tradeoff is captured by the following lower bound, and its proof is deferred to Appendix A.

Theorem 4.1. *For any dimension $d \geq 0$, any embedding $\phi_d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, any $r \in \mathcal{R}_d$, and any $\delta \in (0, \frac{1}{2})$,*

$$\mathcal{L}(r) - \mathcal{L}^* \geq 2e_d + \tfrac{\ell_\delta^2}{2} \bar{P}_d^\delta, \tag{2}$$

where $e_d = \mathbb{E}[(\bar{p} - \tilde{p}_{\phi_d})^2]$, $\bar{P}_d^\delta$ is the curvature-weighted Condorcet cycle probability at margin δ defined in Definition 4.3, $\ell_\delta = F^{-1}(\frac{1}{2} + \delta) = \log \frac{1 + 2\delta}{1 - 2\delta} > 0$.

In the following sections, we show that under a mild assumption, e_d is monotonically non-increasing in d, while a lower bound on $\bar{P}_d^\delta$ is non-decreasing in d. This generalizes and quantifies the intuition built in Example 4.1.

Remark 4.1 (Upper bound)**.** *The upper-bound result is deferred to Appendix B. With the same decomposition, the embedding loss term's upper and lower bounds are matched up to constants. The agreement cost term is upper bounded by a quantity of the same form as $\bar{P}_d^\delta$.*

Remark 4.2 (DPO extension)**.** *The same decomposition and cycle-based lower-bound proof extends to DPO. The full setup, formal statements, and assumptions are deferred to Appendix C.*

4.1 Monotonicity Under Representation Refinement

The embedding loss depends on the embedding's expressiveness. Intuitively, increasing the embedding dimension should improve the embedding's ability to distinguish responses, which ought to reduce the embedding loss. This intuition is indeed correct under a mild condition, namely that if a response pair looks identical under $\phi_d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$, it also looks identical under $\phi_{d-1} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{d-1}$.

This condition is satisfied in many settings. If each ϕ_d is a linear embedding $\phi_d(x, y) = U_d^\top \psi(x, y)$ where $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^D$ is a fixed feature map and the column spaces satisfy $\mathrm{col}(U_d) \subseteq \mathrm{col}(U_{d+1})$,

then higher dimensions add new basis functions and the nesting holds automatically. The same holds when ϕ_d concatenates the first d layers or coordinates of a learned representation, so that increasing d only reveals additional information.

We formalize this as follows.

Assumption 4.1. *For each $d \geq 0$, let $\phi_d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ be measurable and let $\mathcal{G}_d$ be the sigma-field generated by the map $\big(\phi_d(x, y), \phi_d(x, y')\big)$ on $\mathcal{X} \times \mathcal{Y}^2$. Assume $\mathcal{G}_d \subseteq \mathcal{G}_{d+1}$ for every $d \geq 0$.*

Under this assumption, the embedding loss is non-increasing in dimension.

Lemma 4.1. *Under Assumption 4.1, define*

$$e_d = \mathbb{E}_{x\sim\nu}\, \mathbb{E}_{(y,y')\sim\mu_x^{\otimes 2}} \left[\left(\bar{p}(x, y, y') - \tilde{p}_{\phi_d}(x, y, y') \right)^2 \right]$$

then $e_{d+1} \leq e_d$ for all $d \geq 0$.

The proof (deferred to Appendix A.1) uses the tower property as conditioning on a coarser sigma-field yields larger mean-squared error. Since $\mathcal{G}_d \subseteq \mathcal{G}_{d+1}$, conditioning on $\mathcal{G}_{d+1}$ and then on $\mathcal{G}_d$ is the same as conditioning on $\mathcal{G}_d$ alone, which cannot reduce the variance of $\bar{p}$.

Having established the monotonicity of the error, we can lower bound the first term as follows.

Proposition 4.2. *For every measurable ϕ_d and every $d \geq 0$, $\mathcal{E}_{\mathrm{emb}}(\phi_d) \geq 2\, e_d$.*

Proof. Pinsker's inequality gives $D_{\mathrm{KL}}(p\|q) \geq 2\,\mathrm{TV}(p, q)^2$ for distributions p, q. For Bernoulli parameters $p, q \in (0, 1)$, the total variation distance is $\mathrm{TV}(\mathrm{Bern}(p), \mathrm{Bern}(q)) = \frac{1}{2}\big(|p - q| + |(1 - p) - (1 - q)|\big) = |p - q|$, hence $D_{\mathrm{KL}}(p\|q) \geq 2(p - q)^2$. Setting $p = \bar{p}(x, y, y')$ and $q = \tilde{p}_{\phi_d}(x, y, y')$ and taking expectations on both sides yields the result. □

4.2 Cycle-Induced Irrecoverability

The embedding loss captures the information lost to coarsening. But even with a perfectly informative embedding, a more subtle failure mode persists. Suppose ϕ_d is injective, so that $\tilde{p}_{\phi_d} = \bar{p}$. The agreement cost $\mathcal{E}_{\mathrm{agr}}(\phi_d, r)$ measures whether the pooled preferences can be simultaneously rationalized by a scalar reward. When they cannot, we say the preferences are irrecoverable.

The main limitation is that scalar rewards impose transitive structure on preferences. Define

$$W_d(x, y, y') := F^{-1}(\tilde{p}_{\phi_d}(x, y, y')). \tag{3}$$

If ϕ_d is injective, then $\tilde{p}_{\phi_d} = \bar{p}$, and any reward of the form $r(x, y) = g(\phi(x, y))$ that fits the pooled preference must satisfy

$$\Delta r(x, y, y') = g(\phi(x, y)) - g(\phi(x, y')) = W_d(x, y, y') = F^{-1}(\bar{p}(x, y, y')).$$

The right-hand side is a difference of potentials, which implies a cyclic consistency condition.

This phenomenon has a direct analogue in social choice theory. Condorcet's paradox [8] describes the situation where majority voting over three alternatives produces a cycle: A beats B, B beats C, but C beats A.

To formalize when this cyclic inconsistency occurs, we introduce the following definition.

Definition 4.2. *For a triplet $(y_1, y_2, y_3) \in \mathcal{Y}^3$ and fixed $x \in \mathcal{X}$, define the cycle sum*

$$C_d(x, y_1, y_2, y_3) = W_d(x, y_1, y_2) + W_d(x, y_2, y_3) + W_d(x, y_3, y_1)\,.$$

Note that if $W_d(x, y, y') = r(x, y) - r(x, y')$ for some reward r, then $C_d(x, y_1, y_2, y_3) = 0$ for all (y_1, y_2, y_3). Conversely, a non-zero cycle sum prevents any scalar reward from rationalizing the preferences.

We then show that this cycle-sum quantity gives a lower bound on the agreement cost.

Theorem 4.2. *With η_d and $\mathcal{R}_d$ as in Definition 4.1, define $\kappa_d(x, y, y') = \eta_d(x, y, y')\big(1 - \eta_d(x, y, y')\big)$ and $\kappa_d^{\min}(x, y_1, y_2, y_3) = \min\big\{\kappa_d(x, y_1, y_2),\, \kappa_d(x, y_2, y_3),\, \kappa_d(x, y_3, y_1)\big\}$. Then, for any $r \in \mathcal{R}_d$,*

$$\mathcal{E}_{\mathrm{agr}}(\phi_d, r) \geq \frac{1}{18}\, \mathbb{E}_{x\sim\nu}\, \mathbb{E}_{(y_1,y_2,y_3)\sim\mu_x^{\otimes 3}} \left[\kappa_d^{\min}(x, y_1, y_2, y_3)\, C_d(x, y_1, y_2, y_3)^2 \right].$$

Theorem 4.2 shows that the agreement cost is lower-bounded by the minimum-weighted mean squared cycle sum.

To interpret the bound more concretely, we connect the mean squared cycle sum to the familiar notion of Condorcet cycles.

Definition 4.3 (δ-Condorcet cycle)**.** *Fix $\delta \in (0, \frac{1}{2})$. A triplet (y_1, y_2, y_3) forms a positive δ-Condorcet cycle under ϕ_d at prompt x if*

$$\tilde{p}_{\phi_d}(x, y_1, y_2), \tilde{p}_{\phi_d}(x, y_2, y_3), \tilde{p}_{\phi_d}(x, y_3, y_1) \geq \tfrac{1}{2} + \delta\,,$$

and a negative one if each inequality is reversed ($\leq \frac{1}{2} - \delta$). Let $\mathcal{CC}_d^\delta$ be the union of these two events. Define the (unweighted) point-level cycle probability and the κ-weighted population-level cycle probability by

$$P_d^\delta(x, y_1, y_2) = \mathbb{P}_{y_3 \sim \mu_x}\left(\mathcal{CC}_d^\delta(x, y_1, y_2, y_3)\right), \quad P_d^\delta = \mathbb{E}_{x\sim\nu}\mathbb{E}_{(y_1,y_2)\sim\mu_x^{\otimes 2}}\left[P_d^\delta(x, y_1, y_2)\right],$$

$$\bar{P}_d^\delta = \mathbb{E}_{x\sim\nu}\,\mathbb{E}_{(y_1,y_2,y_3)\sim\mu_x^{\otimes 3}}\left[\kappa_d^{\min}(x, y_1, y_2, y_3)\,\mathbf{1}_{\mathcal{CC}_d^\delta}(x, y_1, y_2, y_3)\right].$$

Recall that $\ell_\delta = F^{-1}(\frac{1}{2} + \delta) = \log\frac{1+2\delta}{1-2\delta} > 0$.

The population-level cycle probability $\bar{P}_d^\delta$ is the weighted probability that a random triplet forms a Condorcet cycle, where the weight $\kappa_d^{\min}$ downweights cycles whose weakest pair is near 50/50 (which is inherently ambiguous). When a triplet lies in $\mathcal{CC}_d^\delta$, each of its three pairs has a margin of at least δ, so the cycle sum C_d has magnitude at least $3\ell_\delta$.

With this, we lower bound the agreement cost term as follows.

Lemma 4.2. *For every $\delta \in (0, \frac{1}{2})$, $\inf_{r\in\mathcal{R}_d} \mathcal{E}_{\mathrm{agr}}(\phi_d, r) \geq \frac{\ell_\delta^2}{2}\bar{P}_d^\delta$.*

To show how the cycle probability evolves with embedding dimension, we first make the following definition.

Definition 4.4. *Define P_∞^δ as P_d^δ but with $\bar{p}$ in place of $\tilde{p}_{\phi_d}$ in Definition 4.3 (same margin conditions on the three pairs (y_1, y_2), (y_2, y_3), and (y_3, y_1)). Also define*

$$\bar{P}_\infty^{\inf,\delta} = \inf_{d\geq 0}\ \mathbb{E}_{x\sim\nu}\,\mathbb{E}_{(y_1,y_2,y_3)\sim\mu_x^{\otimes 3}}\left[\kappa_d^{\min}(x, y_1, y_2, y_3)\,\mathbf{1}_{\mathcal{CC}_\infty^\delta}(x, y_1, y_2, y_3)\right].$$

Lemma 4.3. *Assume ϕ_0 is constant. For every $\delta \in (0, \frac{1}{4})$. Then*

1. *$\bar{P}_0^\delta = 0$,*
2. *For all $d \geq 0$, $\bar{P}_d^\delta \geq \bar{P}_\infty^{\inf,2\delta} - \frac{3}{\delta^2}e_d$,*
3. *The map $d \mapsto \bar{P}_\infty^{\inf,2\delta} - \frac{3}{\delta^2}e_d$ is non-decreasing in d.*

The first condition says that a constant embedding sees no cycle, which is trivially true as $\tilde{p}_{\phi_0} = \frac{1}{2}$ on every pair. The second condition gives a direct lower bound on the weighted cycle probability that appears in Lemma 4.2. Roughly, if a triplet is a population-level cycle with margin 2δ, it remains a cycle under ϕ_d unless the embedding error e_d is large. The third item follows from the monotonicity of e_d.

4.3 Quantitative Lower Bound Under Hölder Utilities

The generic lower bound from the previous section holds for any utility functions and any embedding ϕ_d, and it only shows that there is a tradeoff for the excess loss. However, it is not immediately clear whether the minimum of the excess loss is unique. In particular, while the lower bound of $\bar{P}_d^\delta$ is non-decreasing in d, $\bar{P}_d^\delta$ might not be.

To derive more interpretable bounds we now contextualize it under two specific assumptions, a Hölder continuity condition on the annotator utilities and an ϵ-separating property on the embedding.

Fix a metric $d_{\mathcal{Y}}$ on $\mathcal{Y}$, such as normalized Hamming or edit distance. We consider the following class of continuous utilities.

Assumption 4.2 (α-Hölder utilities)**.** *There exist $\alpha \in (0,1]$ and $L_u \geq 0$ such that for all $x \in \mathcal{X}$, all $i \in \{1,\dots,n\}$, and all $y, \tilde{y} \in \mathcal{Y}$, $|u_i(x,y) - u_i(x,\tilde{y})| \leq L_u\, d_{\mathcal{Y}}(y,\tilde{y})^{\alpha}$.*

The parameter $\alpha \in (0,1]$ interpolates between Lipschitz ($\alpha = 1$) and uniformly bounded pairwise differences ($\alpha \to 0^+$).

