Variance-aware Reward Modeling with Anchor Guidance

Anonymous Author(s)
Anonymous Institution
Anonymous Address
anonymous@email.com

Abstract

Standard Bradley–Terry (BT) reward models are limited when human preferences are pluralistic. Although soft preference labels preserve disagreement information, BT can only express it by shrinking reward margins. Gaussian reward models provide an alternative by jointly predicting a reward mean and a reward variance, but suffer from a fundamental non-identifiability from pairwise preferences alone. We propose Anchor Guided Variance-aware Reward Modeling, a framework that resolves this non-identifiability by augmenting preference data with two coarse response-level anchor labels. Building on this, we prove that two anchors are sufficient for identification, develop a joint training objective and establish a non-asymptotic convergence rate for both the estimated reward mean and variance functions. Across simulation studies and four real-world diverging-preference datasets, our method consistently improves reward modeling performance and downstream RLHF, including PPO training and best-of-N selection.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) are widely employed to align LLMs with human preferences and avoid generating harmful content [Bai et al., 2022, Ouyang et al., 2022, Lambert et al., 2024, Hu et al., 2024, Ji et al., 2026, Liu et al., 2026]. A typical RLHF pipeline first trains a reward model on human preference data, and then uses this reward model to guide policy optimization via algorithms such as PPO [Schulman et al., 2017] or GRPO [Shao et al., 2024]. The quality of the reward model is therefore central to the success of the entire pipeline.

As LLMs are used by increasingly diverse populations, alignment should account for pluralistic human preferences [Gordon et al., 2022, Sorensen et al., 2024a,b]. A line of recent work shows that annotators with different backgrounds frequently provide diverging preferences on the same prompt–response pair [Baumler et al., 2023, Mishra, 2023, Siththaranjan et al., 2024, Zhang et al., 2025, Halpern et al., 2026]. Recent multi-annotator preference datasets such as `MultiPref` [Miranda et al., 2025], `HelpSteer2` [Wang et al., 2025a], `HelpSteer3` [Wang et al., 2025b], and `PersonalLLM` [Zollo et al., 2025] further confirm that a substantial fraction of pairwise comparisons receive split votes from human annotators or LLM judges. Standard reward modeling approaches simply aggregate these split votes into a single majority label [Casper et al., 2023, Wang et al., 2024]. This aggregation design fails to capture the underlying diversity of preferences and can further lead to preference collapse [Zhang et al., 2025, Halpern et al., 2026] and reward hacking [Gao et al., 2023, Fu et al., 2025, Sun et al., 2025].

A natural alternative is to adopt soft preference labels that reflect the empirical proportion of annotators favoring one response, thereby preserving annotator disagreement information and providing a better training signal for learning calibrated preference probabilities [Liu et al., 2025b, Halpern et al., 2026]. Although soft preference labels can be directly used in the BT [Bradley and Terry, 1952] loss, the

Submitted to STAI-X 2026 (Full Paper Track). Do not distribute.

standard BT captures only the mean preference probability. As a result, annotator disagreement is absorbed by shrinking the reward margin between two responses, rather than being explicitly represented as uncertainty.

A principled alternative is to model rewards as distributions rather than scalars [Siththaranjan et al., 2024, Lou et al., 2024, Sun et al., 2025, Zhang et al., 2025, Liu et al., 2025a, Frick, 2025, Chen et al., 2026]. A representative instance is the Gaussian reward model, which outputs a reward mean and a reward variance [Siththaranjan et al., 2024, Zhang et al., 2025, Frick, 2025]. It corresponds to the classical Thurstonian formulation [Thurstone, 1927] and generalizes BT from a scalar reward with fixed noise to a heteroscedastic reward distribution. However, this additional modeling flexibility introduces a fundamental identifiability problem [Liu et al., 2025a, Frick, 2025]. Different combinations of reward mean and variance functions can induce the same preference probabilities, making the learned uncertainty difficult to compare and interpret.

To address this issue, we propose Anchor-guided Variance-aware Reward Modeling, which resolves the non-identifiability of Gaussian reward models via an auxiliary anchor dataset: each response receives two coarse binary anchor labels indicating whether its latent utility exceeds two distinct thresholds. We develop a joint training objective that combines a soft-label preference loss and a multinomial cross-entropy loss induced by the two anchors. To verify that performance is robust to the label source, we construct coarse anchor labels from two distinct sources: external reward models and LLM judges. Based on this framework, we prove that the two coarse anchors provide the constraints to uniquely identify both the reward mean and variance functions, and we further establish a non-asymptotic convergence rate for both functions. Empirically, on simulation data and four real-world diverging-preference benchmarks (`MultiPref`, `HelpSteer2`, `HelpSteer3`, `PersonalLLM`), our method consistently outperforms other baselines in both reward modeling evaluation and downstream RLHF performance (PPO training and best-of-N selection) .

To summarize, our contributions include: (1) a formal characterization of the non-identifiability of Gaussian reward models and a two-anchor identification result; (2) a practical anchor-guided variance-aware reward modeling framework with finite-sample convergence guarantees for both the reward mean and variance function; (3) a comprehensive empirical demonstration of its effectiveness on diverging-preference benchmarks and downstream RLHF evaluation.

Related work. We focus on three lines of literature most related to our work. ***(i) Pluralistic alignment.*** Existing methods for pluralistic preferences can be broadly divided into two categories. The first leverages auxiliary information about annotators, such as demographic attributes, group memberships, or user-specific information, to learn group-specific or personalized reward functions [Chakraborty et al., 2024, Poddar et al., 2024, Chen et al., 2024, 2025, Kim and Kim, 2026]. The second represents rewards as distributions rather than scalars, often referred to as distributional or uncertainty-aware reward model. Since this line of work is closely related to ours, we discuss it in detail below. ***(ii) Distributional reward model.*** Existing distributional framework can be divided into three categories: mean-variance reward models [Lou et al., 2024, Siththaranjan et al., 2024, Sun et al., 2025, Zhang et al., 2025, Liu et al., 2025a, Frick, 2025], categorical reward models [Siththaranjan et al., 2024, Zhang et al., 2025, Chen et al., 2026], and ensemble-based reward models [Coste et al., 2023, Lou et al., 2024, Halpern et al., 2026]. Among them, mean-variance reward models are the most related to our work. Frick [2025] only briefly discusses the unidentifiability of Gaussian reward models and their empirical results show comparable performance to the BT model. In contrast, our work resolves this non-identifiability via two coarse anchor labels and obtains clear empirical improvements over other baselines. ***(iii) Auxiliary supervision.*** A growing line of work uses auxiliary fine-grained supervision for BT reward modeling, such as teacher reward values [Zhang et al., 2026a] or attribute-level scores [Zhang et al., 2026b]. In contrast, we employ external reward models or LLM as a judge solely as a low-cost tool to overcome the lack of reponse-level coarse quality labels (e.g., good / normal / bad) in existing preference data. The closest prior work is Chen et al. [2026], which uses such coarse quality labels to train a categorical reward model. From a statistical perspective, we instead use these coarse labels as identification constraints for a Gaussian reward model, and prove both identifiability and a non-asymptotic convergence rate. We further empirically evaluate the robustness of our method across different coarse anchor label sources, and analyze anchor data efficiency by varying proportions of available anchor data.

2 Preliminaries

Let $\mathcal{X}$ be the prompt space and $\mathcal{Y}$ be the response space. We consider a pairwise preference dataset $\mathcal{D}_{\text{pref}} = \{(x_i, y_{i,1}, y_{i,2}, C_i)\}_{i=1}^{n_{\text{pref}}}$ consisting of n_{pref} i.i.d. tuples, where $x \in \mathcal{X}$ is a prompt, $y_{i,1}, y_{i,2} \in \mathcal{Y}$ are two candidate responses and $C_i \in \{0, 1\}$ is the observed binary preference label, with $C_i = 1$ indicating that $y_{i,1}$ is preferred over $y_{i,2}$.

BT Reward Model. Most existing reward-based RLHF algorithms adopt the BT assumption [Bradley and Terry, 1952]. We assume the latent utility of a response y given a prompt x is $U(x, y) = r_\phi(x, y) + \epsilon$, where $r_\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a parametrized scalar reward function and ϵ is i.i.d. standard Gumbel noise. The probability that $y_{i,1}$ is preferred to $y_{i,2}$ is

$$\mathbb{P}(C_i = 1 \mid x_i, y_{i,1}, y_{i,2}) = \mathbb{P}\big(U(x_i, y_{i,1}) > U(x_i, y_{i,2})\big) = \sigma\big(r_\phi(x_i, y_{i,1}) - r_\phi(x_i, y_{i,2})\big) \tag{1}$$

where $\sigma(t) = 1/(1 + e^{-t})$ is the sigmoid function. In practice, we optimize the reward function by minimizing the negative log-likelihood loss over the preference dataset: $\mathcal{L}_{\text{BT}}(\mathcal{D}_{\text{pref}} \mid \phi) = -\frac{1}{n_{\text{pref}}} \sum_{i=1}^{n_{\text{pref}}} \big[C_i \log \sigma(z_i) + (1 - C_i) \log\big(1 - \sigma(z_i)\big)\big]$, where $z_i = r_\phi(x_i, y_{i,1}) - r_\phi(x_i, y_{i,2})$ is the reward margin.

Gaussian Reward Model. Following the Thurstonian model assumption [Thurstone, 1927], we consider a Gaussian reward model that explicitly captures heteroscedastic uncertainty in human preferences. Specifically, for each (x, y), we assume the latent utility follows $U(x, y) \sim \mathcal{N}\big(r_\phi(x, y),\, s_\phi^2(x, y)\big)$, where $r_\phi(x, y)$ is the reward mean function and $s_\phi(x, y) > 0$ is the reward standard deviation function. For a comparison $(x_i, y_{i,1}, y_{i,2})$, the preference probability is

$$\mathbb{P}(C_i = 1 \mid x_i, y_{i,1}, y_{i,2}) = \mathbb{P}\big(U(x_i, y_{i,1}) > U(x_i, y_{i,2})\big) = \Phi\left(\frac{r_\phi(x_i, y_{i,1}) - r_\phi(x_i, y_{i,2})}{\sqrt{s_\phi^2(x_i, y_{i,1}) + s_\phi^2(x_i, y_{i,2})}}\right), \tag{2}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Similar to the BT model, we optimize both functions by minimizing the negative log-likelihood over the preference dataset:

$$\mathcal{L}_{\text{G}}(\mathcal{D}_{\text{pref}} \mid \phi) = -\frac{1}{n_{\text{pref}}} \sum_{i=1}^{n_{\text{pref}}} \big[C_i \log \Phi(z_i^{\text{G}}) + (1 - C_i) \log(1 - \Phi(z_i^{\text{G}}))\big], \tag{3}$$

where $z_i^{\text{G}} = \big(r_\phi(x_i, y_{i,1}) - r_\phi(x_i, y_{i,2})\big) / \sqrt{s_\phi^2(x_i, y_{i,1}) + s_\phi^2(x_i, y_{i,2})}$ is the variance-normalized reward margin. The identifiability issue of this formulation is analyzed in Section 3.

3 Identifiability

We first demonstrate that, given only pairwise preference dataset $\mathcal{D}_{\text{pref}}$, the reward mean and standard deviation functions (r_ϕ, s_ϕ) cannot be uniquely identified.

Proposition 1 (Non-identifiability of the Gaussian reward model)**.** *The Gaussian preference probability defined in* (2) *is invariant under the following two families of function transformations:* (i) Reward translation. *For any constant $b \in \mathbb{R}$, the preference probability is invariant under the mapping $(r_\phi, s_\phi) \mapsto (r_\phi + b,\ s_\phi)$.* (ii) Joint positive scaling. *For any scalar $c > 0$, the preference probability is invariant under the mapping $(r_\phi, s_\phi) \mapsto (c\, r_\phi,\ c\, s_\phi)$.*

Thus, different choices of (r_ϕ, s_ϕ) induce the same distribution over observed preferences. In the BT formulation, the noise distribution is fixed, so r_ϕ is identifiable up to an additive constant. The additional modeling flexibility provided by the learnable variance $s_\phi^2(x, y)$ in the Gaussian model introduces a new identifiability issue. Therefore, without additional constraints or regularization, neither the absolute level of r_ϕ nor the joint scale of (r_ϕ, s_ϕ) is identifiable only from $\mathcal{D}_{\text{pref}}$.