To measure how much ϕ_d coarsens the response space we use the following definition.

Definition 4.5 (ε-separating)**.** *The embedding ϕ_d is ε-*separating *if for all $x \in \mathcal{X}$, $\phi_d(x,y) = \phi_d(x,\tilde{y})$, then $d_{\mathcal{Y}}(y,\tilde{y}) \leq \varepsilon$.*

When $\varepsilon = 0$ this reduces to injectivity. Larger ε allows the embedding to identify responses that are farther apart in $d_{\mathcal{Y}}$. The embedding error e_d is then controlled by the separation parameter. ε-separating embeddings of arbitrarily small ε can be realized by finite-width ReLU MLPs under mild conditions and we defer the proof to Appendix A.3.

Lemma 4.4. *Under Assumption 4.2, if ϕ_d is ε-separating, then $e_d \leq \frac{L_u^2}{4}\varepsilon^{2\alpha}$.*

Combining Lemma 4.4 with the generic lower bound yields the following result.

Theorem 4.3. *Under Assumption 4.2, suppose ϕ_d is ε-separating. Then for any reward $r \in \mathcal{R}_d$ and any $\delta \in (0, \frac{1}{4})$,*

$$\mathcal{L}(r) - \mathcal{L}^* \geq \frac{1}{2}\,\ell_\delta^2\,\bar{P}_\infty^{\inf,2\delta} - \left(\frac{3\,\ell_\delta^2}{2\,\delta^2} - 2\right)^+ \frac{L_u^2}{4\tau^2}\,\varepsilon^{2\alpha},$$

where $(x)^+ = \max(0,x)$.

Hölder regularity converts the lower bound in equation 2 into an explicit quantitative bound with three implications. First, the term $\frac{1}{2}\ell_\delta^2 \bar{P}_\infty^{\inf,2\delta}$ is determined by the population cycle geometry of $\bar{p}$ and therefore defines a data-dependent lower floor that representation design alone cannot eliminate. Second, coarsening can reduce cyclic inconsistency only through an approximation error of order $\varepsilon^{2\alpha}$. Larger α (smoother utilities) yields a more favorable tradeoff, while smaller α limits the benefit of coarsening relative to the induced information loss. Third, δ specifies the cycle margin at which inconsistency is evaluated. Increasing δ restricts attention to stronger cycles while decreasing δ includes weaker cycles but increases sensitivity to representation error through the $1/\delta^2$ scaling. The choice of δ is therefore a calibration parameter tied to the margin distribution of cycles. Consequently, the minimizing representation dimension is intrinsically dataset-dependent.

5 Joint Embedding-Reward Updates

In practice, embeddings and reward heads are typically trained simultaneously, and one might hope that joint training implicitly navigates the representation-rationalizability tradeoff and lands at the sweet spot of our lower bound. We show two structural facts that make this hope fragile.

At iteration t we have a response distribution $\mu_x^{(t)}$ (with joint $\mathcal{D}_t = \mathbb{E}_{x\sim\nu}(\mu_x^{(t)})^{\otimes 2}$), an embedding $\phi_d^{(t)}$ and a reward $r_t = g_t \circ \phi_d^{(t)}$ minimizing $\mathcal{L}(\cdot, \mathcal{D}_t)$, and the soft update for the prompt distribution $d\mu_x^{(t+1)}/d\mu_x^{(t)}(y) \propto e^{\beta r_t(x,y)}$ for a temperature $\beta > 0$. Write $e_d(\mu,\phi) = \mathbb{E}_{x\sim\nu}\mathbb{E}_{(y,y')\sim\mu_x^{\otimes 2}}[(\bar{p} - \tilde{p}_\phi)^2]$ for the embedding error at (μ,ϕ), $P_\infty^{2\delta}(\mu) = \mathbb{P}_{\nu\otimes\mu^{\otimes 3}}(\mathcal{CC}_\infty^{2\delta})$ for the embedding-independent cycle probability under μ, and $\rho_{r,\beta}(x,y) = e^{\beta r(x,y)}/\mathbb{E}_{\mu_x}[e^{\beta r(x,\cdot)}]$ for the soft-update density. Define the tilt covariances

$$\Gamma(\mu,\phi;r,\beta) = \mathbb{E}_\nu\Big[\mathrm{Cov}_{\mu_x^{\otimes 2}}\big((\bar{p} - \tilde{p}_\phi)^2,\ \rho_{r,\beta}(x,y)\rho_{r,\beta}(x,y')\big)\Big],$$

$$\Theta(\mu;r,\beta) = \mathbb{E}_\nu\Big[\mathrm{Cov}_{\mu_x^{\otimes 3}}\big(\mathbf{1}_{\mathcal{CC}_\infty^{2\delta}},\ \rho_{r,\beta}(x,y_1)\rho_{r,\beta}(x,y_2)\rho_{r,\beta}(x,y_3)\big)\Big].$$

Proposition 5.1. *With $\phi = \phi_d^{(t)}$ and r_t measurable, $e_d(\mu^{(t+1)},\phi) - e_d(\mu^{(t)},\phi) = \Gamma_t$, $P_\infty^{2\delta}(\mu^{(t+1)}) - P_\infty^{2\delta}(\mu^{(t)}) = \Theta_t$.*

Proposition 5.1 says the two quantities driving our static lower bound from Lemma 4.2 and Lemma 4.3, namely the embedding error e_d and the cycle probability $P_\infty^{2\delta}$, both shift across rounds via reward-tilt covariances of the same form, differing only in the order of the ρ-product. Round-by-round

optimization fits r_t to $\mathcal{D}_t$ but does not control the sign of the resulting covariances Γ_t, Θ_t, so neither sequence is guaranteed to decrease, and the joint update has no built-in mechanism for shrinking the lower-bound floor.

Proposition 5.2. *At any fixed point (μ^*, ϕ^*, r^*) of the joint update with ϕ^* optimal on μ^* and $r^* \in \mathcal{R}_d(\phi^*)$ loss-optimal on μ^*, both $\Gamma(\mu^*, \phi^*; r^*, \beta) = 0$ and $\Theta(\mu^*; r^*, \beta) = 0$.*

At a fixed point, invariance of μ^* under the soft-update forces $\rho^* = 1$ on $\mathrm{supp}(\mu^*)$, which means r^* is constant on the data and provides no preference signal, and both covariances then vanish trivially. Either the joint dynamics keep moving without converging, or they stabilize only at a degenerate point at which the reward stops ranking responses. In either case, they do not implicitly converge to the sweet spot of the static tradeoff.

6 Experiments

We evaluate our theoretical results on both synthetic and real preference data. For the synthetic experiments, we consider a finite item set with a single prompt x and sample unordered pairs uniformly. There are K annotator types. Each type k follows a logistic pairwise model $P_k(y \succ y' \mid x) = F(U_k(y) - U_k(y'))$, and $\bar{p} = \frac{1}{K}\sum_k P_k$ is the arithmetic mean. The utility functions have three components, a transitive ranking component, a cyclic class component that varies across annotator types, and idiosyncratic noise. When cyclic heterogeneity is present, pooling across types yields a non-Bradley-Terry $\bar{p}$. The embedding ϕ_d is a random projection into binary codes of length $D_{\max}$, retaining the first d coordinates. This satisfies Assumption 4.1 by construction. We report Monte Carlo estimates (for synthetic data) and empirical estimates based on the full dataset (for real datasets) of $\mathcal{E}_{\mathrm{emb}}(\phi_d)$ and $\mathcal{E}_{\mathrm{agr}}(\phi_d, r)$, where r is a Bradley-Terry model with one scalar per embedding group (responses sharing the same ϕ_d value). We compare the result for both fitting a explicit reward (which we refer to as RLHF) and for extracting the implicit reward from DPO (which we just refer to as DPO here). More implementation details are given in Appendix E.

We then evaluate on three real preference datasets: Jester [12], Sushi [15], and MT-Bench [33]. Jester contains individual ratings of jokes on a continuous scale; we construct pairwise preferences by comparing ratings. Sushi provides ranked preferences over sushi types from multiple users, which we convert to pairwise win probabilities. MT-Bench consists of multi-turn dialogue preferences from human evaluations. For each dataset we embed items, namely joke captions, sushi types, or dialogue responses, using a pretrained sentence transformer.

Figure 1 shows the excess loss decomposition as a function of embedding dimension for both RLHF and DPO with the pretrained embedding `all-MiniLM-L6-v2` [24]. The top row gives RLHF on synthetic, Jester, and Sushi, and the bottom row gives the corresponding DPO results. In each panel, the embedding term decreases with dimension while the agreement term increases, consistent with Proposition 4.2 and Theorem 4.2. The total excess loss often has a minimum at an intermediate dimension. We provide more experimental results with pretrained embedding `mpnet` [27] and with the MT-Bench dataset in Appendix E.

Figure 2 provides a geometric visualization of this tradeoff. The top row shows the top 1, 2, and 3 principal components of the embedding used for reward model training on the synthetic data. The dot intensity represents how many data points are mapped to each representation, and the red area highlights the region where points form a Condorcet cycle (Definition 4.3). As the embedding dimension increases, it provides a finer-grained representation of the points, and the Condorcet cycle region grows correspondingly.

7 Conclusion

We show that in RLHF, representation is part of the aggregation problem, not just a modeling detail. The learned embedding determines which preference distinctions are visible to the reward model, inducing a fundamental representation-rationalizability tradeoff. We prove a lower bound on reward excess loss to quantify this tradeoff, where richer representations reduce embedding loss but can increase aggregation inconsistency through cycle structure. Our result extends to DPO and to joint embedding–reward training. In particular, joint updates do not automatically balance the tradeoff,

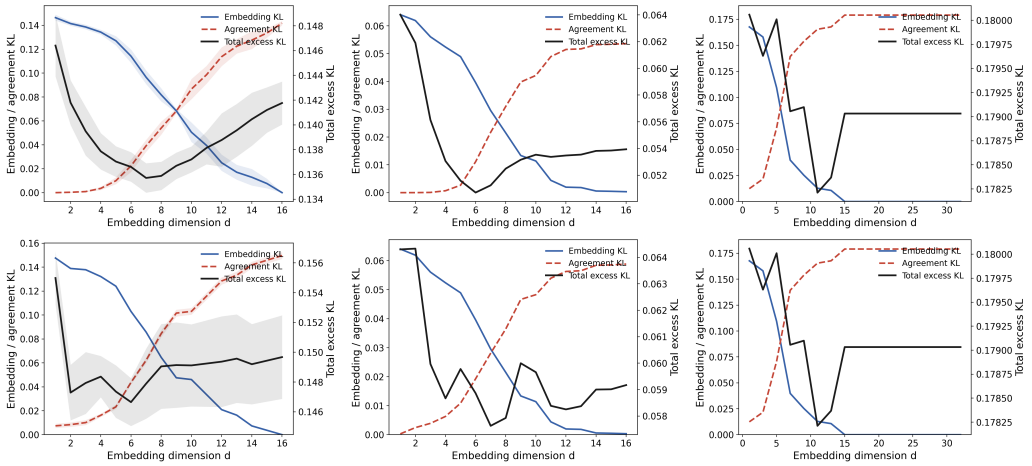


Figure 1: Top row: RLHF on synthetic (left), Jester (middle), and Sushi (right). Bottom row: DPO on synthetic (left), Jester (middle), and Sushi (right).

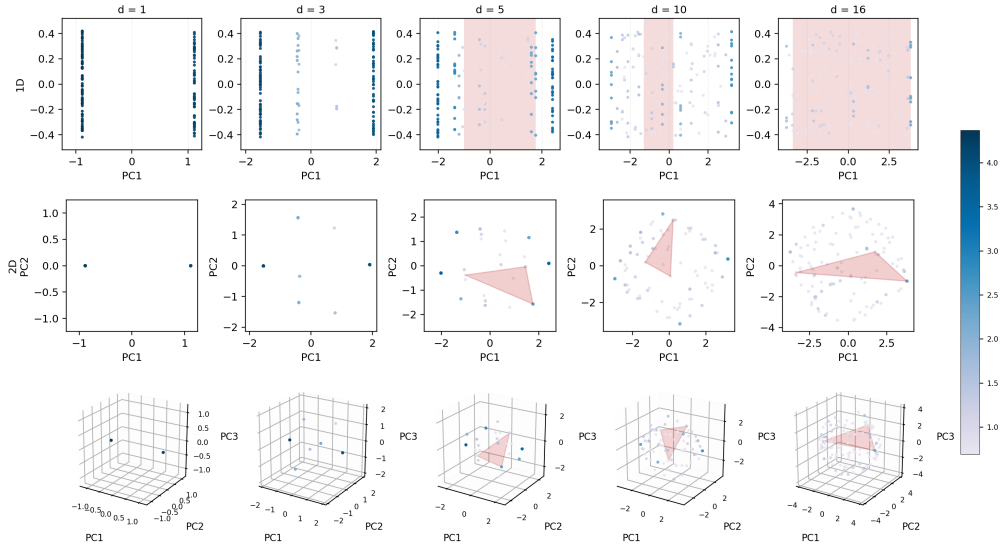


Figure 2: Visualization for the RLHF on the synethtic data. Top 1, 2, and 3 principal components of the embedding used for reward model training on the synthetic data. The intensity of the dots represents how many data points are mapped to each representation, and the red area highlights the region where points form a Condorcet cycle (Definition 4.3).

and can degrade as training proceeds. Experiments on synthetic and real preference datasets confirm our theoretical findings.

AI use statement

In this work, we used generative AI tools to assist in the writing and proofreading of mathematical proofs and to help implement the experimental methods. Additionally, we used generative AI tools to create and edit the experiment code and to edit the manuscript for readability. We have not used generative AI tools to generate research ideas, to discover research topics or identify gaps, or to conduct the research autonomously. We have reviewed all AI-assisted work. The authors checked the AI-assisted proofs, verified and tested the experiment code, and reviewed the polished paper. We take

responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

[1] K. J. Arrow. *Social choice and individual values*. Yale University Press, 2nd edition, 1963.

[2] M. G. Azar, M. Rowland, B. Piot, D. Guo, D. Calandriello, M. Valko, and R. Munos. A general theoretical paradigm to understand learning from human feedback. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024.

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

[4] S. Casper, X. Davies, C. Shi, T. Krendl Gilbert, J. Scheurer, J. Rando Ramirez, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. Bıyık, A. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

[5] V. Conitzer, R. Freedman, J. Heitzig, W. H. Holliday, B. M. Jacobs, N. Lambert, M. Mossé, E. Pacuit, S. Russell, H. Schoelkopf, E. Tewolde, and W. S. Zwicker. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In *Forty-first International Conference on Machine Learning*, 2024.

[6] T. Coste, U. Anwar, R. Kirk, and D. Krueger. Reward model ensembles help mitigate overoptimization. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] J. Dai and E. Fleisig. Mapping social choice theory to RLHF. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024.

[8] N. de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. reprinted by Cambridge University Press, 1784.

[9] L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866, 2023.

[10] L. Ge, D. Halpern, E. Micha, A. D. Procaccia, I. Shapira, Y. Vorobeychik, and J. Wu. Axioms for AI alignment from human feedback. *Advances in Neural Information Processing Systems*, 37:80439–80465, 2024.

[11] A. Gibbard. Manipulation of voting schemes: a general result. *Econometrica*, pages 587–601, 1973.

[12] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.

[13] A. Hollender and S. Kraiczy. Enforcing axioms for AI alignment under loss-based rules. In *The Fourteenth International Conference on Learning Representations*, 2026.

[14] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[15] T. Kamishima. Nantonac collaborative filtering: Recommendation based on order responses. *Knowledge Discovery and Data Mining*, pages 573–579, 2003.

[16] K. Liu, Q. Long, Z. Shi, W. J. Su, and J. Xiao. Statistical impossibility and possibility of aligning LLMs with human preferences: From Condorcet paradox to Nash equilibrium. arXiv preprint arXiv:2503.10990, 2025.

[17] R. D. Luce. *Individual choice behavior: A theoretical analysis*. Wiley, 1959.

[18] D. Mcfadden. Conditional logit analysis of qualitative choice behavior. *Frontier in Econometrics*, 1973.

[19] R. Munos, M. Valko, D. Calandriello, M. G. Azar, M. Rowland, Z. D. Guo, Y. Tang, M. Geist, T. Mesnard, C. Fiegel, A. Michi, M. Selvi, S. Girgin, N. Momchev, O. Bachem, D. J. Mankowitz, D. Precup, and B. Piot. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024.

[20] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Gray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

[21] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

[22] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, pages 53728–53741, 2023.

[23] R. Rafailov, Y. Chittepu, R. Park, H. S. Sikchi, J. Hejna, B. Knox, C. Finn, and S. Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *Advances in Neural Information Processing Systems*, 37:126207–126242, 2024.

[24] N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019.

[25] M. A. Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2): 187–217, 1975.

[26] A. Siththaranjan, C. Laidlaw, and D. Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *The Twelfth International Conference on Learning Representations*, 2024.

[27] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. MPNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[28] H. Sun, Y. Shen, and J.-F. Ton. Rethinking Bradley-Terry models in preference-based reward modeling: Foundations, theory, and alternatives. arXiv preprint arXiv:2411.04991, 2024.

[29] B. Wang, R. Zheng, L. Chen, Y. Liu, S. Dou, C. Huang, W. Shen, S. Jin, E. Zhou, C. Shi, S. Gao, N. Xu, Y. Zhou, X. Fan, Z. Xi, J. Zhao, X. Wang, T. Ji, H. Yan, L. Shen, Z. Chen, T. Gui, Q. Zhang, X. Qiu, X. Huang, Z. Wu, and Y.-G. Jiang. Secrets of RLHF in large language models part II: Reward modeling. arXiv preprint arXiv:2401.06080, 2024.

[30] J. Xiao, Z. Shi, K. Liu, Q. Long, and W. J. Su. Theoretical tensions in RLHF: Reconciling empirical success with inconsistencies in social choice theory. arXiv preprint arXiv:2506.12350, 2025.

[31] R. Yang, R. Ding, Y. Lin, H. Zhang, and T. Zhang. Regularizing hidden states enables learning generalizable reward model for LLMs. *Advances in Neural Information Processing Systems*, 37:62279–62309, 2024.

[32] Y. Zhang, G. Zhang, Y. Wu, K. Xu, and Q. Gu. Beyond Bradley-Terry models: A general preference model for language model alignment. In *International Conference on Machine Learning*, pages 76939–76965, 2025.

[33] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36, 2023.

[34] B. Zhu, M. Jordan, and J. Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067, 2023.

[35] B. Zhu, M. I. Jordan, and J. Jiao. Iterative data smoothing: mitigating reward overfitting and overoptimization in RLHF. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62405–62428, 2024.

A Proofs for Section 4

Proposition 4.1. *Define the embedding-induced win probability as*

$$\tilde{p}_\phi(x, y, y') = \mathbb{E}_{(y_0, y_0') \sim \mu_x^{\otimes 2}} \left[\bar{p}(x, y_0, y_0') \mid \phi(x, y_0) = \phi(x, y),\ \phi(x, y_0') = \phi(x, y') \right].$$

Then for any reward $r(x, y) = g(\phi(x, y))$,

$$\mathcal{L}(r) - \mathcal{L}^* = \underbrace{\mathbb{E}_{x \sim \nu} \mathbb{E}_{(y,y') \sim \mu_x^{\otimes 2}} \left[D_{\mathrm{KL}} \left(\bar{p}(x, y, y') \middle\| \tilde{p}_\phi(x, y, y') \right) \right]}_{\mathcal{E}_{\mathrm{emb}}(\phi),\ \textit{embedding loss}} + \underbrace{\mathbb{E}_{x \sim \nu} \mathbb{E}_{(y,y') \sim \mu_x^{\otimes 2}} \left[D_{\mathrm{KL}} \left(\tilde{p}_\phi(x, y, y') \middle\| F(\Delta r(x, y, y')) \right) \right]}_{\mathcal{E}_{\mathrm{agr}}(\phi, r),\ \textit{agreement cost}}.$$

Proof. For fixed (x, y, y'), minimizing $h(q) = -p \log q - (1 - p) \log(1 - q)$ over $q \in (0, 1)$ with $p = \bar{p}(x, y, y')$ gives the minimizer $q = p$, so the best pointwise fit satisfies $F(\Delta r(x, y, y')) = \bar{p}(x, y, y')$. Plugging $q = p = \bar{p}$ gives the Bayes-optimal loss

$$\mathcal{L}^* = \mathbb{E}_{x \sim \nu} \, \mathbb{E}_{(y,y') \sim \mu_x^{\otimes 2}} \left[H \left(\bar{p}(x, y, y') \right) \right], \qquad H(p) = -p \log p - (1 - p) \log(1 - p).$$

Moreover,

$$\begin{aligned}
&\mathcal{L}(r) - \mathcal{L}^* \\
&= \mathbb{E}_{\substack{x \sim \nu \\ (y,y') \sim \mu_x^{\otimes 2}}} \left[\bar{p}(x, y, y') \log \frac{\bar{p}(x, y, y')}{F(\Delta r(x, y, y'))} + (1 - \bar{p}(x, y, y')) \log \frac{1 - \bar{p}(x, y, y')}{1 - F(\Delta r(x, y, y'))} \right] \\
&= \mathbb{E}_{\substack{x \sim \nu \\ (y,y') \sim \mu_x^{\otimes 2}}} \left[D_{\mathrm{KL}} \left(\bar{p}(x, y, y') \,\middle\|\, F(\Delta r(x, y, y')) \right) \right].
\end{aligned}$$

Define the embedding-induced win probability as

$$\tilde{p}_\phi(x, y, y') = \mathbb{E}_{(y_0, y_0') \sim \mu_x^{\otimes 2}} \left[\bar{p}(x, y_0, y_0') \mid \phi(x, y_0) = \phi(x, y),\ \phi(x, y_0') = \phi(x, y') \right].$$

When ϕ is injective, $\tilde{p}_\phi = \bar{p}$. Otherwise $\tilde{p}_\phi$ averages $\bar{p}$ over pairs that ϕ cannot distinguish.

To decompose D_{KL}, we use

$$\begin{aligned}
\bar{p} \log \frac{\bar{p}}{F(\Delta r)} &= \bar{p} \log \frac{\bar{p}}{\tilde{p}_\phi} + \bar{p} \log \frac{\tilde{p}_\phi}{F(\Delta r)}, \\
(1 - \bar{p}) \log \frac{1 - \bar{p}}{1 - F(\Delta r)} &= (1 - \bar{p}) \log \frac{1 - \bar{p}}{1 - \tilde{p}_\phi} + (1 - \bar{p}) \log \frac{1 - \tilde{p}_\phi}{1 - F(\Delta r)},
\end{aligned}$$

where $\bar{p} = \bar{p}(x, y, y')$, $\tilde{p}_\phi = \tilde{p}_\phi(x, y, y')$, and $F(\Delta r) = F(\Delta r(x, y, y'))$ at each (x, y, y'). Thus

$$\begin{aligned}
&\mathbb{E}_{x \sim \nu} \mathbb{E}_{(y,y') \sim \mu_x^{\otimes 2}} \left[D_{\mathrm{KL}} \left(\bar{p}(x, y, y') \middle\| F(\Delta r(x, y, y')) \right) \right] \\
&= \mathbb{E}_{x \sim \nu} \mathbb{E}_{(y,y') \sim \mu_x^{\otimes 2}} \left[D_{\mathrm{KL}} \left(\bar{p}(x, y, y') \middle\| \tilde{p}_\phi(x, y, y') \right) \right] \\
&\quad + \mathbb{E}_{x \sim \nu} \mathbb{E}_{(y,y') \sim \mu_x^{\otimes 2}} \left[\bar{p} \log \frac{\tilde{p}_\phi}{F(\Delta r)} + (1 - \bar{p}) \log \frac{1 - \tilde{p}_\phi}{1 - F(\Delta r)} \right],
\end{aligned}$$

Let $(Y, Y') \sim \mu_x^{\otimes 2}$ denote the underlying random variables on $\mathcal{X} \times \mathcal{Y}^2$ and set $\mathcal{G} = \sigma(\phi(x, Y), \phi(x, Y'))$.

Then $\tilde{p}_\phi(x, Y, Y') = \mathbb{E}\left[\bar{p}(x, Y, Y') \mid \mathcal{G} \right]$ by the equivalent definition of $\tilde{p}_\phi$ above, hence $\mathbb{E}\left[1 - \bar{p}(x, Y, Y') \mid \mathcal{G} \right] = 1 - \tilde{p}_\phi(x, Y, Y')$. If $r(x, y) = g(\phi(x, y))$, then $F(\Delta r(x, Y, Y'))$ is a measurable function of $(\phi(x, Y), \phi(x, Y'))$, so $F(\Delta r(x, Y, Y'))$ is $\mathcal{G}$-measurable.