To resolve the non-identifiability under preference-only data, we impose additional response-level anchoring constraints. Specifically, suppose we have access to an additional anchor dataset $\mathcal{D}_{\text{anc}} = \{(x_j, y_j, A_{j,1}, A_{j,2})\}_{j=1}^{n_{\text{anc}}}$ consisting of n_{anc} i.i.d. tuples, where each prompt–response pair (x_j, y_j) is associated with two binary labels indicating whether the response quality exceeds two fixed anchor thresholds $\tau_1 < \tau_2$. Formally, for $k \in \{1, 2\}$, the anchor label is defined as $A_{j,k} = \mathbf{1}\{U(x_j, y_j) \geq$

$\tau_k\}$. Under the Gaussian reward model, the probability that the latent utility exceeds τ_k is

$$q_k(x,y) = \mathbb{P}\big(U(x,y) \geq \tau_k \mid x, y\big) = \Phi\left(\frac{r_\phi(x,y) - \tau_k}{s_\phi(x,y)}\right), \qquad k \in \{1,2\}. \tag{4}$$

We now show that two distinct anchor thresholds are sufficient to identify the function pair (r_ϕ, s_ϕ).

Proposition 2 (Identifiability under two anchors)**.** *Let $\tau_1 < \tau_2$ be two distinct anchor thresholds. Under the Gaussian reward model, the two anchor probability functions $q_1(x,y)$ and $q_2(x,y)$ uniquely identify the reward mean function $r_\phi(x,y)$ and the standard deviation function $s_\phi(x,y)$ on the support of the prompt–response distribution.*

The intuition behind Proposition 2 is simple: for each prompt–response pair (x, y), the two anchor probabilities provide two constraints for the two unknowns $r_\phi(x,y)$ and $s_\phi(x,y)$. Specifically, we obtain the following equations $r_\phi(x,y) - \tau_1 = s_\phi(x,y)\, z_{\mathrm{anc},1}(x,y)$ and $r_\phi(x,y) - \tau_2 = s_\phi(x,y)\, z_{\mathrm{anc},2}(x,y)$, where $z_{\mathrm{anc},k}(x,y) = \Phi^{-1}\big(q_k(x,y)\big)$ for $k \in \{1,2\}$. Solving for $s_\phi(x,y)$ and $r_\phi(x,y)$ yields the following closed-form solutions: $s_\phi(x,y) = \frac{\tau_2 - \tau_1}{z_{\mathrm{anc},1}(x,y) - z_{\mathrm{anc},2}(x,y)}$ and $r_\phi(x,y) = \tau_1 + s_\phi(x,y)\, z_{\mathrm{anc},1}(x,y) = \tau_2 + s_\phi(x,y)\, z_{\mathrm{anc},2}(x,y)$. The solution is unique as long as $z_{\mathrm{anc},1}(x,y) \neq z_{\mathrm{anc},2}(x,y)$. Because $\tau_1 \neq \tau_2$ and $s_\phi(x,y) > 0$, the condition $q_1(x,y) \neq q_2(x,y)$ is naturally satisfied. Since the same argument applies to every prompt–response pair on the support, the two anchor probability functions identify both r_ϕ and s_ϕ. Therefore, two distinct anchors are sufficient. Additionally, we design a simulation study to empirically confirm that, unlike preference-only baselines, the two-anchor model recovers both the reward mean and variance (see Figure 6 in Appendix C).

4 Methodology

We introduce our proposed Anchor-guided Variance-aware Reward Modeling framework.

Model architecture. We parameterize the Gaussian reward model using a shared backbone followed by two linear heads. The shared backbone is an instruction-tuned LLM with its language model head removed. Given a prompt–response pair (x, y), the shared backbone with parameter $\theta_{\mathrm{backbone}}$ produces a hidden representation $\psi(x,y) \in \mathbb{R}^d$. Two separate linear heads then map this representation to the reward mean and variance: $r_\phi(x,y) = \mathbf{w}_r^\top \psi(x,y)$ and $s_\phi^2(x,y) = \mathrm{softplus}\big(\mathbf{w}_v^\top \psi(x,y)\big)$, where $\phi = \{\theta_{\mathrm{backbone}}, \mathbf{w}_r, \mathbf{w}_v\}$ denotes the full parameter set. We adopt the $\mathrm{softplus}$ activation to ensure $s_\phi^2(x,y) > 0$, a common choice for enforcing positivity in neural heteroscedastic modeling [Skafte et al., 2019, Seitzer et al., 2022, Stirn et al., 2023].

Anchor label generation. To resolve the identifiability issue discussed in Section 3, we require additional coarse anchor labels at the response level. Such labels can be viewed as coarse quality assessments indicating whether a response is above a low or high quality threshold (e.g., good / normal / bad). However, existing preference datasets mainly focus on pairwise comparison labels and often lack response-level coarse quality labels. We therefore design two strategies to automatically construct coarse anchor labels. (1) ***External reward model***: For a prompt–response pair (x_i, y_i), we compute a scalar score $\tilde{r}(x_i, y_i)$ from an external reward model. (2) ***Human/LLM-as-a-judge.*** Each response can be evaluated by strong language models or human annotators along multiple quality dimensions. These dimension-wise scores are then aggregated into a scalar reward score $\tilde{r}(x_i, y_i)$ via a weighted sum. For both strategies, we then compute the normalized scalar reward $\tilde{r}'(x_i, y_i)$ by subtracting the training set mean. The two thresholds $\tau_1 < \tau_2$ are then set to the q-th and $(1-q)$-th quantiles of the centered score distribution, yielding the anchor labels $A_{i,1} = \mathbf{1}\{\tilde{r}'(x_i, y_i) \geq \tau_1\}$ and $A_{i,2} = \mathbf{1}\{\tilde{r}'(x_i, y_i) \geq \tau_2\}$. Details are provided in Appendix D.3.

Joint training objective. Human preferences often diverge across annotators [Sorensen et al., 2024b, Zhang et al., 2025, Halpern et al., 2026]. Following recent work that treats annotator disagreement as informative soft labels [Halpern et al., 2026], we train the preference model with soft labels rather than collapsing multiple annotations into a single hard label. Specifically, if the i-th comparison $(x_i, y_{i,1}, y_{i,2})$ is annotated by m_i annotators, let $C_{i,a} \in \{0,1\}$ denote the binary preference label provided by the a-th annotator, where $C_{i,a} = 1$ indicates that $y_{i,1}$ is preferred over $y_{i,2}$. We define the soft preference label as $\bar{C}_i = \frac{1}{m_i} \sum_{a=1}^{m_i} C_{i,a}$, where $\bar{C}_i$ represents the empirical fraction of annotators who prefer $y_{i,1}$ over $y_{i,2}$. Given the aggregated soft preference dataset $\{(x_i, y_{i,1}, y_{i,2}, \bar{C}_i)\}_{i=1}^{n_{\mathrm{pref}}}$, the

negative log-likelihood is

$$\mathcal{L}_{\text{pref}}(\phi) = -\frac{1}{n_{\text{pref}}} \sum_{i=1}^{n_{\text{pref}}} \Big[\bar{C}_i \log \Phi(z_i^{\text{G}}) + (1 - \bar{C}_i) \log\big(1 - \Phi(z_i^{\text{G}})\big) \Big], \tag{5}$$

where z_i^{G} is the variance-normalized reward margin defined in Eq. (3) Given an anchor dataset $\mathcal{D}_{\text{anc}} = \{(x_j, y_j, A_{j,1}, A_{j,2})\}_{j=1}^{n_{\text{anc}}}$, where $A_{j,k} = \mathbf{1}\{U(x_j, y_j) \geq \tau_k\}$ and $\tau_1 < \tau_2$, the two anchor labels are derived from the same latent utility $U(x_j, y_j)$ and thus satisfy the ordinal constraint $A_{j,2} \leq A_{j,1}$. Specifically, the valid label configurations are $(A_{j,1}, A_{j,2}) = (0,0)$, $(1,0)$, and $(1,1)$, corresponding to $U(x_j, y_j) < \tau_1$, $\tau_1 \leq U(x_j, y_j) < \tau_2$, and $U(x_j, y_j) \geq \tau_2$. The probabilities of three threshold-induced classes are $\pi_{j,0} = \mathbb{P}(A_{j,1} = 0, A_{j,2} = 0) = 1 - q_1(x_j, y_j)$, $\pi_{j,1} = \mathbb{P}(A_{j,1} = 1, A_{j,2} = 0) = q_1(x_j, y_j) - q_2(x_j, y_j)$ and $\pi_{j,2} = \mathbb{P}(A_{j,1} = 1, A_{j,2} = 1) = q_2(x_j, y_j)$, where $q_k(x, y)$ is the conditional anchor probability defined in Eq. (4). Thus, the anchor negative log-likelihood is the multinomial cross-entropy over the three threshold-induced classes:

$$\mathcal{L}_{\text{anc}}(\phi) = -\frac{1}{n_{\text{anc}}} \sum_{j=1}^{n_{\text{anc}}} \Big[(1 - A_{j,1}) \log \pi_{j,0} + (A_{j,1} - A_{j,2}) \log \pi_{j,1} + A_{j,2} \log \pi_{j,2} \Big]. \tag{6}$$

The joint training objective is to minimize the combined negative log-likelihood: $\mathcal{L}(\phi) = \mathcal{L}_{\text{pref}}(\phi) + \lambda \mathcal{L}_{\text{anc}}(\phi)$, where $\lambda > 0$ is a weighting hyperparameter to control the strength of anchor supervision.

5 Theory

For notational simplicity, we define $u = \psi(x, y) \in \mathcal{U} \subset \mathbb{R}^d$, where $\mathcal{U}$ is a compact domain. We then specify the latent reward mean and log-standard-deviation functions as $r(u)$ and $g(u) = \log s(u)$ in our theoretical analysis. We adopt the log-standard-deviation parameterization instead of the softplus form employed in Section 4, since the log parameterization ensures that g lies in an unconstrained function space and is more convenient for sieve analysis. Let $\{\mathcal{R}_K\}_{K \geq 1}$ and $\{\mathcal{G}_K\}_{K \geq 1}$ be sequences of K-dimensional sieve spaces (e.g., spline spaces) defined on $\mathcal{U}$. For a fixed constant $B_\Theta > B$, we define the truncated sieve space $\Theta_K(B_\Theta) = \{(r, g) \in \mathcal{R}_K \times \mathcal{G}_K : \|r\|_\infty \leq B_\Theta,\ \|g\|_\infty \leq B_\Theta\}$. We consider the aggregated soft-label preference dataset $\{(x_i, y_{i,1}, y_{i,2}, \bar{C}_i)\}_{i=1}^{n_{\text{pref}}}$ defined in Section 4. Under the true functions (r^*, g^*), we assume $\mathbb{E}[\bar{C}_i \mid x_i, y_{i,1}, y_{i,2}] = \Phi\big(z_i^{\text{G}}(r^*, g^*)\big)$. The anchor label pair $(A_{j,1}, A_{j,2})$ follows the ordinal three-class multinomial distribution induced by the two thresholds $\tau_1 < \tau_2$. We estimate the true functions by minimizing the joint empirical loss

$$(\hat{r}, \hat{g}) \in \arg \min_{(r,g) \in \Theta_K(B_\Theta)} \Big\{ \mathcal{L}_{\text{pref}}(r, g) + \lambda \mathcal{L}_{\text{anc}}(r, g) \Big\},$$

where $\mathcal{L}_{\text{pref}}(r, g)$ is the empirical preference loss defined in Eq. (5), $\mathcal{L}_{\text{anc}}(r, g)$ is the empirical anchor loss defined in Eq. (6), and $\lambda > 0$ is a weighting parameter.

We next introduce some assumptions to analyze this estimator.

Assumption 1 (Data generation)**.** *Let $\mathbb{P}_\psi$ denote the marginal distribution of $\psi(x, y)$. For each anchor data (x, y), we have $\psi(x, y) \sim \mathbb{P}_\psi$. For each preference data (x, y_1, y_2), the two responses are conditionally i.i.d. given x, we have $\psi(x, y_1), \psi(x, y_2) \sim \mathbb{P}_\psi$.*

Assumption 2 (Structure and regularity)**.** *The true functions (r^*, g^*) are uniformly bounded by some $B > 0$ and lie in Hölder classes with smoothness indices $\alpha_r, \alpha_g > 0$.*

Assumption 3 (Sieve approximation and stability)**.** *The sieve spaces $\mathcal{R}_K$ and $\mathcal{G}_K$ approximate r^* and g^* at the standard nonparametric rates and the basis functions are uniformly stable.*

Assumption 4 (Sieve dimension)**.** *The sieve dimension K is allowed to grow with the sample sizes and satisfies $\frac{K}{\min\{n_{\text{pref}}, n_{\text{anc}}\}} \to 0$.*

More detailed versions of Assumption 1–3 are provided in Appendix A. Assumption 2 requires the feature representation to be uniformly bounded and non-degenerate, and the true functions to be smooth and uniformly bounded. These conditions are standard in both the RLHF literature [Zhu et al., 2023, Liu et al., 2025a, Zhang et al., 2026b] and nonparametric sieve estimation [Shen and Wong, 1994, Newey, 1997, Chen, 2007]. Assumption 3 ensures that the true functions can be well approximated by finite-dimensional sieve spaces, while Assumption 4 controls the growth of the sieve dimension relative to the sample sizes.

Table 1: Overall performance of different models with the `Llama-3.2-3B-Instruct` as backbone. Bold numbers represent the best performance. Underlined numbers represent the second best.