Using the tower property, we have

$$\begin{aligned}
\mathbb{E}\left[\bar{p}(x,y,y')\log\frac{\tilde{p}_\phi(x,y,y')}{F(\Delta r(x,y,y'))}\right] &= \mathbb{E}\left[\mathbb{E}\left[\bar{p}(x,y,y')\mid\mathcal{G}\right]\log\frac{\tilde{p}_\phi(x,y,y')}{F(\Delta r(x,y,y'))}\right] \\
&= \mathbb{E}\left[\tilde{p}_\phi(x,y,y')\log\frac{\tilde{p}_\phi(x,y,y')}{F(\Delta r(x,y,y'))}\right], \\
\mathbb{E}\left[(1-\bar{p}(x,y,y'))\log\frac{1-\tilde{p}_\phi(x,y,y')}{1-F(\Delta r(x,y,y'))}\right] &= \mathbb{E}\left[\mathbb{E}\left[1-\bar{p}(x,y,y')\mid\mathcal{G}\right]\log\frac{1-\tilde{p}_\phi(x,y,y')}{1-F(\Delta r(x,y,y'))}\right] \\
&= \mathbb{E}\left[(1-\tilde{p}_\phi(x,y,y'))\log\frac{1-\tilde{p}_\phi(x,y,y')}{1-F(\Delta r(x,y,y'))}\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathcal{L}(r)-\mathcal{L}^* = \mathbb{E}_{x\sim\nu}\,\mathbb{E}_{(y,y')\sim\mu_x^{\otimes 2}}\left[D_{\mathrm{KL}}\left(\bar{p}(x,y,y')\,\middle\|\,\tilde{p}_\phi(x,y,y')\right)\right] \\
+\,\mathbb{E}_{x\sim\nu}\,\mathbb{E}_{(y,y')\sim\mu_x^{\otimes 2}}\left[D_{\mathrm{KL}}\left(\tilde{p}_\phi(x,y,y')\,\middle\|\,F(\Delta r(x,y,y'))\right)\right].
\end{aligned}$$

□

A.1 Proofs for SubSection 4.1

Lemma 4.1. *Under Assumption 4.1, define*

$$e_d = \mathbb{E}_{x\sim\nu}\,\mathbb{E}_{(y,y')\sim\mu_x^{\otimes 2}}\left[\left(\bar{p}(x,y,y')-\tilde{p}_{\phi_d}(x,y,y')\right)^2\right]$$

then $e_{d+1}\le e_d$ for all $d\ge 0$.

Proof. Let $X=\bar{p}(x,y,y')$ under $x\sim\nu$, $(y,y')\sim\mu_x^{\otimes 2}$. Then $\mathbb{E}\left[X\mid\mathcal{G}_d\right]=\tilde{p}_{\phi_d}(x,y,y')$ a.s., hence $e_d=\|X-\mathbb{E}\left[X\mid\mathcal{G}_d\right]\|_2^2$.

By Assumption 4.1, $\mathcal{G}_d\subseteq\mathcal{G}_{d+1}$, thus the tower property gives $\mathbb{E}\left[X\mid\mathcal{G}_d\right]=\mathbb{E}\left[\mathbb{E}\left[X\mid\mathcal{G}_{d+1}\right]\mid\mathcal{G}_d\right]$, and

$$\begin{aligned}
\|X-\mathbb{E}\left[X\mid\mathcal{G}_d\right]\|_2^2 &= \|X-\mathbb{E}\left[X\mid\mathcal{G}_{d+1}\right]\|_2^2+\|\mathbb{E}\left[X\mid\mathcal{G}_{d+1}\right]-\mathbb{E}\left[X\mid\mathcal{G}_d\right]\|_2^2 \\
&\ge \|X-\mathbb{E}\left[X\mid\mathcal{G}_{d+1}\right]\|_2^2.
\end{aligned}$$

□

A.2 Proofs for SubSection 4.2

Theorem 4.2. *With η_d and $\mathcal{R}_d$ as in Definition 4.1, define $\kappa_d(x,y,y') = \eta_d(x,y,y')\big(1-\eta_d(x,y,y')\big)$ and $\kappa_d^{\min}(x,y_1,y_2,y_3) = \min\big\{\kappa_d(x,y_1,y_2),\, \kappa_d(x,y_2,y_3),\, \kappa_d(x,y_3,y_1)\big\}$. Then, for any $r \in \mathcal{R}_d$,*

$$\mathcal{E}_{\mathrm{agr}}(\phi_d, r) \geq \frac{1}{18}\, \mathbb{E}_{x\sim\nu}\, \mathbb{E}_{(y_1,y_2,y_3)\sim\mu_x^{\otimes 3}}\Big[\kappa_d^{\min}(x,y_1,y_2,y_3)\, C_d(x,y_1,y_2,y_3)^2\Big].$$

Proof. For each (x,y,y') set $\eta = \eta_d(x,y,y') \in (0,\frac{1}{2}]$, $\kappa = \kappa_d(x,y,y') = \eta(1-\eta)$, $p = \tilde{p}_{\phi_d}(x,y,y')$, and

$$I_{d,(x,y,y')} = \Big[\log\tfrac{\eta}{1-\eta},\, \log\tfrac{1-\eta}{\eta}\Big].$$

The map $f_p(t) = D_{\mathrm{KL}}(p\|F(t))$ satisfies $f_p'(t) = F(t) - p$ and $f_p''(t) = F(t)(1-F(t))$. On $I_{d,(x,y,y')}$, $F(t) \in [\eta, 1-\eta]$, so $f_p''(t) \geq \kappa$. The unique minimizer $t_p = F^{-1}(p)$ lies in $I_{d,(x,y,y')}$ with $f_p(t_p) = 0$, and strong convexity gives

$$f_p(t) \geq \tfrac{\kappa}{2}\,(t-t_p)^2, \qquad \forall t \in I_{d,(x,y,y')}.$$

Fix $r \in \mathcal{R}_d$ and set $t = \Delta r(x,y,y')$. By definition of $\mathcal{R}_d$, $F(t) \in [\eta, 1-\eta]$, so $t \in I_{d,(x,y,y')}$. Applying the pointwise bound and integrating,

$$\mathcal{E}_{\mathrm{agr}}(\phi_d, r) \geq \frac{1}{2}\, \mathbb{E}_{x\sim\nu}\, \mathbb{E}_{(y,y')\sim\mu_x^{\otimes 2}}\Big[\kappa_d(x,y,y')\,\big(W_d(x,y,y') - \Delta r(x,y,y')\big)^2\Big]. \tag{4}$$

For $(y_1,y_2,y_3) \sim \mu_x^{\otimes 3}$ write $a_{ij}(x,y_i,y_j) = W_d(x,y_i,y_j) - \Delta r(x,y_i,y_j)$. Since the cyclic sum of Δr vanishes,

$$C_d(x,y_1,y_2,y_3) = W_d(x,y_1,y_2) + W_d(x,y_2,y_3) + W_d(x,y_3,y_1) = a_{12} + a_{23} + a_{31}.$$

For each triplet, $(a_{12}+a_{23}+a_{31})^2 \leq 3(a_{12}^2 + a_{23}^2 + a_{31}^2)$, and $\kappa_d^{\min} \leq \kappa_d(x,y_i,y_j)$ for each pair $(i,j) \in \{(1,2),(2,3),(3,1)\}$, so

$$\begin{aligned}\kappa_d^{\min}(x,y_1,y_2,y_3)\, C_d^2 &\leq 3\,\kappa_d^{\min}(x,y_1,y_2,y_3)\,(a_{12}^2 + a_{23}^2 + a_{31}^2)\\ &\leq 3\big(\kappa_d(x,y_1,y_2)\, a_{12}^2 + \kappa_d(x,y_2,y_3)\, a_{23}^2 + \kappa_d(x,y_3,y_1)\, a_{31}^2\big).\end{aligned} \tag{5}$$

Take $\mathbb{E}_{(y_1,y_2,y_3)\sim\mu_x^{\otimes 3}}$ on both sides. Each a_{ij}^2 is a function of only two of the three coordinates, so by exchangeability of $\mu_x^{\otimes 3}$ each of the three terms on the right marginalizes to the same pair expectation $\mathbb{E}_{(y,y')\sim\mu_x^{\otimes 2}}\big[\kappa_d(x,y,y')(W_d - \Delta r)^2\big]$. Therefore

$$\mathbb{E}_{(y_1,y_2,y_3)\sim\mu_x^{\otimes 3}}\big[\kappa_d^{\min}\, C_d^2\big] \leq 9\, \mathbb{E}_{(y,y')\sim\mu_x^{\otimes 2}}\big[\kappa_d(x,y,y')\,(W_d - \Delta r)^2\big]. \tag{6}$$

Taking $\mathbb{E}_x$ in equation 6 and combining with equation 4,

$$\mathcal{E}_{\mathrm{agr}}(\phi_d, r) \geq \frac{1}{18}\, \mathbb{E}_{x\sim\nu}\, \mathbb{E}_{(y_1,y_2,y_3)\sim\mu_x^{\otimes 3}}\Big[\kappa_d^{\min}(x,y_1,y_2,y_3)\, C_d(x,y_1,y_2,y_3)^2\Big].$$

Since the right-hand side does not depend on $r \in \mathcal{R}_d$, the bound applies to $\inf_{r\in\mathcal{R}_d} \mathcal{E}_{\mathrm{agr}}(\phi_d, r)$. □

Lemma 4.2. *For every $\delta \in (0, \frac{1}{2})$, $\inf_{r\in\mathcal{R}_d} \mathcal{E}_{\mathrm{agr}}(\phi_d, r) \geq \frac{\ell_\delta^2}{2}\bar{P}_d^\delta$.*

Proof. Since $(F^{-1})'(p) = 1/(p(1-p)) > 0$, the map F^{-1} is strictly increasing on $(0,1)$.

On a positive δ-cycle, each term has $\tilde{p}_{\phi_d} \geq \frac{1}{2} + \delta$, so

$$\begin{aligned}C_d(x,y_1,y_2,y_3) &= F^{-1}\left(\tilde{p}_{\phi_d}(x,y_1,y_2)\right) + F^{-1}\left(\tilde{p}_{\phi_d}(x,y_2,y_3)\right) + F^{-1}\left(\tilde{p}_{\phi_d}(x,y_3,y_1)\right)\\ &\geq 3F^{-1}\left(\frac{1}{2}+\delta\right) = 3\ell_\delta > 0\,.\end{aligned}$$

Similarly, on a negative δ-cycle each term is at most $-\ell_\delta$ and $C_d \leq -3\ell_\delta$. Therefore $C_d^2 \geq 9\ell_\delta^2$ on $\mathcal{CC}_d^\delta$, hence

$$\mathbb{E}_{x,(y_1,y_2,y_3)}\big[\kappa_d^{\min}\, C_d^2\big] \geq \mathbb{E}_{x,(y_1,y_2,y_3)}\big[\kappa_d^{\min}\, C_d^2\, \mathbf{1}_{\mathcal{CC}_d^\delta}\big] \geq 9\ell_\delta^2\, \mathbb{E}_{x,(y_1,y_2,y_3)}\big[\kappa_d^{\min}\, \mathbf{1}_{\mathcal{CC}_d^\delta}\big] = 9\ell_\delta^2\, \bar{P}_d^\delta.$$

Combining with Theorem 4.2,

$$\inf_{r\in\mathcal{R}_d} \mathcal{E}_{\mathrm{agr}}(\phi_d, r) \geq \tfrac{1}{18}\,\mathbb{E}\big[\kappa_d^{\min}\, C_d^2\big] \geq \tfrac{9\ell_\delta^2}{18}\,\bar{P}_d^\delta = \tfrac{\ell_\delta^2}{2}\,\bar{P}_d^\delta.$$

□

Lemma 4.3. *Assume ϕ_0 is constant. For every $\delta \in (0, \frac{1}{4})$. Then*

1. *$\bar{P}_0^\delta = 0$,*
2. *For all $d \geq 0$, $\bar{P}_d^\delta \geq \bar{P}_\infty^{\inf,2\delta} - \frac{3}{\delta^2} e_d$,*
3. *The map $d \mapsto \bar{P}_\infty^{\inf,2\delta} - \frac{3}{\delta^2} e_d$ is non-decreasing in d.*

Proof. **(1)** Under ϕ_0, conditioning is trivial and $\tilde{p}_{\phi_0}(x, y, y') = \mathbb{E}_{(y_0,y_0')\sim\mu_x^{\otimes 2}}[\bar{p}(x, y_0, y_0')]$. Symmetry of $\mu_x^{\otimes 2}$ in (y_0, y_0') and $\bar{p}(x, y, y') + \bar{p}(x, y', y) = 1$ imply $\tilde{p}_{\phi_0} = \frac{1}{2}$ everywhere, hence $\mathcal{CC}_0^\delta$ has probability zero and therefore $\bar{P}_0^\delta = 0$.

(2) Let $\mathcal{CC}_\infty^{2\delta}$ be the population analogue of Definition 4.3 with margin 2δ and $\bar{p}$ in place of $\tilde{p}_{\phi_d}$, so $P_\infty^{2\delta} = \mathbb{P}(\mathcal{CC}_\infty^{2\delta})$.