Method	RewardBench			PPE-P			Method	RewardBench			PPE-P		
	Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓		Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓
Dataset: MultiPref-8K							**Dataset: HelpSteer3-20K**						
BT	71.7%	0.209	0.604	59.3%	0.236	0.665	BT	79.4%	0.169	0.590	60.9%	0.267	0.862
BT(Hard)	75.9%	0.194	0.703	57.1%	0.295	0.962	BT(Hard)	77.7%	0.178	0.639	60.3%	0.285	1.015
Gaussian	73.2%	0.201	0.584	57.7%	0.242	0.680	Gaussian	78.7%	0.168	0.644	60.3%	0.264	0.932
Two-anchor-Skywork	**82.1%**	**0.163**	**0.490**	**60.0%**	**0.234**	**0.664**	Two-anchor-Skywork	82.7%	0.147	0.506	**63.0%**	0.244	0.782
Two-anchor-DeepSeek-Pro	78.4%	0.180	0.534	56.3%	0.244	0.682	Two-anchor-DeepSeek-Pro	79.9%	0.167	0.551	61.4%	**0.244**	**0.732**
Two-anchor-DeepSeek-Flash	80.4%	0.167	0.502	58.2%	0.239	0.672	Two-anchor-DeepSeek-Flash	**83.8%**	**0.141**	**0.457**	61.4%	0.246	0.742
Dataset: HelpSteer2-9K							**Dataset: PersonalLLM-40K**						
BT	84.8%	0.140	0.444	58.8%	0.251	0.719	BT	76.8%	0.209	0.625	59.2%	0.252	0.729
BT(Hard)	81.1%	0.149	0.509	58.4%	0.284	0.932	BT(Hard)	74.2%	0.245	1.114	59.5%	0.305	1.251
Gaussian	81.9%	0.161	0.520	58.2%	0.251	0.729	Gaussian	73.5%	0.220	0.644	59.5%	0.253	0.725
Two-anchor-Skywork	**85.5%**	**0.132**	**0.413**	**60.2%**	0.241	0.694	Two-anchor-Skywork	**80.3%**	**0.181**	**0.546**	**61.2%**	0.240	0.696
Two-anchor-DeepSeek-Pro	82.1%	0.157	0.483	59.4%	**0.234**	**0.660**	Two-anchor-DeepSeek-Pro	77.1%	0.192	0.566	60.7%	0.241	0.688
Two-anchor-DeepSeek-Flash	83.0%	0.151	0.467	60.0%	0.237	0.672	Two-anchor-DeepSeek-Flash	78.4%	0.188	0.558	60.8%	**0.238**	**0.680**

Proposition 3 (Global quadratic curvature of the two-anchor population loss)**.** *We define the population anchor loss $\mathcal{M}_{\text{anchor}}(r, g)$ and the optimal population anchor loss $\mathcal{M}_{\text{anchor}}(r^*, g^*)$. Under Assumptions 2–3, there exists a constant $c_{\text{anc}} > 0$, depending only on B_Θ, τ_1, τ_2, such that for all $(r, g) \in \Theta_K(B_\Theta)$,*

$$\mathcal{M}_{\text{anchor}}(r, g) - \mathcal{M}_{\text{anchor}}(r^*, g^*) \geq c_{\text{anc}} \left(\|r - r^*\|^2_{L_2(\mathbb{P}_\psi)} + \|g - g^*\|^2_{L_2(\mathbb{P}_\psi)} \right).$$

Without two-anchor labels, the preference loss alone is invariant to certain transformations discussed in Section 3. Two distinct anchor thresholds break the degeneracy, and Proposition 3 provides a quadratic lower bound on the anchor population excess loss, which extends to the joint population excess loss (See Corollary 1 in Appendix). This quadratic curvature converts the empirical fluctuation between the empirical and population excess losses into estimation error bounds, yielding the following non-asymptotic rate.

Theorem 1 (Non-asymptotic rate for the sieve MLE)**.** *Under Assumptions 1, 2 and 3, there exists a constant $C > 0$, depending only on B_Θ, τ_1, τ_2, such that for every $t \geq 1$, with probability at least $1 - 2e^{-t}$,*

$$\|\hat{r} - r^*\|_{L_2(\mathbb{P}_\psi)} + \|\hat{g} - g^*\|_{L_2(\mathbb{P}_\psi)} \leq C \Big[\sqrt{1 + \lambda^{-1}} \left(K^{-\alpha_r/d} + K^{-\alpha_g/d} \right) + \lambda^{-1} \sqrt{\tfrac{K+t}{n_{\text{pref}}}} + \sqrt{\tfrac{K+t}{n_{\text{anc}}}} \Big] \tag{7}$$

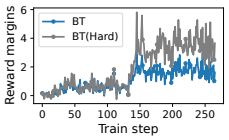
In particular, with $\alpha := \min\{\alpha_r, \alpha_g\}$,

$$\|\hat{r} - r^*\|_{L_2(\mathbb{P}_\psi)} + \|\hat{g} - g^*\|_{L_2(\mathbb{P}_\psi)} \leq C \Big[\sqrt{1 + \lambda^{-1}}\, K^{-\alpha/d} + \lambda^{-1} \sqrt{\tfrac{K+t}{n_{\text{pref}}}} + \sqrt{\tfrac{K+t}{n_{\text{anc}}}} \Big] \tag{8}$$

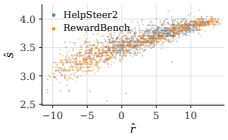
The bound in Theorem 1 consists of two parts: a deterministic approximation error $\sqrt{1 + \lambda^{-1}} K^{-\alpha/d}$ arising from the sieve approximation of (r^*, g^*), and a stochastic estimation error $\lambda^{-1} \sqrt{(K+t)/n_{\text{pref}}} + \sqrt{(K+t)/n_{\text{anc}}}$ controlled by the smaller of the two sample sizes. Notably, (i) the approximation error vanishes as $K \to \infty$ at the standard nonparametric rate; (ii) Under Assumption 4, the stochastic term vanishes as $n_{\min} \to \infty$, and is governed by $n_{\min}$ rather than $n_{\text{pref}} + n_{\text{anc}}$, reflecting the complementary roles of the two datasets. Balancing the two parts via $K \asymp n_{\min}^{d/(2\alpha+d)}$ yields the optimal rate $\mathcal{O}_p(n_{\min}^{-\alpha/(2\alpha+d)})$, which matches the classical minimax rate for nonparametric regression [Stone, 1982]. Whereas existing RLHF analyses [Zhu et al., 2023] focus on the reward mean function, the proposed estimator additionally recovers the variance function, without sacrificing the classical nonparametric convergence rate under the two-anchor identification condition. Auxiliary results and Detailed proofs are provided in Appendices A and B, respectively.

6 Experiments

Datasets. We evaluate our method on public preference datasets with annotator disagreement. By leveraging multiple annotations per comparison, we construct empirical soft labels that provide a natural signal for modeling preference uncertainty. After preprocessing public datasets that satisfy this requirement, we obtain four diverging preference datasets with empirical soft labels: `MultiPref-8K` [Miranda et al., 2025], `HelpSteer2-9K` [Wang et al., 2025a], `HelpSteer3-20K` [Wang et al., 2025b], `PersonalLLM-40K` [Zollo et al., 2025]. These datasets vary in both size and soft-label



(a)



(b)

(c)

Method	Pearson($\hat{r}$, $\hat{s}$)			
	MP	HS2	HS3	PLLM
Gaussian	0.291	**0.859**	**0.592**	**0.787**
Two-anchor-Skywork	**0.412**	0.041	-0.252	0.356
Two-anchor-DeepSeek-Pro	-0.116	-0.553	-0.137	-0.454
Two-anchor-DeepSeek-Flash	0.158	-0.110	0.006	-0.061

Notes: MP, HS2, HS3, PLLM denote MultiPref-8K, HelpSteer2-9K, HelpSteer3-20K, PersonalLLM-40K.

(d)

Figure 1: **Uncertainty representation. (a)** BT vs BT(Hard) reward margin. **(b)** Estimated $\hat{r}$ versus $\hat{s}$ for Gaussian trained on `HelpSteer2`. **(c)** Estimated $\hat{r}$ versus $\hat{s}$ for Two-anchor-DeepSeek-Flash trained on `HelpSteer2`. **(d)** Pearson correlation between estimated $\hat{r}$ and $\hat{s}$. Results are all based on the Llama-3.2-3B backbone.

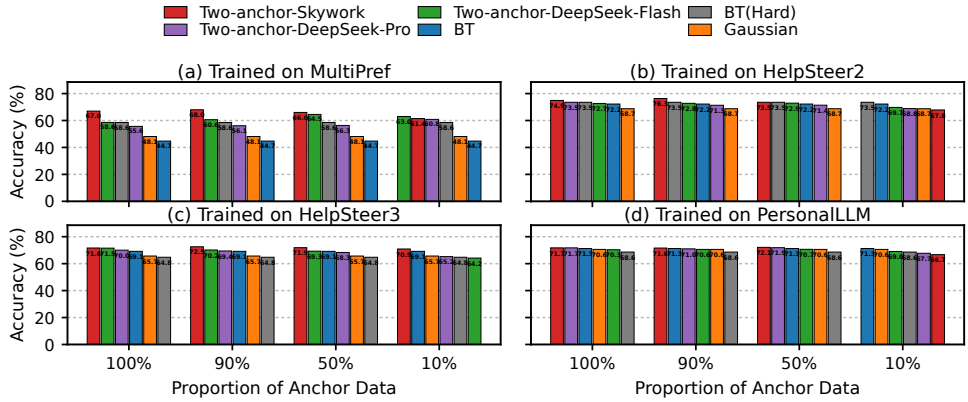


Figure 2: **Anchor data efficiency on RewardBench Accuracy.** Results are all based on the Llama-3.2-1B backbone.

distributions, covering different levels of preference disagreement (see Figure 7 in Appendix). Each dataset is split into training and held-out validation sets. Further details about dataset preprocessing and training are provided in Appendix D.

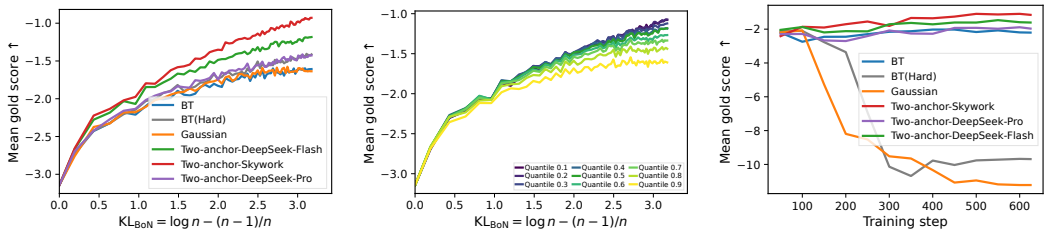
Baselines. We employ the Llama-3 series of models (`Llama-3.1-8B-Instruct`, `Llama-3.2-3B-Instruct` and `Llama-3.2-1B-Instruct`) as the backbone for training reward model. We compare four reward modeling approaches. (1) **BT**: a standard reward model with the BT loss. (2) **BT(Hard)**: a BT reward model trained on hard label derived from majority voting (3) **Gaussian**: a standard gaussian reward model predicting both mean r_ϕ and standard deviation s_ϕ. (4) **Two-anchor variants**: our proposed model, extending Gaussian with two anchors. We implement three variants: **Two-anchor-Skywork**, utilizing `Skywork-Reward-V2-Llama-3.2-3B` as the external reward model; **Two-anchor-DeepSeek-Pro** and **Two-anchor-DeepSeek-Flash**, employing `DeepSeek-V4-Pro` and `DeepSeek-V4-Flash` as LLM judges.

6.1 Evaluation on reward modeling

Following prior work, we train reward models on four datasets and then evaluate the learned models under two reward benchmarks, RewardBench [Lambert et al., 2025] and PPE-P [Frick et al., 2025]. We report pairwise accuracy and calibration metrics, including Cross-Entropy and Brier score.

Main results. Table 1 shows that Two-anchor variants achieve superior performance, not only achieving higher accuracy but also yielding better calibration metrics compared to other baselines. The improvements are consistent when using anchor labels produced by different strategies, including Skywork and DeepSeek-based variants. Meanwhile, Gaussian and BT demonstrate comparable performance, which supports our motivation that variance modeling requires additional identification constraints. Notably, BT(Hard) tends to degrade calibration metrics despite moderate accuracy gains. Full results across all backbones are provided in Appendix D.4

Uncertainty representation. Figure 1 illustrates how preference uncertainty is represented by different models. As shown in Figure 1(a), BT trained on soft label produces smaller reward margins than BT(Hard), which confirms that annotator disagreement is represented by shrinking the reward



(a) (b) (c)

Figure 3: **Mean gold score of BoN and PPO. (a)** BoN performance VS baselines. **(b)** BoN performance of Two-anchor under different reward distribution quantiles. **(c)** PPO performance VS baselines. Proxy reward models are all trained on the `MultiPref`.

margin. The Gaussian model introduces a variance head but $\hat{r}$ and $\hat{s}$ remain strongly correlated without anchor supervision (Figures 1(b)). In contrast, Two-anchor variants weaken this dependence, as shown in Figures 1(c) and Table 1(d). These results demonstrate that even coarse anchor labels provide sufficient identification signal to disentangle reward quality from preference uncertainty.