If a triplet is in $\mathcal{CC}_{\infty,+}^{2\delta}$ and each of the three pairs (y_1, y_2), (y_2, y_3), (y_3, y_1) satisfies $|\tilde{p}_{\phi_d} - \bar{p}| \leq \delta$, then $\tilde{p}_{\phi_d} \geq \frac{1}{2} + \delta$ on each pair, so the triplet lies in $\mathcal{CC}_d^\delta$. The same implication holds for $\mathcal{CC}_{\infty,-}^{2\delta}$. Hence every triplet in $\mathcal{CC}_\infty^{2\delta} \setminus \mathcal{CC}_d^\delta$ has at least one of those three pairs with $|\tilde{p}_{\phi_d} - \bar{p}| > \delta$, i.e.

$$\mathbf{1}_{\mathcal{CC}_\infty^{2\delta}} \leq \mathbf{1}_{\mathcal{CC}_d^\delta} + \sum_{(i,j)\in\{(1,2),(2,3),(3,1)\}} \mathbf{1}_{\{|\tilde{p}_{\phi_d}(y_i,y_j) - \bar{p}(y_i,y_j)|>\delta\}}.$$

Multiplying by $\kappa_d^{\min} \in [0, 1/4]$ and taking expectations,

$$\mathbb{E}\left[\kappa_d^{\min} \mathbf{1}_{\mathcal{CC}_\infty^{2\delta}}\right] \leq \bar{P}_d^\delta + 3\,\mathbb{P}\left(|\tilde{p}_{\phi_d}(y, y') - \bar{p}(y, y')| > \delta\right).$$

By definition of $\bar{P}_\infty^{\inf,2\delta}$, we have $\mathbb{E}[\kappa_d^{\min} \mathbf{1}_{\mathcal{CC}_\infty^{2\delta}}] \geq \bar{P}_\infty^{\inf,2\delta}$, and Markov's inequality gives $\mathbb{P}(|\tilde{p}_{\phi_d} - \bar{p}| > \delta) \leq \delta^{-2}\|\tilde{p}_{\phi_d} - \bar{p}\|_2^2 \leq e_d/\delta^2$, which proves item (2).

(3) If $d_1 \leq d_2$ then $e_{d_1} \geq e_{d_2}$ by Lemma 4.1, so $\bar{P}_\infty^{\inf,2\delta} - \frac{3}{\delta^2} e_{d_1} \leq \bar{P}_\infty^{\inf,2\delta} - \frac{3}{\delta^2} e_{d_2}$. □

A.3 Proofs for SubSection 4.3

Constructing ε-separating embeddings. For completeness we note that ε-separating embeddings of arbitrarily small ε can be realized by finite-width ReLU MLPs whenever the response space is compact and the embedding map $y \mapsto e(y)$ is continuous and injective in y.

Let $\mathcal{N}_{\mathcal{Y}}(\epsilon)$ denote the ϵ-covering number of $(\mathcal{Y}, d_{\mathcal{Y}})$. A ReLU MLP is a function $f : \mathbb{R}^D \to \mathbb{R}^m$ of the form $f = W_L \circ \sigma \circ W_{L-1} \circ \cdots \circ \sigma \circ W_1$, where W_ℓ are affine maps and $\sigma(t) = \max(0, t)$ is applied coordinate-wise.

Theorem A.1 (Universal approximation [14])**.** *Let $K \subset \mathbb{R}^D$ be compact and $g : K \to \mathbb{R}^m$ be continuous. For any $\eta > 0$, there exists a ReLU MLP $f : \mathbb{R}^D \to \mathbb{R}^m$ of finite depth and width such that $\sup_{x \in K} \|f(x) - g(x)\|_\infty \le \eta$.*

Lemma A.1. *Let $(\mathcal{Y}, d_{\mathcal{Y}})$ be compact and $e : \mathcal{Y} \to \mathbb{R}^D$ be continuous and injective. Set $N_\varepsilon = \mathcal{N}_{\mathcal{Y}}(\varepsilon/4)$. For every $d \ge N_\varepsilon$, there exists a ReLU MLP $f : \mathbb{R}^D \to \mathbb{R}^d$ such that the embedding $y \mapsto f(e(y))$ is ε-separating.*

Lemma 4.4. *Under Assumption 4.2, if ϕ_d is ε-separating, then $e_d \le \frac{L_u^2}{4} \varepsilon^{2\alpha}$.*

Proof. Fix (x, y, y'). For any (y_0, y_0') in the conditioning event defining $\tilde{p}_{\phi_d}(x, y, y')$, the ε-separating condition gives $d_{\mathcal{Y}}(y_0, y) \le \varepsilon$ and $d_{\mathcal{Y}}(y_0', y') \le \varepsilon$.

Recalling $\bar{p}(x, y, y') = \frac{1}{n} \sum_{i=1}^n F\big(\frac{u_i(x,y) - u_i(x,y')}{\tau}\big)$ and $F'(t) = F(t)(1 - F(t)) \le \frac{1}{4}$, the mean-value theorem gives

$$\begin{aligned} |\bar{p}(x, y_0, y_0') - \bar{p}(x, y, y')| &\le \frac{1}{4\tau n} \sum_{i=1}^n \big[|u_i(x, y_0) - u_i(x, y)| + |u_i(x, y_0') - u_i(x, y')| \big] \\ &\le \frac{L_u}{4\tau} \big[d_{\mathcal{Y}}(y_0, y)^\alpha + d_{\mathcal{Y}}(y_0', y')^\alpha \big] \le \frac{L_u}{2}\, \varepsilon^\alpha. \end{aligned}$$

Since this holds for every (y_0, y_0') in the conditioning event, the conditional expectation satisfies $|\tilde{p}_{\phi_d}(x, y, y') - \bar{p}(x, y, y')| \le \frac{L_u}{2} \varepsilon^\alpha$, whence $e_d = \mathbb{E}[(\bar{p} - \tilde{p}_{\phi_d})^2] \le \frac{L_u^2}{4} \varepsilon^{2\alpha}$. □

Theorem 4.3. *Under Assumption 4.2, suppose ϕ_d is ε-separating. Then for any reward $r \in \mathcal{R}_d$ and any $\delta \in (0, \frac{1}{4})$,*

$$\mathcal{L}(r) - \mathcal{L}^* \ge \frac{1}{2}\, \ell_\delta^2\, \bar{P}_\infty^{\inf, 2\delta} - \left(\frac{3\, \ell_\delta^2}{2\, \delta^2} - 2 \right)^+ \frac{L_u^2}{4\tau^2}\, \varepsilon^{2\alpha},$$

where $(x)^+ = \max(0, x)$.

Proof. Combining Lemma 4.2 with Lemma 4.3 (i.e., $\bar{P}_d^\delta \ge \bar{P}_\infty^{\inf, 2\delta} - \frac{3}{\delta^2} e_d$) yields

$$\inf_{r \in \mathcal{R}_d} \mathcal{E}_{\mathrm{agr}}(\phi_d, r) \ge \tfrac{\ell_\delta^2}{2} \bar{P}_d^\delta \ge \tfrac{\ell_\delta^2}{2} \Big(\bar{P}_\infty^{\inf, 2\delta} - \tfrac{3}{\delta^2} e_d \Big).$$

Together with Proposition 4.2 ($\mathcal{E}_{\mathrm{emb}} \ge 2 e_d$) and the decomposition $\mathcal{L}(r) - \mathcal{L}^* = \mathcal{E}_{\mathrm{emb}} + \mathcal{E}_{\mathrm{agr}}$,

$$\begin{aligned} \mathcal{L}(r) - \mathcal{L}^* &\ge 2\, e_d + \tfrac{\ell_\delta^2}{2} \Big(\bar{P}_\infty^{\inf, 2\delta} - \tfrac{3}{\delta^2} e_d \Big) \\ &= (2 - \beta)\, e_d + \tfrac{1}{2}\, \ell_\delta^2\, \bar{P}_\infty^{\inf, 2\delta}, \end{aligned}$$

where $\beta = \frac{3\, \ell_\delta^2}{2\, \delta^2}$.

Define the affine function $f(t) = (2 - \beta)\, t + \frac{1}{2}\, \ell_\delta^2\, \bar{P}_\infty^{\inf, 2\delta}$ for $t \ge 0$. By Lemma 4.4, $e_d \in [0,\, \frac{L_u^2}{4\tau^2} \varepsilon^{2\alpha}]$, so $\mathcal{L}(r) - \mathcal{L}^* \ge \min_{t \in [0,\, L_u^2 \varepsilon^{2\alpha} / (4\tau^2)]} f(t)$. Since f is affine, the minimum is attained at an endpoint. At $t = 0$ if $\beta \le 2$ and at $t = \frac{L_u^2}{4\tau^2} \varepsilon^{2\alpha}$ if $\beta > 2$. In either case,

$$\mathcal{L}(r) - \mathcal{L}^* \ge \tfrac{1}{2}\, \ell_\delta^2\, \bar{P}_\infty^{\inf, 2\delta} - (\beta - 2)^+\, \tfrac{L_u^2}{4\tau^2}\, \varepsilon^{2\alpha}.$$

Substituting β completes the proof. □

Lemma A.1. *Let $(\mathcal{Y}, d_{\mathcal{Y}})$ be compact and $e : \mathcal{Y} \to \mathbb{R}^D$ be continuous and injective. Set $N_\varepsilon = \mathcal{N}_{\mathcal{Y}}(\varepsilon/4)$. For every $d \geq N_\varepsilon$, there exists a ReLU MLP $f : \mathbb{R}^D \to \mathbb{R}^d$ such that the embedding $y \mapsto f(e(y))$ is ε-separating.*

Proof. Let $\{y_1, \ldots, y_{N_\varepsilon}\}$ be an $\varepsilon/4$-net in $\mathcal{Y}$, so the open balls $B_i = \{y \in \mathcal{Y} : d_{\mathcal{Y}}(y, y_i) < \varepsilon/4\}$ cover $\mathcal{Y}$.

For $i = 1, \ldots, N_\varepsilon$, define the continuous tent function $u_i(y) = (\varepsilon/4 - d_{\mathcal{Y}}(y, y_i))_+$, which is positive precisely on B_i. Let $S(y) = \sum_{j=1}^{N_\varepsilon} u_j(y)$; the cover property gives $S(y) > 0$ for every $y \in \mathcal{Y}$, and by compactness $\inf_{y \in \mathcal{Y}} S(y) > 0$. Define

$$\hat{\psi} : \mathcal{Y} \to \Delta^{N_\varepsilon - 1} \subseteq \mathbb{R}^{N_\varepsilon}, \qquad \hat{\psi}(y)_i = u_i(y)/S(y).$$

Then $\hat{\psi}$ is continuous on $\mathcal{Y}$, $\hat{\psi}(y)_i > 0$ iff $y \in B_i$, and $\sum_i \hat{\psi}(y)_i = 1$ for every y.

Since $\bar{\psi} := \hat{\psi} \circ e^{-1}$ is continuous, we can extend it to a continuous function $\bar{\psi} : \mathbb{R}^D \to \mathbb{R}^{N_\varepsilon}$.

Apply Theorem A.1 to $\bar{\psi}$ on the compact set $e(\mathcal{Y})$ with tolerance $\eta = \frac{1}{3N_\varepsilon}$. There exists a ReLU MLP $\tilde{f} : \mathbb{R}^D \to \mathbb{R}^{N_\varepsilon}$ such that $\sup_{z \in e(\mathcal{Y})} \|\tilde{f}(z) - \bar{\psi}(z)\|_\infty \leq \eta$. If $d > N_\varepsilon$, pad $\tilde{f}$ with zero coordinates to obtain $f : \mathbb{R}^D \to \mathbb{R}^d$.

Suppose $f(e(y)) = f(e(\tilde{y}))$. The padded coordinates carry no information, so $\tilde{f}(e(y)) = \tilde{f}(e(\tilde{y}))$. Hence,

$$\begin{aligned}\|\hat{\psi}(y) - \hat{\psi}(\tilde{y})\|_\infty &= \|\bar{\psi}(e(y)) - \bar{\psi}(e(\tilde{y}))\|_\infty \\ &\leq \|\bar{\psi}(e(y)) - \tilde{f}(e(y))\|_\infty + \|\tilde{f}(e(y)) - \tilde{f}(e(\tilde{y}))\|_\infty + \|\tilde{f}(e(\tilde{y})) - \bar{\psi}(e(\tilde{y}))\|_\infty .\end{aligned}$$

Since $\tilde{f}(e(y)) = \tilde{f}(e(\tilde{y}))$, and both $e(y), e(\tilde{y}) \in e(\mathcal{Y})$, we have $\|\tilde{f}(e(y)) - \tilde{f}(e(\tilde{y}))\|_\infty = 0$, $\|\bar{\psi}(e(y)) - \tilde{f}(e(y))\|_\infty \leq \eta$, $\|\tilde{f}(e(\tilde{y})) - \bar{\psi}(e(\tilde{y}))\|_\infty \leq \eta$. Therefore

$$\|\hat{\psi}(y) - \hat{\psi}(\tilde{y})\|_\infty \leq 2\eta = \frac{2}{3N_\varepsilon}.$$

We claim that there exists i^* with $\hat{\psi}(y)_{i^*} > 0$ and $\hat{\psi}(\tilde{y})_{i^*} > 0$. Suppose for contradiction that $\hat{\psi}(y)$ and $\hat{\psi}(\tilde{y})$ have disjoint supports. Both are probability vectors in $\mathbb{R}^{N_\varepsilon}$, so $\max_i \hat{\psi}(y)_i \geq 1/N_\varepsilon$ at some index i_0. At that index, $\hat{\psi}(\tilde{y})_{i_0} = 0$, hence $|\hat{\psi}(y)_{i_0} - \hat{\psi}(\tilde{y})_{i_0}| \geq 1/N_\varepsilon > 2/(3N_\varepsilon)$, a contradiction.