Anchor data efficiency. Figure 2 illustrates the effect of varying anchor data availability (from 10% to 100%) on RewardBench accuracy. Even with only 50% of anchor data, Two-anchor variants still outperform nearly all baselines with little performance drop, demonstrating that our method is robust to limited anchor data. Such stability effectively reduces the cost of anchor data collection in practice.

6.2 Evaluation on RLHF

We evaluate the downstream performance of reward models through both PPO training and best-of-N (BoN) evaluation. Following prior work [Gao et al., 2023, Coste et al., 2023, Yang et al., 2024, Zhang et al., 2026b], we select `Llama-3.2-1B-Instruct` as the base policy model and `UltraRM-13b`, a strong open-source reward model trained on `UltraFeedback` with reported state-of-the-art performance among open-source reward models, as the gold reward model. The learned reward model based on the `Llama-3.2-3B-Instruct` backbone serves as the proxy reward model. In PPO experiment, we downsample 20K samples from the `Unified-Feedback` dataset to optimize the policy, reserving an additional 500 samples as a held-out set for evaluation. The optimized policy then generates responses for the held-out prompts, which are evaluated by the gold reward model. For BoN, we apply the same 500 held-out prompts, selecting the best of n responses per prompt based on proxy reward models. The selected responses are then evaluated by the gold reward model. For both PPO and BoN, we average gold scores across all prompts to assess true quality.

Figure 3 presents PPO and BoN results on the `UltraFeedback` dataset. As shown in Figure 3(a), Two-anchor variants consistently outperform all baselines in BoN, with Two-anchor-Skywork achieving the best performance. We further evaluate the effect of quantile-based reward, which is defined as $\hat{r}_q(u) = \hat{r}(u) + \Phi^{-1}(q)\hat{s}(u)$. Figure 3(b) shows that lower quantile rewards yield better mean gold scores. This suggests that penalizing response uncertainty leads to more reliable selection. As shown in Figure 3(c), Two-anchor variants maintain stable and strong performance throughout PPO training. Notably, BT, BT(Hard) and Gaussian perform similarly in BoN but BT(Hard) and Gaussian quickly degrade during PPO training due to reward hacking.

7 Conclusion

In this paper, we propose an effective framework for identifiable variance-aware reward modeling under pluralistic human preferences. On the theoretical side, the two anchors induce a global quadratic curvature of the population loss that drives the estimator to the classical nonparametric minimax rate. On the empirical side, our method decouples the learned reward mean from the learned variance, remains effective with only 50% of the anchor data and enhances the Gaussian model against reward hacking during PPO training. These findings suggest that two coarse anchors label are not only auxiliary supervision, but also provide a practical and effective way to make Gaussian reward models identifiable.

References

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

Connor Baumler, Anna Sotnikova, and Hal Daumé III. Which examples should be multiply annotated? active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, 2023.

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

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification, Featured Certification.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. In *International Conference on Machine Learning*, pages 6116–6135. PMLR, 2024.

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

Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. Pal: Sample-efficient personalized reward modeling for pluralistic alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

Longze Chen, Lu Wang, Renke Shan, Ze Gong, Run Luo, Jiaming Li, Jing Luo, Qiyao Wang, and Min Yang. Learning ordinal probabilistic reward from preferences. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0Vf5trUAVF.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

Evan Frick. *Reward Modeling for Human Preferences*. PhD thesis, Master's thesis, EECS Department, University of California, Berkeley, May, 2025.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cbttLtO94Q.

Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

Daniel Halpern, Evi Micha, Ariel D. Procaccia, and Itai Shapira. Pairwise calibrated rewards for pluralistic alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=dtH7hOwTeS.

Jian Hu, Xibin Wu, Zilin Zhu, Weixun Wang, Dehao Zhang, Yu Cao, et al. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 6, 2024.

Wenlong Ji, Weizhe Yuan, Emily Getzen, Kyunghyun Cho, Michael I Jordan, Song Mei, Jason Weston, Weijie J Su, Jing Xu, and Linjun Zhang. An overview of large language models for statisticians. *The American Statistician*, (just-accepted):1–106, 2026.

Gihoon Kim and Euntai Kim. Swap-guided preference learning for personalized reinforcement learning from human feedback. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nc28mSbyVG.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, 2025.

Pangpang Liu, Junwei Lu, and Will Wei Sun. Uncertainty quantification for large language model reward learning under heterogeneous human feedback, 2025a. URL https://arxiv.org/abs/2512.03208.

Pangpang Liu, Chengchun Shi, and Will Wei Sun. Reinforcement learning from human feedback: A statistical perspective. *arXiv preprint arXiv:2604.02507*, 2026.

Shang Liu, Yu Pan, Guanting Chen, and Xiaocheng Li. Reward modeling with ordinal feedback: Wisdom of the crowd. In *Forty-second International Conference on Machine Learning*, 2025b. URL https://openreview.net/forum?id=2JRrmzPQSc.

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*, 2024.

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. ai feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7162–7200, 2025.

Abhilash Mishra. Ai alignment and social choice: Fundamental limitations and policy implications. *arXiv preprint arXiv:2310.16048*, 2023.

Whitney K Newey. Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168, 1997.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=gRG6SzbW9p.

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

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, pages 580–615, 1994.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. In *The Twelfth International Conference on Learning Representations*, 2024.

Nicki Skafte, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947, 2024a.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=gQpBnRHwxM.

Andrew Stirn, Harm Wessels, Megan Schertzer, Laura Pereira, Neville Sanjana, and David Knowles. Faithful heteroscedastic regression with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 5593–5613. PMLR, 2023.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

Wangtao Sun, Xiang Cheng, Xing Yu, Haotian Xu, Zhao Yang, Shizhu He, Jun Zhao, and Kang Liu. Probabilistic uncertain reward model. *arXiv preprint arXiv:2503.22480*, 2025.

Louis L Thurstone. The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21(4):384, 1927.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models, 2024.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=MnfHxPP5gs.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages, 2025b. URL https://arxiv.org/abs/2505.11475.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for llms. *Advances in Neural Information Processing Systems*, 37:62279–62309, 2024.

Michael Jq Zhang, Zhilin Wang, Jena D Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. Diverging preferences: When do annotators disagree and do models know? In *International Conference on Machine Learning*, pages 76193–76212. PMLR, 2025.

Yudi Zhang, Lu Wang, Meng Fang, Yali Du, Chenghua Huang, Jun Wang, Qingwei Lin, Mykola Pechenizkiy, Dongmei Zhang, and Saravan Rajmohan. Distill not only data but also rewards: Can smaller language models surpass larger ones?, 2026a. URL https://openreview.net/forum?id=tK6VZy5RYr.

Zhiwei Zhang, Hui Liu, Xiaomin Li, Zhenwei Dai, Jingying Zeng, Fali Wang, Minhua Lin, Ramraj Chandradevan, Linlin Wu, Zhen Li, Chen Luo, Zongyu Wu, Xianfeng Tang, Qi He, and Suhang Wang. Bradley-terry and multi-objective reward modeling are complementary. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=3QHKJcwnpb.

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

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. Personal-LLM: Tailoring LLMs to individual preferences. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2R7498e2Tx.

A Addition auxiliary results for the Theory Section

This appendix collects the auxiliary results supporting the proof of Theorem 1. Proposition 4 ensures that all true preference and anchor probabilities are bounded away from 0 and 1. Proposition 5 characterizes the nondegeneracy of the two-anchor map and provides the geometric intuition behind the curvature in Proposition 3. Corollary 1 extends the anchor curvature to the joint population loss, while Lemma 1 provides a quadratic upper bound under Assumption 1. Lemma 2 bounds the empirical-process term uniformly over the truncated sieve space. Combining Corollary 1, Lemma 2 and the sieve approximation in Assumption 3, we obtain Theorem 1. All proofs are deferred to Appendix B.

Assumption (Full version of Assumption 1)**.** *Let $\mathbb{P}_\psi$ denote a marginal distribution of the feature representation $\psi(x, y)$. The anchor and preference datasets are generated as follows.* (i) Anchor. *For each anchor sample (x, y), the feature $\psi(x, y) \sim \mathbb{P}_\psi$.* (ii) Preference. *For each preference sample (x, y_1, y_2), the two responses are conditionally i.i.d. given x, $y_1, y_2 \mid x \sim \pi(\cdot \mid x)$. The feature representation $u_1 = \psi(x, y_1)$ and $u_2 = \psi(x, y_2)$ satisfy $u_1 \sim \mathbb{P}_\psi$ and $u_2 \sim \mathbb{P}_\psi$.*

Assumption (Full version of Assumption 2)**.** *The fixed feature mapping $\psi(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathcal{U} \subset [-\psi_{\max}, \psi_{\max}]^d$ admits a density bounded above and below by positive constants. The true mean and log-standard-deviation functions satisfy $r^* \in \mathcal{H}(\alpha_r, L_r; \mathcal{U})$ and $g^* \in \mathcal{H}(\alpha_g, L_g; \mathcal{U})$ for some Hölder smoothness indices $\alpha_r, \alpha_g > 0$ and constants $L_r, L_g > 0$. Moreover, there exists a positive constant B such that $\|r^*\|_\infty \le B$ and $\|g^*\|_\infty \le B$.*

Assumption (Full version of Assumption 3)**.** *The sieve spaces $\mathcal{R}_K$ and $\mathcal{G}_K$ satisfy:* (i) Approximation. *There exists a constant $C > 0$, independent of K, such that $\inf_{(r,g)\in\Theta_K(B_\Theta)} \|r - r^*\|_{L_2(\mathbb{P}_\psi)} \le CK^{-\alpha_r/d}$ and $\inf_{(r,g)\in\Theta_K(B_\Theta)} \|g - g^*\|_{L_2(\mathbb{P}_\psi)} \le CK^{-\alpha_g/d}$;* (ii) Basis functions stability. *Each $r \in \mathcal{R}_K$ and $g \in \mathcal{G}_K$ admits a basis expansion $r(u) = \beta_r^\top B_{r,K}(u)$, $g(u) = \beta_g^\top B_{g,K}(u)$, and there exist constants $0 < c_B < C_B < \infty$, independent of K, satisfying $c_B\|\beta_r - \beta_r'\|_2 \le \|r - r'\|_{L_2(\mathbb{P}_\psi)} \le C_B\|\beta_r - \beta_r'\|_2$ and $c_B\|\beta_g - \beta_g'\|_2 \le \|g - g'\|_{L_2(\mathbb{P}_\psi)} \le C_B\|\beta_g - \beta_g'\|_2$ for all $r, r' \in \mathcal{R}_K$ and $g, g' \in \mathcal{G}_K$.*

Proposition 4 (Overlap boundedness)**.** *Under Assumption 2, there exists a constant $c_0 \in (0, 1/3)$, depending only on B and τ_1, τ_2, such that for all $u_1, u_2, u \in \mathcal{U}$, $c_0 \le \Phi\left(\frac{r^*(u_1) - r^*(u_2)}{\sqrt{e^{2g^*(u_1)} + e^{2g^*(u_2)}}}\right) \le 1 - c_0$. Moreover, let true anchor probability $q_k^*(u) := \Phi\left(\frac{r^*(u) - \tau_k}{e^{g^*(u)}}\right)$ for $k \in \{1, 2\}$, and define the three threshold-induced anchor classes class probabilities $\pi_0^*(u) = 1 - q_1^*(u)$, $\pi_1^*(u) = q_1^*(u) - q_2^*(u)$ and $\pi_2^*(u) = q_2^*(u)$. Then $\pi_\ell^*(u) \ge c_0$ for all $\ell \in \{0, 1, 2\}$*

Proposition 4 ensures that the true preference probability is bounded away from 0 and 1, and that each true anchor class has a uniformly positive probability.

Proposition 5 (Two-anchor map nondegeneracy on bounded sets)**.** *For each $u \in \mathcal{U}$, define the two-anchor map on pointwise values $(r(u), g(u)) \in \mathbb{R}^2$ by*

$$\Gamma_u(r(u), g(u)) = \left(\Phi\left(\tfrac{r(u) - \tau_1}{e^{g(u)}}\right),\ \Phi\left(\tfrac{r(u) - \tau_2}{e^{g(u)}}\right)\right),$$

and let $J_{\Gamma_u}(r(u), g(u))$ denote its Jacobian. The Jacobian satisfies

$$\det J_{\Gamma_u}(r(u), g(u)) = \varphi(z_1)\,\varphi(z_2)\,e^{-2g(u)}\,(\tau_2 - \tau_1), \qquad z_k := (r(u) - \tau_k)e^{-g(u)},$$

where φ is the standard normal density. Moreover, for every $R > 0$, there exists a constant $c_J = c_J(R, \tau_1, \tau_2) > 0$, independent of u, such that

$$J_{\Gamma_u}(r(u), g(u))^\top J_{\Gamma_u}(r(u), g(u)) \succeq c_J\, I_2 \qquad \text{for all } u \in \mathcal{U} \text{ and } (r(u), g(u)) \in [-R, R]^2.$$

Remark 1 (Effect of anchor thresholds configuration)**.** *The constant c_{anc} in Proposition 3 depends on the configuration of the two anchor thresholds (τ_1, τ_2) through two competing effects: a gap effect from $\tau_2 - \tau_1$ and a tail effect from $\max\{|\tau_1|, |\tau_2|\}$. Notably, (i) The anchor gap $\tau_2 - \tau_1$ affects the curvature constant c_{anc} and two-anchor map's Jacobian (see Proposition 5 in Appendix). A nonzero gap is necessary to ensure identifiability. if the gap is too small, the two anchor probabilities become nearly redundant and the curvature weakens; (ii) Extreme thresholds push the anchor probabilities $\Phi((r^*(u) - \tau_k)/e^{g^*(u)})$ toward 0 or 1, reducing the information carried by the anchor labels and again weakening the curvature.*

Corollary 1 (Curvature lower bound for the joint population loss)**.** *We denote the population preference loss by*

$$\mathcal{M}_{\mathrm{pref}}(r,g) = -\,\mathbb{E}_{(u_1,u_2)\sim\mathbb{P}_\psi^{\mathrm{pref}}}\left[p^*(u_1,u_2)\log p(u_1,u_2) + \big(1-p^*(u_1,u_2)\big)\log\big(1-p(u_1,u_2)\big)\right],$$

where $p(u_1,u_2) = \Phi\left(\frac{r(u_1)-r(u_2)}{\sqrt{e^{2g(u_1)}+e^{2g(u_2)}}}\right)$ *and* $p^*(u_1,u_2)$ *is defined similarly. The population anchor loss* $\mathcal{M}_{\mathrm{anchor}}$ *was defined in Proposition 3. We define the joint population loss as*

$$\mathcal{M}(r,g) := \mathcal{M}_{\mathrm{pref}}(r,g) + \lambda\,\mathcal{M}_{\mathrm{anchor}}(r,g).$$

Under Assumptions 2–3, there exists a constant $c_{\mathrm{pop}} > 0$*, depending only on* $B_\Theta, \tau_1, \tau_2, \lambda$*, such that for all* $(r,g) \in \Theta_K(B_\Theta)$*,*

$$\mathcal{M}(r,g) - \mathcal{M}(r^*,g^*) \;\geq\; c_{\mathrm{pop}}\left(\|r-r^*\|_{L_2(\mathbb{P}_\psi)}^2 + \|g-g^*\|_{L_2(\mathbb{P}_\psi)}^2\right).$$

Lemma 1 (Quadratic upper bound on the population excess loss)**.** *Under Assumptions 1,2 and 3, there exists a constant* $C_{\mathrm{pop}} > 0$*, depending only on* $B_\Theta, \tau_1, \tau_2, \lambda$*, such that for all* $(r,g) \in \Theta_K(B_\Theta)$*,*

$$\mathcal{M}(r,g) - \mathcal{M}(r^*,g^*) \;\leq\; C_{\mathrm{pop}}\left(\|r-r^*\|_{L_2(\mathbb{P}_\psi)}^2 + \|g-g^*\|_{L_2(\mathbb{P}_\psi)}^2\right).$$

Lemma 2 (Empirical-process bound on the truncated sieve)**.** *We denote*

$$R(r,g) := \|r-r^*\|_{L_2(\mathbb{P}_\psi)} + \|g-g^*\|_{L_2(\mathbb{P}_\psi)}$$

for the joint L_2 *error, and*

$$G(r,g) := \big[\mathcal{L}(r,g) - \mathcal{L}(r^*,g^*)\big] - \big[\mathcal{M}(r,g) - \mathcal{M}(r^*,g^*)\big]$$

for the centered empirical fluctuation between the empirical and population excess losses. Under Assumptions 1, 2 and 3, for every $t \geq 1$*, with probability at least* $1-2e^{-t}$*, for all* $(r,g) \in \Theta_K(B_\Theta)$*,*

$$\big|G(r,g)\big| \leq C\big\{R(r,g)\lambda\Delta_t + \lambda\Delta_t^2\big\}.$$

where $\Delta_t := \lambda^{-1}\sqrt{(K+t)/n_{\mathrm{pref}}} + \sqrt{(K+t)/n_{\mathrm{anc}}}$*. The constant* $C > 0$ *depends only on* B_Θ, τ_1, τ_2*, and is independent of* $K, n_{\mathrm{pref}}, n_{\mathrm{anc}}$*, and* λ*.*

B Proofs for the Theory Section

B.1 Proof of Proposition 4

Proof of Proposition 1. For notational simplicity, we define the true preference margin and the true anchor threshold scores as $z^*_{\mathrm{pref}}(u_1,u_2) = \frac{r^*(u_1)-r^*(u_2)}{\sqrt{e^{2g^*(u_1)}+e^{2g^*(u_2)}}}$ and $z^*_{\mathrm{anc},k}(u) := \frac{r^*(u)-\tau_k}{e^{g^*(u)}}$ for $k \in \{1,2\}$.

By Assumption 2, we have $|r^*(u)| \leq B$ and $|g^*(u)| \leq B$ for all $u \in \mathcal{U}$. It follows that $|r^*(u_1) - r^*(u_2)| \leq 2B$ and $\sqrt{e^{2g^*(u_1)} + e^{2g^*(u_2)}} \geq \sqrt{2}\,e^{-B}$.

Therefore, combining these two bounds yields

$$|z^*_{\mathrm{pref}}(u_1,u_2)| \;\leq\; \frac{2B}{\sqrt{2}\,e^{-B}} \;=\; \sqrt{2}\,Be^{B}. \tag{A.1}$$

Again by Assumption 2, we obtain

$$|z^*_{\mathrm{anc},k}(u)| \leq e^B(B + |\tau_k|), \qquad k \in \{1,2\}. \tag{A.2}$$

Then (A.1) and (A.2) imply the boundedness of the true preference margin and the true anchor threshold scores:

$$|z^*_{\mathrm{pref}}(u_1,u_2)| \leq M, \qquad |z^*_{\mathrm{anc},k}(u)| \leq M, \quad k \in \{1,2\},$$

where $M = \max\big\{\sqrt{2}\,Be^B,\, e^B(B+|\tau_1|),\, e^B(B+|\tau_2|)\big\}$.

Since the standard normal cumulative distribution function Φ is strictly increasing, it follows that $\Phi(-M) \le \Phi\Big(z^*_{\mathrm{pref}}(u_1,u_2)\Big) \le \Phi(M) = 1-\Phi(-M)$. This proves the boundedness of the true preference margin.

We next prove the lower bounds for the three threshold-induced anchor class probabilities. By the boundedness of $z^*_{\mathrm{anc},k}(u)$, we have the boundedness for true anchor probability

$$\Phi(-M) \le q^*_k(u) \le 1-\Phi(-M), \qquad k \in \{1,2\}.$$

Therefore, we have

$$\pi^*_0(u) = 1-q^*_1(u) \ge \Phi(-M), \qquad \pi^*_2(u) = q^*_2(u) \ge \Phi(-M).$$

For the middle class probability, since $\tau_1 < \tau_2$, we have

$$z^*_{\mathrm{anc},1}(u) - z^*_{\mathrm{anc},2}(u) = \frac{\tau_2-\tau_1}{e^{g^*(u)}} \ge (\tau_2-\tau_1)e^{-B}.$$

By the mean value theorem and $z^*_{\mathrm{anc},1}(u), z^*_{\mathrm{anc},2}(u) \in [-M,M]$, we obtain

$$\pi^*_1(u) = q^*_1(u) - q^*_2(u) = \Phi\big(z^*_{\mathrm{anc},1}(u)\big) - \Phi\big(z^*_{\mathrm{anc},2}(u)\big) \ge \varphi(M)(\tau_2-\tau_1)e^{-B},$$

where φ is the standard normal density.

Thus, we have $c_0 \le \Phi\Big(z^*_{\mathrm{pref}}(u_1,u_2)\Big) \le 1-c_0$ and $\pi^*_\ell(u) \ge c_0$ for $\ell \in \{0,1,2\}$ by setting $c_0 := \frac{1}{2}\min\big\{\frac{1}{3},\, \Phi(-M),\, \varphi(M)(\tau_2-\tau_1)e^{-B}\big\} \in (0,1/3)$ This completes the proof. □

B.2 Proof of Proposition 3

Proof of Global quadratic curvature of the two-anchor population loss. We define the population anchor loss $\mathcal{M}_{\mathrm{anchor}}(r,g) = \mathbb{E}_{u\sim\mathbb{P}_\psi}\Big[-\sum_{\ell=0}^{2}\pi^*_\ell(u)\log\pi_\ell(u)\Big]$, where $q_k(u) := \Phi\Big(\frac{r(u)-\tau_k}{e^{g(u)}}\Big)$ and $q^*_k(u) := \Phi\Big(\frac{r^*(u)-\tau_k}{e^{g^*(u)}}\Big)$ for $k \in \{1,2\}$ and the three threshold-induced anchor class probabilities are defined as $\pi_0(u) = 1-q_1(u)$, $\pi_1(u) = q_1(u)-q_2(u)$, and $\pi_2(u) = q_2(u)$ with $\pi^*_\ell(u)$ defined similarly. For $k=1,2$ and $u \in \mathcal{U}$, we define $z_{\mathrm{anc},k}(u) = \frac{r(u)-\tau_k}{e^{g(u)}} = (r(u)-\tau_k)e^{-g(u)}$, $z^*_{\mathrm{anc},k}(u) = \frac{r^*(u)-\tau_k}{e^{g^*(u)}}$ and write $q_k(u) = \Phi(z_{\mathrm{anc},k}(u))$ and $q^*_k(u) = \Phi(z^*_{\mathrm{anc},k}(u))$. Let $\tau_{\max} := \max\{|\tau_1|,|\tau_2|\}$ and define $M := e^{B_\Theta}(B_\Theta + \tau_{\max})$ and $M^* := e^B(B+\tau_{\max})$.

By Assumption 2 and the truncation $(r,g) \in \Theta_K(B_\Theta)$, we have $|r(u)|, |g(u)| \le B_\Theta$ and $|r^*(u)|, |g^*(u)| \le B$. Following the same strategy as used in the proof of Proposition 4, we obtain

$$|z_{\mathrm{anc},k}(u)| \le M, \qquad |z^*_{\mathrm{anc},k}(u)| \le M^*, \qquad k \in \{1,2\}. \tag{A.3}$$

We first establish the Lipschitz bound from q_k to $z_{\mathrm{anc},k}$. Combining (A.3) and the strictly increasing property of Φ, we have

$$\Phi(-M) \le q_k(u), q^*_k(u) \le \Phi(M).$$

Since Φ^{-1} is continuously differentiable with derivative $1/\varphi(\Phi^{-1}(\cdot))$ and attains its maximum at the endpoints, Φ^{-1} is L_Φ-Lipschitz on $[\Phi(-M),\Phi(M)]$ with $L_\Phi = \frac{1}{\varphi(M)}$. It follows that, for every $u \in \mathcal{U}$ and $k=1,2$,

$$|z_{\mathrm{anc},k}(u) - z^*_{\mathrm{anc},k}(u)| \le L_\Phi |q_k(u) - q^*_k(u)|. \tag{A.4}$$

We next establish a pointwise Lipschitz bound from $(z_{\mathrm{anc},1}(u), z_{\mathrm{anc},2}(u))$ to $(r(u),g(u))$. For notational simplicity, $z_1(u) = z_{\mathrm{anc},1}(u)$ and $z_2(u) = z_{\mathrm{anc},2}(u)$ in the following calculation. Since $\tau_2 > \tau_1$, for every $u \in \mathcal{U}$, $z_1(u) - z_2(u) = (\tau_2-\tau_1)e^{-g(u)} > 0$. Moreover, $r(u)$ and $g(u)$ follow the map in terms of $z_1(u)$ and $z_2(u)$:

$$r(u) = \frac{\tau_2 z_1(u) - \tau_1 z_2(u)}{z_1(u)-z_2(u)}, \qquad g(u) = \log\frac{\tau_2-\tau_1}{z_1(u)-z_2(u)}. \tag{A.5}$$

The bounds $|g| \le B_\Theta$ and $|z_{\mathrm{anc},k}| \le M$ imply that the corresponding pairs $(z_{\mathrm{anc},1}(u), z_{\mathrm{anc},2}(u))$ and $(z^*_{\mathrm{anc},1}(u), z^*_{\mathrm{anc},2}(u))$ both belong to the compact convex set

$$D := \Big\{(z_1(u), z_2(u)) \in \mathbb{R}^2 : (\tau_2 - \tau_1)e^{-B_\Theta} \le z_1(u) - z_2(u) \le (\tau_2 - \tau_1)e^{B_\Theta}\Big\}, \tag{A.6}$$

where $\max\{|z_1(u)|, |z_2(u)|\} \le M$.