Therefore $y \in B_{i^*}$ and $\tilde{y} \in B_{i^*}$, so $d_{\mathcal{Y}}(y, \tilde{y}) \leq d_{\mathcal{Y}}(y, y_{i^*}) + d_{\mathcal{Y}}(y_{i^*}, \tilde{y}) < \varepsilon/2 \leq \varepsilon$. □

B Excess Loss Upper bound

Theorem B.1. *Suppose $\tilde{p}_{\phi_d}(x, y, y') \in (0, 1)$ almost surely and define*

$$\kappa_d(x, y, y') = \tilde{p}_{\phi_d}(x, y, y')\big(1 - \tilde{p}_{\phi_d}(x, y, y')\big).$$

Define $\nu_{d,x} := \phi_{d\star}\mu_x$ and $\nu_{d,x}^{\otimes 2} := \nu_{d,x} \otimes \nu_{d,x}$. Define the score function $W_d(x, y, y') = F^{-1}(\tilde{p}_{\phi_d}(x, y, y'))$ and the subspace

$$\mathcal{V}_x = \overline{\big\{\Delta h \text{ where } h : \mathcal{Z}_d(x) \to \mathbb{R}\big\}}^{L^2(\nu_{d,x}^{\otimes 2})}, \qquad (\Delta h)(z, z') = h(z) - h(z'),$$

let $\Pi_{\mathcal{V}_x}$ denote orthogonal projection onto $\mathcal{V}_x \subseteq L^2(\nu_{d,x}^{\otimes 2})$, and set

$$U_x := \widetilde{W}_{d,x} - \Pi_{\mathcal{V}_x}\widetilde{W}_{d,x}, \qquad \widetilde{W}_{d,x}(z, z') = W_d(x, y, y') \text{ where } z = \phi_d(x, y)\,, z' = \phi_d(x, y').$$

Then for every $\varepsilon > 0$, there exists a measurable prompt-indexed score head $g_x^\varepsilon : \mathbb{R}^d \to \mathbb{R}$ such that $r_d^\varepsilon(x, y) = g_x^\varepsilon(\phi_d(x, y))$ satisfies

$$\mathcal{L}(r_d^\varepsilon) - \mathcal{L}^* \le \mathbb{E}\left[\frac{\big(\bar{p} - \tilde{p}_{\phi_d}\big)^2}{\kappa_d}\right] + \frac{1}{8}\,\mathbb{E}_x\big[\|U_x\|^2_{L^2(\nu_{d,x}^{\otimes 2})}\big] + \varepsilon\,.$$

Proof. Throughout let $(y, y') \sim \mu_x^{\otimes 2}$, $(z, z') = \big(\phi_d(x, y), \phi_d(x, y')\big) \sim \nu_{d,x}^{\otimes 2}$, and $\nu_{d,x} = \phi_{d\star}\mu_x$ be the pushforward.

For $u \in [0, 1]$ and $v \in (0, 1)$, the elementary inequality $\log(1 + t) \le t$ yields $D_{\mathrm{KL}}(u\|v) \le \chi^2(u, v) = \frac{(u-v)^2}{v(1-v)}$. With $v = \tilde{p}_{\phi_d}(x, y, y')$ and $\kappa_d = v(1 - v)$,

$$\mathcal{E}_{\mathrm{emb}}(\phi_d) = \mathbb{E}\big[D_{\mathrm{KL}}(\bar{p}\|\tilde{p}_{\phi_d})\big] \le \mathbb{E}\left[\frac{(\bar{p} - \tilde{p}_{\phi_d})^2}{\kappa_d}\right].$$

Define $W_d(x, y, y') = F^{-1}(\tilde{p}_{\phi_d}(x, y, y'))$ and

$$\mathcal{V}_x = \overline{\big\{\Delta h : h : \mathcal{Z}_d(x) \to \mathbb{R}\big\}}^{L^2(\nu_{d,x}^{\otimes 2})}, \qquad (\Delta h)(z, z') = h(z) - h(z').$$

Let $\Pi_{\mathcal{V}_x}$ denote orthogonal projection onto $\mathcal{V}_x \subseteq L^2(\nu_{d,x}^{\otimes 2})$ and set $U_x = \widetilde{W}_{d,x} - \Pi_{\mathcal{V}_x}\widetilde{W}_{d,x}$, where $\widetilde{W}_{d,x}(z, z') = W_d(x, y, y')$ is well-defined because W_d is $\sigma(\phi_d(x, y), \phi_d(x, y'))$-measurable.

Fix $\varepsilon > 0$. Since $\Pi_{\mathcal{V}_x}\widetilde{W}_{d,x} \in \mathcal{V}_x = \overline{\{\Delta h\}}$, for each x there exists measurable $h_x^\varepsilon : \mathcal{Z}_d(x) \to \mathbb{R}$ such that

$$\|\Pi_{\mathcal{V}_x}\widetilde{W}_{d,x} - \Delta h_x^\varepsilon\|^2_{L^2(\nu_{d,x}^{\otimes 2})} \le 8\varepsilon.$$

Extend h_x^ε arbitrarily to $g_x^\varepsilon : \mathbb{R}^d \to \mathbb{R}$ and set $r_d^\varepsilon(x, y) = g_x^\varepsilon(\phi_d(x, y))$. Then $\Delta r_d^\varepsilon(x, y, y') = \Delta h_x^\varepsilon\big(\phi_d(x, y), \phi_d(x, y')\big)$, and by the Pythagorean identity for orthogonal projection,

$$\begin{aligned}
\mathbb{E}_{(y,y')\sim\mu_x^{\otimes 2}}\big[(W_d - \Delta r_d^\varepsilon)^2\big] &= \|\widetilde{W}_{d,x} - \Delta h_x^\varepsilon\|^2_{L^2(\nu_{d,x}^{\otimes 2})} \\
&= \|U_x\|^2_{L^2(\nu_{d,x}^{\otimes 2})} + \|\Pi_{\mathcal{V}_x}\widetilde{W}_{d,x} - \Delta h_x^\varepsilon\|^2_{L^2(\nu_{d,x}^{\otimes 2})} \\
&\le \|U_x\|^2_{L^2(\nu_{d,x}^{\otimes 2})} + 8\varepsilon\,.
\end{aligned}$$

Set $v = \tilde{p}_{\phi_d}(x, y, y')$, $t_v = F^{-1}(v) = W_d(x, y, y')$ and $t' = \Delta r_d^\varepsilon(x, y, y')$. Let $f_v(t) = D_{\mathrm{KL}}(v\|F(t))$; then $f_v'(t) = F(t) - v$ vanishes at t_v and $f_v''(t) = F(t)(1 - F(t)) \le \frac{1}{4}$ for all $t \in \mathbb{R}$. Taylor's theorem with remainder at t_v gives, for some ξ between t' and t_v,

$$f_v(t') = f_v(t_v) + f_v'(t_v)(t' - t_v) + \frac{1}{2}f_v''(\xi)\,(t' - t_v)^2 \le \tfrac{1}{8}\,(t' - t_v)^2,$$

since $f_v(t_v) = 0$ and $f_v'(t_v) = 0$. Applying this pointwise and integrating,

$$\mathcal{E}_{\mathrm{agr}}(\phi_d, r_d^\varepsilon) = \mathbb{E}\big[D_{\mathrm{KL}}(\tilde{p}_{\phi_d}\|F(\Delta r_d^\varepsilon))\big] \le \tfrac{1}{8}\,\mathbb{E}_x\big[\|W_d - \Delta r_d^\varepsilon\|^2_{L^2(\nu_{d,x}^{\otimes 2})}\big] \le \tfrac{1}{8}\,\mathbb{E}_x\left[\|U_x\|^2_{L^2(\nu_{d,x}^{\otimes 2})}\right] + \varepsilon.$$

Adding the embedding and agreement bounds yields the stated inequality. □

C Extension to Direct Preference Optimization

Fix a finite vocabulary $\mathcal{V}$ and identify $\mathcal{Y}$ with the set of finite sequences over $\mathcal{V}$. A policy π is a conditional distribution over $\mathcal{Y}$ given $x \in \mathcal{X}$, with the autoregressive factorization

$$\pi(y \mid x) = \prod_{t=1}^{|y|} \pi(y_t \mid y_{<t}, x), \qquad y_{<t} = (y_1, \dots, y_{t-1}).$$

The reference policy π_{ref} is a fixed autoregressive model and $\lambda > 0$ is the KL-regularization coefficient. Define

$$\Delta_\pi(x, y, y') = \lambda \left(\log \frac{\pi(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} - \log \frac{\pi(y' \mid x)}{\pi_{\mathrm{ref}}(y' \mid x)} \right).$$

Then the DPO loss can be written as

$$\mathcal{L}_{\mathrm{DPO}}(\pi) = \mathbb{E}_{x\sim\nu}\, \mathbb{E}_{(y,y')\sim\mu_x^{\otimes 2}} \Big[-\bar{p}(x,y,y') \log F\big(\Delta_\pi(x,y,y')\big) - \big(1 - \bar{p}(x,y,y')\big) \log\big(1 - F\big(\Delta_\pi(x,y,y')\big)\big) \Big]. \tag{7}$$

Its implicit reward is

$$r_\pi(x, y) = \lambda \sum_{t=1}^{|y|} \log \frac{\pi(y_t \mid y_{<t}, x)}{\pi_{\mathrm{ref}}(y_t \mid y_{<t}, x)}. \tag{8}$$

For each position $t \geq 1$, let $\psi_t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ be a prefix-conditional feature extractor. Define

$$\Phi_\psi(x, y) = \big(\psi_1(x, y_{\leq 1}), \dots, \psi_{|y|}(x, y_{\leq |y|})\big) \in (\mathbb{R}^d)^{|y|},$$

and

$$\mathcal{R}_{\mathrm{seq}}(\psi) := \Big\{ r : \mathcal{X}\times\mathcal{Y} \to \mathbb{R} \,\Big|\, r(x,y) = \sum_{t=1}^{|y|} g_t\big(\Phi_\psi(x, y_{\leq t})\big),\ g_t : (\mathbb{R}^d)^t \to \mathbb{R} \text{ measurable for each } t \geq 1 \Big\}.$$

For $(y_0, y_0') \sim \mu_x^{\otimes 2}$, set

$$\tilde{p}_\psi(x, y, y') = \mathbb{E}\big[\bar{p}(x, y_0, y_0') \,\big|\, \Phi_\psi(x, y_0) = \Phi_\psi(x, y), \Phi_\psi(x, y_0') = \Phi_\psi(x, y')\big].$$

Define P_ψ^δ as P_d^δ in Definition 4.3, with $\tilde{p}_\psi$ in place of $\tilde{p}_{\phi_d}$, and denote the corresponding cycle event by $\mathcal{CC}_\psi^\delta$.

Assumption C.1 (Prefix-covering)**.** *For every $x \in \mathcal{X}$ and any pair of trajectories (y, y'), the set of positions at which they share the same prefix,*

$$\mathrm{pref}(y, y') = \big\{ t \in \mathcal{N} : y_{\leq t} = y'_{\leq t} \big\},$$

is determined by the pair of trajectory embeddings $(\Phi_\psi(x, y), \Phi_\psi(x, y'))$. For every fixed x and $y \in \mathcal{Y}$, the map $t \mapsto \Phi_\psi(x, y_{\leq t})$ is injective on $\{0, 1, \dots, |y|\}$. Equivalently, if $\Phi_\psi(x, y_1) = \Phi_\psi(x, y_2)$ then $\mathrm{pref}(y, y_1) = \mathrm{pref}(y, y_2)$ for every trajectory y, and two prefixes of the same trajectory at different depths never share an embedding.