The Jacobian of the map in (A.5) is

$$J_{\mathrm{inv}}(z_1(u), z_2(u)) = \begin{pmatrix} -\dfrac{(\tau_2 - \tau_1)z_2(u)}{(z_1(u) - z_2(u))^2} & \dfrac{(\tau_2 - \tau_1)z_1(u)}{(z_1(u) - z_2(u))^2} \\ -\dfrac{1}{z_1(u) - z_2(u)} & \dfrac{1}{z_1(u) - z_2(u)} \end{pmatrix}.$$

Based on the convex set D (A.6), we have $z_1(u) - z_2(u) \ge \delta_\Theta > 0$ and $\max\{|z_1(u)|, |z_2(u)|\} \le M$, where $\delta_\Theta := (\tau_2 - \tau_1)e^{-B_\Theta}$. Therefore, $\sup_{(z_1(u), z_2(u)) \in D} \|J_{\mathrm{inv}}(z_1(u), z_2(u))\|_F \le L_z$, where

$$\begin{aligned} \|J_{\mathrm{inv}}(z_1(u), z_2(u))\|_F^2 &= \left(\frac{(\tau_2 - \tau_1)z_2(u)}{(z_1(u) - z_2(u))^2}\right)^2 + \left(\frac{(\tau_2 - \tau_1)z_1(u)}{(z_1(u) - z_2(u))^2}\right)^2 + \\ &\quad \left(\frac{1}{z_1(u) - z_2(u)}\right)^2 + \left(\frac{1}{z_1(u) - z_2(u)}\right)^2 \\ &\le 2\left(\frac{(\tau_2 - \tau_1)M}{\delta_\Theta^2}\right)^2 + 2\left(\frac{1}{\delta_\Theta}\right)^2 \\ &=: L_z^2, \end{aligned}$$

and $L_z = L_z(B_\Theta, \tau_1, \tau_2) < \infty$.

Given the convex set D defined in (A.6), applying the mean value inequality to the map defined in (A.5) yields that, for every $u \in \mathcal{U}$,

$$\begin{aligned} |r(u) - r^*(u)|^2 + |g(u) - g^*(u)|^2 &= \left\| \begin{pmatrix} r(u) \\ g(u) \end{pmatrix} - \begin{pmatrix} r^*(u) \\ g^*(u) \end{pmatrix} \right\|_2^2 \\ &\le L_z^2 \left\| \begin{pmatrix} z_1(u) \\ z_2(u) \end{pmatrix} - \begin{pmatrix} z_1^*(u) \\ z_2^*(u) \end{pmatrix} \right\|_2^2 \\ &= L_z^2 \left[|z_1(u) - z_1^*(u)|^2 + |z_2(u) - z_2^*(u)|^2\right]. \end{aligned} \tag{A.7}$$

Combining the two Lipschitz bounds (A.4) and (A.7), we obtain

$$|r(u) - r^*(u)|^2 + |g(u) - g^*(u)|^2 \le L_z^2 L_\Phi^2 \left[|q_1(u) - q_1^*(u)|^2 + |q_2(u) - q_2^*(u)|^2\right]. \tag{A.8}$$

For the anchor part, conditional on u, the pointwise population loss difference equals $\mathrm{KL}(\pi^*(u) \,\|\, \pi(u))$, where $\pi(u) = (\pi_0(u), \pi_1(u), \pi_2(u))$ and $\pi^*(u) = (\pi_0^*(u), \pi_1^*(u), \pi_2^*(u))$.

For $k \in \{1, 2\}$, we define $\delta_k(u) := q_k(u) - q_k^*(u)$. By the definitions of the three threshold-induced anchor class probabilities, we have $\pi_0(u) - \pi_0^*(u) = -\delta_1(u)$, $\pi_1(u) - \pi_1^*(u) = \delta_1(u) - \delta_2(u)$ and $\pi_2(u) - \pi_2^*(u) = \delta_2(u)$. Therefore, we obtain

$$\begin{aligned} \|\pi(u) - \pi^*(u)\|_2^2 &= |\delta_1(u)|^2 + |\delta_1(u) - \delta_2(u)|^2 + |\delta_2(u)|^2 \\ &\ge |\delta_1(u)|^2 + |\delta_2(u)|^2 \\ &= |q_1(u) - q_1^*(u)|^2 + |q_2(u) - q_2^*(u)|^2. \end{aligned} \tag{A.9}$$

By Pinsker's inequality and (A.8), we have

$$\begin{aligned} \mathrm{KL}(\pi^*(u) \,\|\, \pi(u)) &\ge \frac{1}{2}\|\pi(u) - \pi^*(u)\|_1^2 \\ &\ge \frac{1}{2}\|\pi(u) - \pi^*(u)\|_2^2 \\ &\ge \frac{1}{2}\left[|q_1(u) - q_1^*(u)|^2 + |q_2(u) - q_2^*(u)|^2\right] \\ &\ge \frac{1}{2L_z^2 L_\Phi^2}\left\{|r(u) - r^*(u)|^2 + |g(u) - g^*(u)|^2\right\}. \end{aligned} \tag{A.10}$$

Finally, integrating (A.10) with respect to $u \sim \mathbb{P}_\psi$ gives

$$\begin{aligned}\mathcal{M}_{\mathrm{anchor}}(r,g) - \mathcal{M}_{\mathrm{anchor}}(r^*,g^*) &= \mathbb{E}_{u\sim\mathbb{P}_\psi}\left[\mathrm{KL}(\pi^*(u)\,\|\,\pi(u))\right] \\ &\geq c_{\mathrm{anc}}\left(\|r-r^*\|^2_{L_2(\mathbb{P}_\psi)} + \|g-g^*\|^2_{L_2(\mathbb{P}_\psi)}\right),\end{aligned}$$

where $c_{\mathrm{anc}} := \frac{1}{2L_z^2 L_\Phi^2} > 0$. The constant c_{anc} depends only on B_Θ, τ_1, τ_2, and is independent of (r,g) and K. This completes the proof. □

B.3 Proof of Corollary 1

Proof. For any $(r,g) \in \Theta_K(B_\Theta)$,

$$\begin{aligned}&\mathcal{M}(r,g) - \mathcal{M}(r^*,g^*) \\ &\quad= \Big[\mathcal{M}_{\mathrm{pref}}(r,g) - \mathcal{M}_{\mathrm{pref}}(r^*,g^*)\Big] \\ &\qquad+ \lambda\Big[\mathcal{M}_{\mathrm{anchor}}(r,g) - \mathcal{M}_{\mathrm{anchor}}(r^*,g^*)\Big].\end{aligned}$$

For the preference component, we have

$$\mathcal{M}_{\mathrm{pref}}(r,g) - \mathcal{M}_{\mathrm{pref}}(r^*,g^*) = \mathbb{E}_{(u_1,u_2)\sim\mathbb{P}_\psi^{\mathrm{pref}}}\left[\mathrm{KL}(\mathrm{Bern}(p^*(u_1,u_2))\,\|\,\mathrm{Bern}(p(u_1,u_2)))\right] \geq 0. \tag{A.11}$$

For the anchor component, Proposition 3 yields

$$\begin{aligned}&\mathcal{M}_{\mathrm{anchor}}(r,g) - \mathcal{M}_{\mathrm{anchor}}(r^*,g^*) \\ &\quad\geq c_{\mathrm{anc}}\left(\|r-r^*\|^2_{L_2(\mathbb{P}_\psi)} + \|g-g^*\|^2_{L_2(\mathbb{P}_\psi)}\right).\end{aligned} \tag{A.12}$$

Combining (A.11) and (A.12), and using $\lambda > 0$, we have

$$\begin{aligned}&\mathcal{M}(r,g) - \mathcal{M}(r^*,g^*) \\ &\quad\geq c_{\mathrm{pop}}\left(\|r-r^*\|^2_{L_2(\mathbb{P}_\psi)} + \|g-g^*\|^2_{L_2(\mathbb{P}_\psi)}\right),\end{aligned} \tag{A.13}$$

where $c_{\mathrm{pop}} := \lambda c_{\mathrm{anc}} > 0$. □

B.4 Proof of Lemma 1

Proof of Quadratic upper bound on the population excess loss. The joint excess loss decomposes into preference and anchor parts:

$$\begin{aligned}&\mathcal{M}(r,g) - \mathcal{M}(r^*,g^*) \\ &\quad= \Big[\mathcal{M}_{\mathrm{pref}}(r,g) - \mathcal{M}_{\mathrm{pref}}(r^*,g^*)\Big] \\ &\qquad+ \lambda\Big[\mathcal{M}_{\mathrm{anchor}}(r,g) - \mathcal{M}_{\mathrm{anchor}}(r^*,g^*)\Big].\end{aligned} \tag{A.14}$$

We bound each component in (A.14). We first establish uniform KL upper bounds. Since $(r,g) \in \Theta_K(B_\Theta)$ and $\|r^*\|_\infty, \|g^*\|_\infty \leq B < B_\Theta$, the same argument as in Proposition 4 ensures that the preference probabilities $p(u_1,u_2)$ and $p^*(u_1,u_2)$ are uniformly bounded away from 0 and 1. Therefore, there exists a constant $C_{\mathrm{KL}} > 0$, depending only on B_Θ, such that for every (u_1,u_2),

$$\mathrm{KL}(\mathrm{Bern}(p^*(u_1,u_2))\,\|\,\mathrm{Bern}(p(u_1,u_2))) \leq C_{\mathrm{KL}}\left(p(u_1,u_2) - p^*(u_1,u_2)\right)^2. \tag{A.15}$$

Moreover, the three threshold-induced anchor class probabilities $\pi_\ell(u)$ and $\pi^*_\ell(u)$, $\ell \in \{0,1,2\}$, are uniformly bounded away from zero. Hence there exists a constant $C_\pi > 0$, depending only on B_Θ, τ_1, τ_2, such that

$$\mathrm{KL}(\pi^*(u)\,\|\,\pi(u)) \leq C_\pi \|\pi(u) - \pi^*(u)\|_2^2. \tag{A.16}$$

We next establish a pointwise Lipschitz bound from $(r(u), g(u))$ to the anchor probabilities. As in the proof of Proposition 3, define $z_{\mathrm{anc},k}(u) = (r(u) - \tau_k)e^{-g(u)}$ and $z^*_{\mathrm{anc},k}(u) = (r^*(u) - \tau_k)e^{-g^*(u)}$,

so that $q_k(u) = \Phi(z_{\mathrm{anc},k}(u))$ and $q_k^*(u) = \Phi(z_{\mathrm{anc},k}^*(u))$. The pointwise map $(r(u), g(u)) \mapsto z_{\mathrm{anc},k}(u) = (r(u) - \tau_k)e^{-g(u)}$ has gradient

$$\nabla_{(r(u),g(u))} z_{\mathrm{anc},k}(u) = \big(e^{-g(u)},\ -(r(u) - \tau_k)e^{-g(u)}\big).$$

Let $\tau_{\max} := \max\{|\tau_1|, |\tau_2|\}$. Since $|r(u)| \le B_\Theta$ and $|g(u)| \le B_\Theta$, its Euclidean norm satisfies

$$\big\|\nabla_{(r(u),g(u))} z_{\mathrm{anc},k}(u)\big\|_2 \le e^{B_\Theta}\sqrt{1 + (B_\Theta + \tau_{\max})^2} =: L_a nc. \tag{A.17}$$

Combining (A.17) with $\|\Phi'\|_\infty = \varphi(0) \le 1/\sqrt{2\pi}$ and the chain rule yields, for every $u \in \mathcal{U}$ and $k \in \{1, 2\}$,

$$\begin{aligned}
|q_k(u) - q_k^*(u)| &= |\Phi(z_{\mathrm{anc},k}(u)) - \Phi(z_{\mathrm{anc},k}^*(u))| \\
&\le \|\Phi'\|_\infty \, |z_{\mathrm{anc},k}(u) - z_{\mathrm{anc},k}^*(u)| \\
&\le \frac{1}{\sqrt{2\pi}} \sup_{(r,g)\in[-B_\Theta,B_\Theta]^2} \|\nabla_{(r,g)} z_{\mathrm{anc},k}(r,g)\|_2 \cdot \sqrt{|r(u) - r^*(u)|^2 + |g(u) - g^*(u)|^2} \\
&\le \frac{L_a nc}{\sqrt{2\pi}} \sqrt{|r(u) - r^*(u)|^2 + |g(u) - g^*(u)|^2}.
\end{aligned}$$

Squaring both sides yields, for every $u \in \mathcal{U}$ and $k \in \{1, 2\}$,

$$|q_k(u) - q_k^*(u)|^2 \le \frac{L_{anc}^2}{2\pi}\big(|r(u) - r^*(u)|^2 + |g(u) - g^*(u)|^2\big). \tag{A.18}$$

By the definitions in (A.9), we have

$$\begin{aligned}
\|\pi(u) - \pi^*(u)\|_2^2 &= |\delta_1(u)|^2 + |\delta_1(u) - \delta_2(u)|^2 + |\delta_2(u)|^2 \\
&\le 3\left(|\delta_1(u)|^2 + |\delta_2(u)|^2\right) \\
&= 3\left[|q_1(u) - q_1^*(u)|^2 + |q_2(u) - q_2^*(u)|^2\right].
\end{aligned} \tag{A.19}$$

Combining (A.16), (A.18), and (A.19), and integrating with respect to $u \sim \mathbb{P}_\psi$ (Assumption 1), we obtain

$$\begin{aligned}
&\mathcal{M}_{\mathrm{anchor}}(r, g) - \mathcal{M}_{\mathrm{anchor}}(r^*, g^*) \\
&\quad= \mathbb{E}_{u\sim\mathbb{P}_\psi}\left[\mathrm{KL}(\pi^*(u) \,\|\, \pi(u))\right] \\
&\quad\le C_\pi \mathbb{E}_{u\sim\mathbb{P}_\psi}\left[\|\pi(u) - \pi^*(u)\|_2^2\right] \\
&\quad\le 3C_\pi \sum_{k=1}^{2} \mathbb{E}_{u\sim\mathbb{P}_\psi}\left[|q_k(u) - q_k^*(u)|^2\right] \\
&\quad\le C_{\mathrm{anc}}\left(\|r - r^*\|_{L_2(\mathbb{P}_\psi)}^2 + \|g - g^*\|_{L_2(\mathbb{P}_\psi)}^2\right),
\end{aligned} \tag{A.20}$$

where $C_{\mathrm{anc}} := 3C_\pi L_{anc}^2/\pi$ depends only on B_Θ, τ_1, τ_2.

We now establish the Lipschitz bound from $(r(u), g(u))$ to the preference probability. Define

$$z_{\mathrm{pref}}(u_1, u_2) = \frac{r(u_1) - r(u_2)}{\sqrt{e^{2g(u_1)} + e^{2g(u_2)}}}, \qquad z_{\mathrm{pref}}^*(u_1, u_2) = \frac{r^*(u_1) - r^*(u_2)}{\sqrt{e^{2g^*(u_1)} + e^{2g^*(u_2)}}},$$

so that $p(u_1, u_2) = \Phi(z_{\mathrm{pref}}(u_1, u_2))$ and $p^*(u_1, u_2) = \Phi(z_{\mathrm{pref}}^*(u_1, u_2))$.

Write $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) := \big(r(u_1), r(u_2), g(u_1), g(u_2)\big)$, so that z_{pref} can be viewed as a function of $\theta \in \mathbb{R}^4$. A direct computation gives

$$\partial_{\theta_1} z_{\mathrm{pref}} = \frac{1}{\sqrt{e^{2\theta_3} + e^{2\theta_4}}}, \qquad \partial_{\theta_2} z_{\mathrm{pref}} = -\frac{1}{\sqrt{e^{2\theta_3} + e^{2\theta_4}}},$$

$$\partial_{\theta_3} z_{\mathrm{pref}} = -\frac{(\theta_1 - \theta_2)e^{2\theta_3}}{\big(e^{2\theta_3} + e^{2\theta_4}\big)^{3/2}}, \qquad \partial_{\theta_4} z_{\mathrm{pref}} = -\frac{(\theta_1 - \theta_2)e^{2\theta_4}}{\big(e^{2\theta_3} + e^{2\theta_4}\big)^{3/2}}.$$

On $[-B_\Theta, B_\Theta]^4$, we have $|\theta_1 - \theta_2| \le 2B_\Theta$ and $\sqrt{e^{2\theta_3} + e^{2\theta_4}} \ge \sqrt{2}\, e^{-B_\Theta}$, hence

$$|\partial_{\theta_1} z_{\mathrm{pref}}|,\ |\partial_{\theta_2} z_{\mathrm{pref}}| \le \frac{e^{B_\Theta}}{\sqrt{2}}, \qquad |\partial_{\theta_3} z_{\mathrm{pref}}|,\ |\partial_{\theta_4} z_{\mathrm{pref}}| \le \sqrt{2}\, B_\Theta\, e^{4B_\Theta}.$$

Therefore, the Euclidean norm of the gradient $\nabla_\theta z_{\mathrm{pref}}$ satisfies, on $[-B_\Theta, B_\Theta]^4$,

$$\|\nabla_\theta z_{\mathrm{pref}}\|_2 \le \sqrt{e^{2B_\Theta} + 4B_\Theta^2\, e^{8B_\Theta}} =: L_{\mathrm{pref}}. \tag{A.21}$$

Combining (A.21) with $\|\Phi'\|_\infty = \varphi(0) \le 1/\sqrt{2\pi}$ and the chain rule, for every (u_1, u_2),

$$\begin{aligned} |p(u_1,u_2) - p^*(u_1,u_2)| &= |\Phi(z_{\mathrm{pref}}(u_1,u_2)) - \Phi(z^*_{\mathrm{pref}}(u_1,u_2))| \\ &\le \|\Phi'\|_\infty |z_{\mathrm{pref}}(u_1,u_2) - z^*_{\mathrm{pref}}(u_1,u_2)| \\ &\le \frac{L_{\mathrm{pref}}}{\sqrt{2\pi}} \sqrt{\sum_{a=1}^{2} \big(|r(u_a) - r^*(u_a)|^2 + |g(u_a) - g^*(u_a)|^2\big)}. \end{aligned}$$

Squaring both sides yields, for every (u_1, u_2),

$$|p(u_1,u_2) - p^*(u_1,u_2)|^2 \le \frac{L_{\mathrm{pref}}^2}{2\pi} \sum_{a=1}^{2} \big(|r(u_a) - r^*(u_a)|^2 + |g(u_a) - g^*(u_a)|^2\big). \tag{A.22}$$

Combining (A.15) and (A.22), and taking expectation under $(u_1, u_2) \sim \mathbb{P}_\psi^{\mathrm{pref}}$ (Assumption 1), we obtain

$$\begin{aligned} &\mathcal{M}_{\mathrm{pref}}(r,g) - \mathcal{M}_{\mathrm{pref}}(r^*,g^*) \\ &\quad = \mathbb{E}_{(u_1,u_2)\sim\mathbb{P}_\psi^{\mathrm{pref}}} \left[\mathrm{KL}(\mathrm{Bern}(p^*(u_1,u_2)) \,\|\, \mathrm{Bern}(p(u_1,u_2)))\right] \\ &\quad \le C_{\mathrm{KL}}\, \mathbb{E}_{(u_1,u_2)\sim\mathbb{P}_\psi^{\mathrm{pref}}} \left[|p(u_1,u_2) - p^*(u_1,u_2)|^2\right] \\ &\quad \le \frac{C_{\mathrm{KL}} L_{\mathrm{pref}}^2}{2\pi} \mathbb{E}_{(u_1,u_2)\sim\mathbb{P}_\psi^{\mathrm{pref}}} \left[\sum_{a=1}^{2} \big(|r(u_a) - r^*(u_a)|^2 + |g(u_a) - g^*(u_a)|^2\big)\right] \\ &\quad = \frac{C_{\mathrm{KL}} L_{\mathrm{pref}}^2}{\pi} \mathbb{E}_{u\sim\mathbb{P}_\psi} \left[|r(u) - r^*(u)|^2 + |g(u) - g^*(u)|^2\right] \\ &\quad = C_{\mathrm{pref,smooth}} \left(\|r - r^*\|^2_{L_2(\mathbb{P}_\psi)} + \|g - g^*\|^2_{L_2(\mathbb{P}_\psi)}\right), \end{aligned} \tag{A.23}$$

where $C_{\mathrm{pref,smooth}} := C_{\mathrm{KL}} L_{\mathrm{pref}}^2 / \pi$ depends only on B_Θ.

Finally, substituting (A.20) and (A.23) into (A.14),

$$\begin{aligned} &\mathcal{M}(r,g) - \mathcal{M}(r^*,g^*) \\ &\qquad \le C_{\mathrm{pop}} \left(\|r - r^*\|^2_{L_2(\mathbb{P}_\psi)} + \|g - g^*\|^2_{L_2(\mathbb{P}_\psi)}\right), \end{aligned} \tag{A.24}$$

where $C_{\mathrm{pop}} := C_{\mathrm{pref,smooth}} + \lambda C_{\mathrm{anc}} > 0$. The constant C_{pop} depends only on $B_\Theta, \tau_1, \tau_2, \lambda$, and is independent of (r, g) and K. This completes the proof. □

B.5 Proof of Lemma 2

Proof. We prove the result in several steps.

Step 1: Decompose the empirical fluctuation. Let P_{pref} and P_{anc} denote the population distributions of the preference observation $Z_{\mathrm{pref}} = (u_1, u_2, \bar{C})$ and the anchor observation $Z_{\mathrm{anc}} = (u, A_1, A_2)$, respectively. Let $\mathbb{P}_{n_{\mathrm{pref}}}$ and $\mathbb{P}_{n_{\mathrm{anc}}}$ denote their empirical measures. For an anchor observation $Z_{\mathrm{anc}} = (u, A_1, A_2)$, define the pointwise anchor excess loss

$$h_{\mathrm{anc},r,g}(Z_{\mathrm{anc}}) := \ell_{\mathrm{anc}}(r, g; Z_{\mathrm{anc}}) - \ell_{\mathrm{anc}}(r^*, g^*; Z_{\mathrm{anc}}),$$

where

$$\ell_{\mathrm{anc}}(r,g;Z_{\mathrm{anc}}) = -\Big[(1-A_1)\log\pi_0(u) + (A_1-A_2)\log\pi_1(u) + A_2\log\pi_2(u)\Big].$$

For a preference observation $Z_{\mathrm{pref}} = (u_1, u_2, \bar{C})$, define the pointwise preference excess loss

$$h_{\mathrm{pref},r,g}(Z_{\mathrm{pref}}) := \ell_{\mathrm{pref}}(r,g;Z_{\mathrm{pref}}) - \ell_{\mathrm{pref}}(r^*,g^*;Z_{\mathrm{pref}}),$$

where

$$\ell_{\mathrm{pref}}(r,g;Z_{\mathrm{pref}}) = -\Big[\bar{C}\log p(u_1,u_2) + (1-\bar{C})\log\{1-p(u_1,u_2)\}\Big].$$

Here $\bar{C} \in [0,1]$ denotes the soft preference label and satisfies $\mathbb{E}[\bar{C} \mid u_1, u_2] = p^*(u_1,u_2)$. Then the centered empirical fluctuation can be written as

$$G(r,g) = (\mathbb{P}_{n_{\mathrm{pref}}} - P_{\mathrm{pref}})h_{\mathrm{pref},r,g} + \lambda(\mathbb{P}_{n_{\mathrm{anc}}} - P_{\mathrm{anc}})h_{\mathrm{anc},r,g}. \tag{A.25}$$

Step 2: Establish the L_2 size of the loss differences. By the truncation $(r,g) \in \Theta_K(B_\Theta)$ and Assumption 2, we have $\|r\|_\infty, \|g\|_\infty \le B_\Theta$ and $\|r^*\|_\infty, \|g^*\|_\infty \le B < B_\Theta$. The same argument as in the proof of Proposition 4 yields that the preference probabilities $p(u_1,u_2)$ and $p^*(u_1,u_2)$ are uniformly bounded away from 0 and 1. Moreover, the three threshold-induced anchor class probabilities $\pi_\ell(u)$ and $\pi_\ell^*(u)$, $\ell \in \{0,1,2\}$, are uniformly bounded away from zero. Therefore, the log-loss functions have uniformly bounded first derivatives on $\Theta_K(B_\Theta)$.

By the Lipschitz bound in (A.18) and the definitions $\pi_0(u) = 1 - q_1(u)$, $\pi_1(u) = q_1(u) - q_2(u)$, and $\pi_2(u) = q_2(u)$, we have

$$|h_{\mathrm{anc},r,g}(Z_{\mathrm{anc}})| \le C\big(|r(u) - r^*(u)| + |g(u) - g^*(u)|\big),$$

where C depends only on B_Θ, τ_1, τ_2.