Theorem C.1. *Let π satisfy $r_\pi \in \mathcal{R}_{\mathrm{seq}}(\psi)$ and suppose $\tilde{p}_\psi(x, y, y') \in (0, 1)$ almost surely. Define*

$$\alpha_\psi(x, y, y') = \tilde{p}_\psi(x, y, y')\big(1 - \tilde{p}_\psi(x, y, y')\big),$$

$$\alpha_\psi^{\min}(x, y_1, y_2, y_3) = \min\{\alpha_\psi(x, y_1, y_2), \alpha_\psi(x, y_2, y_3), \alpha_\psi(x, y_3, y_1)\},$$

and

$$\bar{P}_\psi^\delta = \mathbb{E}_{x\sim\nu} \mathbb{E}_{(y_1,y_2,y_3)\sim\mu_x^{\otimes 3}} \Big[\alpha_\psi^{\min}(x, y_1, y_2, y_3)\, \mathbf{1}_{\mathcal{CC}_\psi^\delta}(x, y_1, y_2, y_3) \Big].$$

Let $e_\psi = \mathbb{E}[(\bar{p} - \tilde{p}_\psi)^2]$. Then for every $\delta \in (0, \frac{1}{4})$,

$$\mathcal{L}_{\mathrm{DPO}}(\pi) - \mathcal{L}^* \geq 8 e_\psi + \frac{\ell_\delta^2}{2} \bar{P}_\psi^\delta.$$

Proof. Set $\mathcal{H} = \sigma\big(\Phi_\psi(x, Y), \Phi_\psi(x, Y')\big)$ for $(Y, Y') \sim \mu_x^{\otimes 2}$ and

$$\tilde{p}_\psi(x, y, y') = \mathbb{E}\big[\bar{p}(x, Y, Y') \mid \mathcal{H}\big], \qquad e_\psi = \mathbb{E}\big[(\bar{p} - \tilde{p}_\psi)^2\big].$$

With $W_\psi = F^{-1}(\tilde{p}_\psi)$ and $\alpha_\psi = \tilde{p}_\psi(1 - \tilde{p}_\psi)$, the same decomposition argument as in Proposition 4.1 gives

$$\mathcal{L}_{\mathrm{DPO}}(\pi) - \mathcal{L}^* = \mathcal{E}_{\mathrm{emb}}(\Phi_\psi) + \mathcal{E}_{\mathrm{agr}}(\Phi_\psi, r_\pi),$$
$$\mathcal{E}_{\mathrm{emb}}(\Phi_\psi) \geq 8 e_\psi.$$

For the agreement term, pointwise strong convexity of $t \mapsto D_{\mathrm{KL}}(p \| F(t))$ at $t_p = F^{-1}(p)$ with curvature $p(1-p)$ yields

$$D_{\mathrm{KL}}\big(\tilde{p}_\psi \| F(\Delta r_\pi)\big) \geq \frac{\alpha_\psi}{2} \big(W_\psi - \Delta r_\pi\big)^2.$$

Integrating and taking the infimum over $r_\pi \in \mathcal{R}_{\mathrm{seq}}(\psi)$,

$$\inf_{r_\pi} \mathcal{E}_{\mathrm{agr}}(\Phi_\psi, r_\pi) \geq \frac{1}{2} \mathbb{E}\left[\alpha_\psi \operatorname{dist}\big(W_\psi, \mathcal{V}_{\mathrm{seq}}(x)\big)^2\right].$$

Dropping the additivity restriction can only decrease distance, so

$$\operatorname{dist}\big(W_\psi, \mathcal{V}_{\mathrm{seq}}(x)\big) \geq \operatorname{dist}\big(W_\psi, \overline{\mathcal{V}_x}\big),$$

and applying the weighted cycle argument from Proposition 4.2 and Lemma 4.2 with $\tilde{p}_\psi$ in place of $\tilde{p}_{\phi_d}$ gives

$$\inf_{r_\pi} \mathcal{E}_{\mathrm{agr}}(\Phi_\psi, r_\pi) \geq \frac{\ell_\delta^2}{2} \bar{P}_\psi^\delta.$$

Combining with $\mathcal{E}_{\mathrm{emb}}(\Phi_\psi) \geq 8 e_\psi$ yields the claim. □

Lemma C.1. *If Assumption C.1 holds, then $\mathcal{V}_{\mathrm{seq}}(x) = \overline{\mathcal{V}_x}$ for all $x \in \mathcal{X}$, so the sequential-additivity constraint introduces no additional projection gap.*

Proof. It suffices to show that every trajectory reward $r(x, y) = h(\Phi_\psi(x, y))$ can be written as a depth-indexed sum of per-token contributions, i.e., that the resulting Δr lies in $\mathcal{V}_{\mathrm{seq}}(x)$. Combined with the inclusion $\mathcal{V}_{\mathrm{seq}}(x) \subseteq \overline{\mathcal{V}_x}$ noted earlier, this gives $\overline{\mathcal{V}_x} = \mathcal{V}_{\mathrm{seq}}(x)$.

For each $t \geq 0$ let

$$\mathcal{S}_x^{(t)} = \big\{\Phi_\psi(x, y_{\leq t}) : y \in \mathcal{Y},\ |y| \geq t\big\} \subseteq (\mathbb{R}^d)^t,$$

with $\mathcal{S}_x^{(0)} = \{\perp\}$ a singleton root. Assumption C.1 states that the map $t \mapsto \Phi_\psi(x, y_{\leq t})$ is injective on $\{0, 1, \ldots, |y|\}$ for every y and that prefix-coincidence is determined by the embeddings. In particular, any two trajectories sharing $\Phi_\psi(x, y_{\leq t}) = \Phi_\psi(x, y'_{\leq t'})$ must satisfy $t = t'$ and $y_{\leq t} = y'_{\leq t}$. Therefore we work on the disjoint union of depth layers $\bigsqcup_{t \geq 0} \mathcal{S}_x^{(t)}$, whose elements are depth-tagged states (t, s) with $s \in \mathcal{S}_x^{(t)}$. The parent map takes one such node as input

$$\pi : \bigsqcup_{t \geq 1} \mathcal{S}_x^{(t)} \to \bigsqcup_{t \geq 0} \mathcal{S}_x^{(t)}, \qquad \pi(t, s) = \big(t - 1, \Phi_\psi(x, y_{\leq t-1})\big),$$

where $s = \Phi_\psi(x, y_{\leq t})$.

Let $r(x, y) = h(\Phi_\psi(x, y))$ be any trajectory reward and write $h_t : \mathcal{S}_x^{(t)} \to \mathbb{R}$ for its restriction to depth-t states (extended by $h_0(\perp) = 0$ at the root). Define $g_t : \mathcal{S}_x^{(t)} \to \mathbb{R}$ for $t \geq 1$ by

$$g_t(s) = h_t(s) - h_{t-1}(\pi(t, s)).$$

Extend each g_t arbitrarily (e.g. by zero) to a measurable function on $(\mathbb{R}^d)^t$. Telescoping along any prefix path $(s_0, s_1, \ldots, s_T)$ with $s_t = \Phi_\psi(x, y_{\leq t})$ gives, for every trajectory y with $|y| = T$,

$$h(\Phi_\psi(x, y)) = h_T(s_T) - h_0(\perp) = \sum_{t=1}^{T} g_t\big(\Phi_\psi(x, y_{\leq t})\big).$$

Hence $r(x, y) = \sum_{t=1}^{|y|} g_t(\Phi_\psi(x, y_{\leq t}))$, which gives $r \in \mathcal{R}_{\mathrm{seq}}(\psi)$. Consequently $\Delta r \in \mathcal{V}_{\mathrm{seq}}(x)$, and since r was arbitrary, every trajectory difference lies in $\mathcal{V}_{\mathrm{seq}}(x)$. Taking L^2 closures yields $\overline{\mathcal{V}_x} \subseteq \mathcal{V}_{\mathrm{seq}}(x)$, and the reverse inclusion is by definition, so $\overline{\mathcal{V}_x} = \mathcal{V}_{\mathrm{seq}}(x)$. □

D Proofs for Section 5

Proposition D.1. *Let μ' be obtained from μ by the soft-update with reward r and temperature $\beta > 0$, and let ϕ'_d be an embedding for which r is $\sigma(\phi'_d)$-measurable, i.e. $r(x,y) = h(\phi'_d(x,y))$ for some measurable h. Set $Z_x = \mathbb{E}_{\mu_x}[e^{\beta r(x,\cdot)}]$, $\rho(x,y) = e^{\beta r(x,y)}/Z_x$, and $\Delta p(x,y,y') = \bar{p}(x,y,y') - \tilde{p}_{\phi'_d}(x,y,y')$. Define*

$$\Gamma(\mu, \phi'_d; r, \beta) \;=\; \mathbb{E}_{\mathcal{D}_\mu}\Big[(\Delta p)^2\big(\rho(x,y)\rho(x,y') - 1\big)\Big] \;=\; \mathrm{Cov}_{\mathcal{D}_\mu}\big((\Delta p)^2,\, \rho(x,y)\rho(x,y')\big).$$

Then

$$e_d(\mu', \phi'_d) \;=\; e_d(\mu, \phi'_d) + \Gamma(\mu, \phi'_d; r, \beta)\,.$$

Proof. Fix x, and let Y, Y' be i.i.d. with law μ_x on $\mathcal{Y}$. Set $G_x = \sigma\big(\phi'_d(x,Y),\, \phi'_d(x,Y')\big)$. The conditional density of Y under μ'_x given $\phi'_d(x,Y) = z$ relative to that under μ_x is

$$\frac{e^{\beta r(x,y)}}{\mathbb{E}_{\mu_x}\big[e^{\beta r(x,Y)} \mid \phi'_d(x,Y) = z\big]} = \frac{e^{\beta h(z)}}{e^{\beta h(z)}} = 1,$$

where we used that $r(x,y) = h(\phi'_d(x,y)) = h(z)$ is constant on the level set $\{\phi'_d(x,\cdot) = z\}$. Hence the conditional law of Y given $\phi'_d(x,Y)$ is the same under μ_x and μ'_x; by independence of (Y, Y') under both product measures, the same holds for the joint conditional given G_x. Consequently

$$\mathbb{E}_{\mu'}[\bar{p} \mid \mathcal{G}_x] = \mathbb{E}_{\mu}[\bar{p} \mid \mathcal{G}_x] = \tilde{p}_{\phi'_d},$$

so Δp is well-defined as the same residual under either measure. Using the change of measure $\frac{d\mathcal{D}_{\mu'}}{d\mathcal{D}_\mu}(x,y,y') = \rho(x,y)\rho(x,y')$,

$$\begin{aligned} e_d(\mu', \phi'_d) &= \mathbb{E}_{\mathcal{D}_{\mu'}}[(\Delta p)^2] = \mathbb{E}_{\mathcal{D}_\mu}\big[(\Delta p)^2\, \rho(x,y)\rho(x,y')\big] \\ &= \mathbb{E}_{\mathcal{D}_\mu}\big[(\Delta p)^2\big] + \mathbb{E}_{\mathcal{D}_\mu}\big[(\Delta p)^2(\rho(x,y)\rho(x,y') - 1)\big] \\ &= e_d(\mu, \phi'_d) + \Gamma(\mu, \phi'_d; r, \beta). \end{aligned}$$

The covariance form follows from $\mathbb{E}_{\mathcal{D}_\mu}[\rho(x,y)\rho(x,y')] = \mathbb{E}_{x\sim\nu}\big[(\mathbb{E}_{\mu_x}\rho)^2\big] = 1$, since $\mathbb{E}_{\mu_x}\rho = Z_x^{-1}\mathbb{E}_{\mu_x}[e^{\beta r}] = 1$. □

The change in embedding error decomposes exactly as a covariance under $\mathcal{D}_\mu$ between the squared embedding residual $(\Delta p)^2$ and the relative likelihood $\rho(x,y)\rho(x,y')$ that the soft-update assigns to the pair. Embedding error grows under retraining if and only if these two quantities are non-negatively correlated. The next proposition gives sufficient conditions for that to happen.

Proposition D.2. *Under the assumptions of Proposition D.1, $\Gamma(\mu, \phi'_d; r, \beta) \geq 0$ holds in either of the following two settings:*

1. $\mathrm{Cov}_{\mathcal{D}_\mu}\big((\bar{p} - \tilde{p}_{\phi'_d})^2,\, \rho(x,y)\rho(x,y')\big) \geq 0$.
2. *For ν-almost every x, whenever one pair (y,y') has larger unresolved embedding error $(\bar{p} - \tilde{p}_{\phi'_d})^2$ than another pair, it also has no smaller total reward score $r(x,y) + r(x,y')$, after grouping pairs by the same embedding cells $(\phi'_d(x,y), \phi'_d(x,y'))$.*

Proof. Setting (1) is the covariance form of Γ in Proposition D.1.

For (2), the function $S(x,y,y') = \rho(x,y)\rho(x,y') = Z_x^{-2}\, e^{\beta(r(x,y)+r(x,y'))}$ is, conditional on x, a non-decreasing function of $r(x,y) + r(x,y')$ since $\beta > 0$. Co-monotonicity with $(\Delta p)^2$ and the Chebyshev sum inequality applied to the pushforward of $\mu_x^{\otimes 2}$ on the common ordering give

$$\mathbb{E}_{\mu_x^{\otimes 2}}\big[(\Delta p)^2\, S\big] \geq \mathbb{E}_{\mu_x^{\otimes 2}}[(\Delta p)^2] \cdot \mathbb{E}_{\mu_x^{\otimes 2}}[S] = \mathbb{E}_{\mu_x^{\otimes 2}}[(\Delta p)^2],$$

using $\mathbb{E}_{\mu_x}[\rho] = 1$. Integrating over $x \sim \nu$ yields $\Gamma \geq 0$. □

Proposition 5.1. *With $\phi = \phi_d^{(t)}$ and r_t measurable, $e_d(\mu^{(t+1)}, \phi) - e_d(\mu^{(t)}, \phi) = \Gamma_t$, $P_\infty^{2\delta}(\mu^{(t+1)}) - P_\infty^{2\delta}(\mu^{(t)}) = \Theta_t$.*

Proof. The Γ term is the statement of Proposition D.1. For Θ, fix x and use the derivative

$$\frac{d(\mu_x^{(t+1)})^{\otimes 3}}{d(\mu_x^{(t)})^{\otimes 3}}(y_1, y_2, y_3) = \rho_t(x, y_1)\rho_t(x, y_2)\rho_t(x, y_3),$$

which integrates to 1 since $\mathbb{E}_{\mu_x^{(t)}}[\rho_t(x, \cdot)] = 1$ and the three coordinates are independent. Because $\mathcal{CC}_\infty^{2\delta}$ is defined via $\bar{p}$ and does not depend on $\mu^{(t)}$,

$$\begin{aligned} P_\infty^{2\delta,(t+1)} - P_\infty^{2\delta,(t)} &= \mathbb{E}_\nu\Big[\mathbb{E}_{\mu_x^{(t),\otimes 3}}\left[(\rho_t(x, y_1)\rho_t(x, y_2)\rho_t(x, y_3) - 1)\,\mathbf{1}_{\mathcal{CC}_\infty^{2\delta}}\right]\Big] \\ &= \mathbb{E}_\nu\Big[\mathrm{Cov}_{\mu_x^{(t),\otimes 3}}\left(\mathbf{1}_{\mathcal{CC}_\infty^{2\delta}}, \rho_t(x, y_1)\rho_t(x, y_2)\rho_t(x, y_3)\right)\Big], \end{aligned}$$

where the second equality uses $\mathbb{E}_{\mu_x^{(t),\otimes 3}}[\rho_t(x, y_1)\rho_t(x, y_2)\rho_t(x, y_3)] = 1$. □

Proposition 5.2. *At any fixed point (μ^*, ϕ^*, r^*) of the joint update with ϕ^* optimal on μ^* and $r^* \in \mathcal{R}_d(\phi^*)$ loss-optimal on μ^*, both $\Gamma(\mu^*, \phi^*; r^*, \beta) = 0$ and $\Theta(\mu^*; r^*, \beta) = 0$.*

Proof. Invariance of μ^* under the soft-update means $d\mu_x^{*,\prime}/d\mu_x^* = 1$, i.e. $\rho^*(x, y) = 1$ for ν-a.e. x and μ_x^*-a.e. y, in which case both products $\rho^*(x, y)\rho^*(x, y')$ and $\rho^*(x, y_1)\rho^*(x, y_2)\rho^*(x, y_3)$ are constant (equal to 1) on the supports of the corresponding product measures. Constants have zero covariance with any function, so $\Gamma(\mu^*, \phi^*; r^*, \beta) = 0$ and $\Theta(\mu^*; r^*, \beta) = 0$. □

E More Experiment Details and Results

Fitting the Reward For each d, items are partitioned into groups by ϕ_d. We fit one scalar score per group by full-batch gradient descent on the pairwise logistic loss to the coarsened matrix $\tilde{P}_{\phi_d}$, with learning rate 5×10^{-2} and a decaying step-size schedule.

DPO We define implicit rewards $r_i = \lambda\big(\log \pi_\theta(y_i) - \log \pi_{\text{ref}}(y_i)\big)$, average r within each ϕ_d-group, and set $q^{\pi}_{ij} = \sigma(r_i - r_j)$. The policy π_θ and reference π_{ref} are fixed in this evaluation.

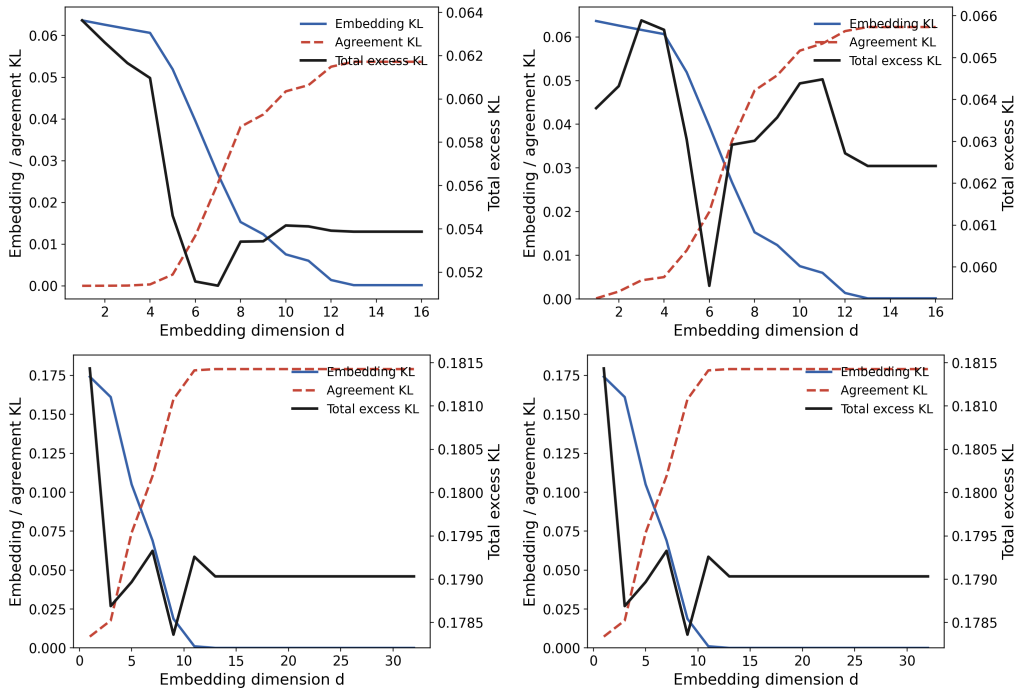


Figure 3: Top row: results on Jester, Bottom row: results on Sushi. Left panel: results with RLHF, right panel: results with DPO.

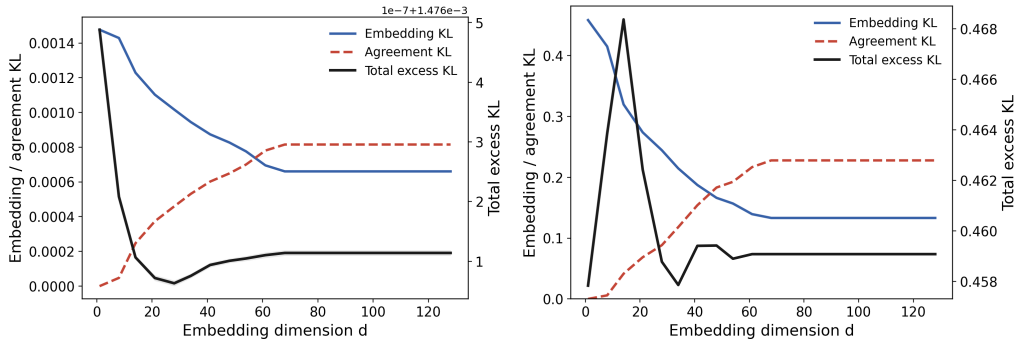


Figure 4: Excess loss decomposition on the MT-Bench preference dataset with `all-MiniLM-L6-v2` embedding. Left: RLHF. Right: DPO.

Licenses and terms of use. Pretrained embeddings are `sentence-transformers/all-MiniLM-L6-v2` and `sentence-transformers/all-mpnet-base-v2` from the Hugging Face Hub (`https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`, `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`). Each model card lists Apache License, version 2.0, and we use the

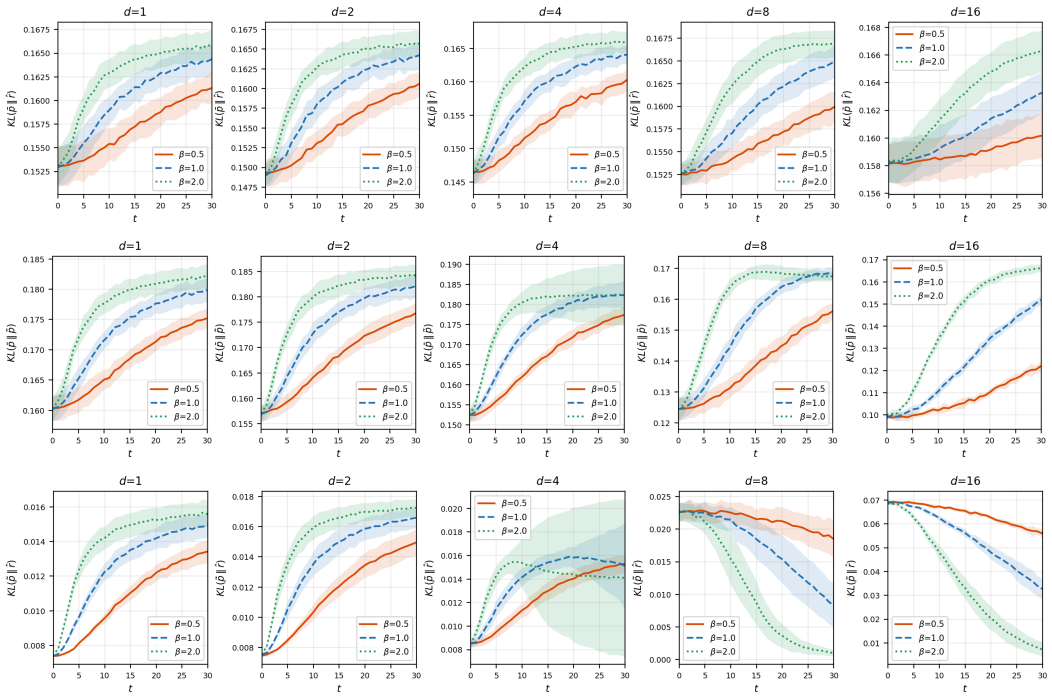


Figure 5: Joint embedding–reward updates on synthetic data. Top: total excess loss over update rounds. Middle: embedding error $\mathcal{E}_{\text{emb}}(\phi_d)$ over rounds. Bottom: agreement cost $\mathcal{E}_{\text{agr}}(\phi_d, r)$ over rounds.

checkpoints accordingly. MT-Bench pairwise judgments are taken from the Hugging Face dataset `lmsys/mt_bench_human_judgments` (`https://huggingface.co/datasets/lmsys/mt_bench_human_judgments`), which lists CC BY 4.0 on its dataset card. The Jester and Sushi data are used under the terms of the original releases cited above.

Figure 3 reports results on Jester and Sushi for both RLHF and DPO with pretrained embedding from `mpnet` [27].

Figure 4 shows the results on MT-Bench. Again, the embedding loss decreases and the agreement cost increases with embedding dimension, and their sum may be minimized at an intermediate dimension.

We also evaluate joint embedding-reward updates using the same synthetic setup. We initialize from a constant embedding and run $T = 10$ rounds of alternating reward fitting and distribution tilting. At each round we measure the embedding error $\mathcal{E}_{\text{emb}}(\phi_d^{(t)})$, the agreement cost $\mathcal{E}_{\text{agr}}(\phi_d^{(t)}, r_t)$, and their sum. The Top panel of Figure 5 shows that total excess loss increases over round. The middle panel shows that the embedding error grows over rounds. The bottom panel shows agreement cost can either increase or decrease, with different dimension.

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 and introduction match the paper with the main contibution being a lower bound to demonstrate tradeoff of embedding in RLHF

   Guidelines:

   - The answer [N/A] 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 [N/A] 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 pooled preference model assumes the random-utility/Bradley Terry model.

   Guidelines:

   - The answer [N/A] 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: Proofs to every theoretical claims made in the paper are provided in the appendix.

Guidelines:

- The answer [N/A] 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: Our experiments are mainly for the purpose of validating our theoretical claims. Therefore, the experiments are relatively simple and easy to replicate. We have included the main implementations details in the paper and provided the code in the supplemental materials.

Guidelines:

- The answer [N/A] 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 used publicly available dataset and synthetically generated data. The code used to generate the data and to replicate the main experimental results are provided in the supplemental material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The full implemention details are given in the supplemntal material.

   Guidelines:

   - The answer [N/A] 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: We reported the error bars for the synthetic experiments.

   Guidelines:

   - The answer [N/A] 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 described in the experiment section thatall experiments ran on one NVIDIA L40 GPU with 48 GB GDDR6 memory. Workloads are lightweight relative to this hardware, so wall-clock is dominated by dataset-dependent encoder passes and was not separately logged.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have reviewed the code of ethics and confirm that the paper conform to it.

   Guidelines:

   - The answer [N/A] 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 have included a broader impact paragraph following the conclusion to discuss the societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release new models.

    Guidelines:

    - The answer [N/A] 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: We have cited the paper that provided the Jester, Sushi and MT-Bench dataset. We explicitly mentioned the license for the pretrained checkpoints and dataset we have used in the appendix.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We have mainly used LLM for code assistance and writing assistance.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.