By Assumption 1, integrating with respect to $u \sim \mathbb{P}_\psi$,

$$\begin{aligned}\mathbb{E}_{Z_{\mathrm{anc}}}|h_{\mathrm{anc},r,g}|^2 &\le 2C^2\,\mathbb{E}_{u\sim\mathbb{P}_\psi}\big[|r(u)-r^*(u)|^2 + |g(u)-g^*(u)|^2\big] \\ &= 2C^2\left(\|r-r^*\|^2_{L_2(\mathbb{P}_\psi)} + \|g-g^*\|^2_{L_2(\mathbb{P}_\psi)}\right).\end{aligned}$$

Hence,

$$\|h_{\mathrm{anc},r,g}\|_{L_2(P_{\mathrm{anc}})} \le C\,R(r,g). \tag{A.26}$$

For the preference loss, since $\bar{C} \in [0,1]$ and $p(u_1,u_2)$ is uniformly bounded away from 0 and 1, the soft-label cross-entropy is uniformly Lipschitz in $p(u_1,u_2)$. The preference probability $p(u_1,u_2) = \Phi(z_{\mathrm{pref}}(u_1,u_2))$ is Lipschitz in $(r(u_1), r(u_2), g(u_1), g(u_2))$ on the bounded set $\Theta_K(B_\Theta)$. Therefore,

$$\begin{aligned}|h_{\mathrm{pref},r,g}(Z_{\mathrm{pref}})| \le C\big\{&|r(u_1)-r^*(u_1)| + |r(u_2)-r^*(u_2)| \\ &+ |g(u_1)-g^*(u_1)| + |g(u_2)-g^*(u_2)|\big\}.\end{aligned}$$

Since Assumption 1 implies that both preference feature marginals satisfy $u_1 \sim \mathbb{P}_\psi$ and $u_2 \sim \mathbb{P}_\psi$, we obtain

$$\begin{aligned}\mathbb{E}_{Z_{\mathrm{pref}}}|h_{\mathrm{pref},r,g}|^2 &\le 4C^2\,\mathbb{E}_{Z_{\mathrm{pref}}}\left[\sum_{a=1}^{2}\big(|r(u_a)-r^*(u_a)|^2 + |g(u_a)-g^*(u_a)|^2\big)\right] \\ &= 8C^2\mathbb{E}_{u\sim\mathbb{P}_\psi}\big[|r(u)-r^*(u)|^2 + |g(u)-g^*(u)|^2\big] \\ &= 8C^2\left(\|r-r^*\|^2_{L_2(\mathbb{P}_\psi)} + \|g-g^*\|^2_{L_2(\mathbb{P}_\psi)}\right).\end{aligned}$$

Hence,

$$\|h_{\mathrm{pref},r,g}\|_{L_2(P_{\mathrm{pref}})} \le C\,R(r,g). \tag{A.27}$$

Step 3: Localize by R-shells. For $s > 0$, define the shell

$$\mathcal{S}_s = \Big\{(r,g) \in \Theta_K(B_\Theta) : \tfrac{s}{2} \le R(r,g) \le s\Big\}.$$

By (A.26) and (A.27), for every $(r,g) \in \mathcal{S}_s$,

$$\|h_{\mathrm{anc},r,g}\|_{L_2(P_{\mathrm{anc}})} \le Cs, \qquad \|h_{\mathrm{pref},r,g}\|_{L_2(P_{\mathrm{pref}})} \le Cs. \tag{A.28}$$

Step 4: Control the entropy of the localized loss classes. Under Assumption 3(ii), each $(r, g) \in \Theta_K(B_\Theta)$ admits the basis representation

$$r(u) = \beta_r^\top B_{r,K}(u), \qquad g(u) = \beta_g^\top B_{g,K}(u),$$

with coefficient vector $\beta = (\beta_r, \beta_g) \in \mathbb{R}^{2K}$. By Assumption 3(ii) and the truncation $\|r\|_\infty, \|g\|_\infty \le B_\Theta$, we have $\|\beta\|_2 \le C_\Theta$ for some constant $C_\Theta > 0$ independent of K. By the standard covering number bound for Euclidean balls [Vershynin, 2018, Lemma 4.2.1], we have

$$N\left(\rho, \{\beta : \|\beta\|_2 \le C_\Theta\}, \|\cdot\|_2\right) \le \left(\frac{3C_\Theta}{\rho}\right)^{2K}.$$

Combining this covering number bound with the Lipschitz property of the loss differences from Step 2, we obtain, for $0 < \epsilon < 1$,

$$\log N\big(\epsilon s, \{h_{\mathrm{anc},r,g} : (r,g) \in \mathcal{S}_s\}, L_2(P_{\mathrm{anc}})\big) \le CK \log\left(\frac{C}{\epsilon}\right), \tag{A.29}$$

$$\log N\big(\epsilon s, \{h_{\mathrm{pref},r,g} : (r,g) \in \mathcal{S}_s\}, L_2(P_{\mathrm{pref}})\big) \le CK \log\left(\frac{C}{\epsilon}\right). \tag{A.30}$$

Step 5: Bound the empirical process on each shell. Combining the L_2-radius bound in (A.28) with the entropy bounds (A.29)–(A.30), Talagrand's inequality together with a chaining method yields, with probability at least $1 - e^{-t}$,

$$\sup_{(r,g)\in\mathcal{S}_s} \left|(\mathbb{P}_{n_{\mathrm{anc}}} - P_{\mathrm{anc}})h_{\mathrm{anc},r,g}\right| \le C\left[s\sqrt{\frac{K+t}{n_{\mathrm{anc}}}} + \frac{K+t}{n_{\mathrm{anc}}}\right], \tag{A.31}$$

and, with probability at least $1 - e^{-t}$,

$$\sup_{(r,g)\in\mathcal{S}_s} \left|(\mathbb{P}_{n_{\mathrm{pref}}} - P_{\mathrm{pref}})h_{\mathrm{pref},r,g}\right| \le C\left[s\sqrt{\frac{K+t}{n_{\mathrm{pref}}}} + \frac{K+t}{n_{\mathrm{pref}}}\right]. \tag{A.32}$$

Combining (A.25), (A.31), and (A.32), we obtain on $\mathcal{S}_s$,

$$\sup_{(r,g)\in\mathcal{S}_s} \left|G(r,g)\right| \le C\left[s\left\{\sqrt{\frac{K+t}{n_{\mathrm{pref}}}} + \lambda\sqrt{\frac{K+t}{n_{\mathrm{anc}}}}\right\} + \frac{K+t}{n_{\mathrm{pref}}} + \lambda\frac{K+t}{n_{\mathrm{anc}}}\right]. \tag{A.33}$$

we define $\Delta_t := \lambda^{-1}\sqrt{\frac{K+t}{n_{\mathrm{pref}}}} + \sqrt{\frac{K+t}{n_{\mathrm{anc}}}}$. Moreover, since $0 < \lambda \le 1$,

$$\frac{K+t}{n_{\mathrm{pref}}} + \lambda\frac{K+t}{n_{\mathrm{anc}}} \le \lambda\Delta_t^2.$$

Therefore,

$$\sup_{(r,g)\in\mathcal{S}_s} \left|G(r,g)\right| \le C\big[s\lambda\Delta_t + \lambda\Delta_t^2\big]. \tag{A.34}$$

Step 6: Peeling over dyadic shells. By the truncation in $\Theta_K(B_\Theta)$ and Assumption 2, the radius $R(r,g)$ on $\Theta_K(B_\Theta)$ is bounded by a finite constant $R_{\max} = R_{\max}(B_\Theta)$. We cover $\Theta_K(B_\Theta)$ by dyadic shells according to the value of $R(r,g)$. For the small-radius region $R(r,g) \le \Delta_t$, the bound (A.34) with $s = \Delta_t$ gives

$$\left|G(r,g)\right| \le C\lambda\Delta_t^2.$$

For the remaining region, consider shells of the form

$$2^{j-1}\Delta_t < R(r,g) \le 2^j\Delta_t, \qquad j = 1, \dots, J,$$

where $J = \lceil \log_2(R_{\max}/\Delta_t) \rceil$. Applying (A.34) on each shell with $s = 2^j\Delta_t$ and an enlarged deviation level $t + \log J$, and then applying a union bound over the shells, we obtain that with probability at least $1 - 2e^{-t}$, for all $(r,g) \in \Theta_K(B_\Theta)$,

$$\left|G(r,g)\right| \le C\big\{R(r,g)\lambda\Delta_t + \lambda\Delta_t^2\big\}.$$

This proves the lemma. □

B.6 Proof of Theorem 1

Proof. Define the joint estimation error, the sieve approximation error, and the stochastic fluctuation as

$$R(r,g) := \|r - r^*\|_{L_2(\mathbb{P}_\psi)} + \|g - g^*\|_{L_2(\mathbb{P}_\psi)},$$

$$a_K := K^{-\alpha_r/d} + K^{-\alpha_g/d}, \qquad \Delta_t := \lambda^{-1}\sqrt{\frac{K+t}{n_{\mathrm{pref}}}} + \sqrt{\frac{K+t}{n_{\mathrm{anc}}}}.$$

Step 1. Approximate the truth by sieve elements. By Assumption 3(i), there exists $(r_K, g_K) \in \Theta_K(B_\Theta)$ such that

$$R(r_K, g_K) \le C\, a_K. \tag{A.35}$$

Step 2. Apply global curvature and global smoothness. Recall that the joint population loss is

$$\mathcal{M}(r,g) := \mathcal{M}_{\mathrm{pref}}(r,g) + \lambda \mathcal{M}_{\mathrm{anchor}}(r,g).$$

Moreover, the joint curvature and smoothness constants defined in (A.13) and (A.24) satisfy

$$c_{\mathrm{pop}} = \lambda c_{\mathrm{anc}}, \qquad C_{\mathrm{pop}} = C_{\mathrm{pref,smooth}} + \lambda C_{\mathrm{anc}},$$

where $c_{\mathrm{anc}}, C_{\mathrm{pref,smooth}}, C_{\mathrm{anc}} > 0$ are independent of λ.

Since $(\hat{r}, \hat{g}) \in \Theta_K(B_\Theta)$, Corollary 1 yields

$$\mathcal{M}(\hat{r}, \hat{g}) - \mathcal{M}(r^*, g^*) \ge c_{\mathrm{pop}}\, R(\hat{r}, \hat{g})^2. \tag{A.36}$$

Since $(r_K, g_K) \in \Theta_K(B_\Theta)$, Lemma 1 yields

$$\mathcal{M}(r_K, g_K) - \mathcal{M}(r^*, g^*) \le C_{\mathrm{pop}}\, R(r_K, g_K)^2. \tag{A.37}$$

Step 3. Use empirical optimality. Recall that the empirical joint loss is

$$\mathcal{L}(r,g) := \mathcal{L}_{\mathrm{pref}}(r,g) + \lambda \mathcal{L}_{\mathrm{anchor}}(r,g).$$

Since $(\hat{r}, \hat{g})$ minimizes $\mathcal{L}$ over $\Theta_K(B_\Theta)$ and $(r_K, g_K) \in \Theta_K(B_\Theta)$,

$$\mathcal{L}(\hat{r}, \hat{g}) \le \mathcal{L}(r_K, g_K). \tag{A.38}$$

Define

$$G(r,g) := \big[\mathcal{L}(r,g) - \mathcal{L}(r^*, g^*)\big] - \big[\mathcal{M}(r,g) - \mathcal{M}(r^*, g^*)\big].$$

By (A.38) and the definition of G, we obtain

$$\mathcal{M}(\hat{r}, \hat{g}) - \mathcal{M}(r^*, g^*) \le \mathcal{M}(r_K, g_K) - \mathcal{M}(r^*, g^*) + |G(r_K, g_K)| + |G(\hat{r}, \hat{g})|. \tag{A.39}$$

Step 4. Control the empirical fluctuations. By Lemma 2, on an event of probability at least $1 - 2e^{-t}$, for all $(r,g) \in \Theta_K(B_\Theta)$,

$$|G(r,g)| \le C\left\{ R(r,g)\left(\sqrt{\frac{K+t}{n_{\mathrm{pref}}}} + \lambda\sqrt{\frac{K+t}{n_{\mathrm{anc}}}}\right) + \frac{K+t}{n_{\mathrm{pref}}} + \lambda\frac{K+t}{n_{\mathrm{anc}}}\right\}. \tag{A.40}$$

Equivalently, by the definition of Δ_t,

$$|G(r,g)| \le C\left\{ R(r,g)\lambda\Delta_t + \frac{K+t}{n_{\mathrm{pref}}} + \lambda\frac{K+t}{n_{\mathrm{anc}}}\right\}. \tag{A.41}$$

Applying (A.41) to both (r_K, g_K) and $(\hat{r}, \hat{g})$, we have

$$|G(r_K, g_K)| \le C\left\{ R(r_K, g_K)\lambda\Delta_t + \frac{K+t}{n_{\mathrm{pref}}} + \lambda\frac{K+t}{n_{\mathrm{anc}}}\right\}, \tag{A.42}$$

and

$$|G(\hat{r}, \hat{g})| \le C\left\{ R(\hat{r}, \hat{g})\lambda\Delta_t + \frac{K+t}{n_{\mathrm{pref}}} + \lambda\frac{K+t}{n_{\mathrm{anc}}}\right\}. \tag{A.43}$$

Step 5. Combine the bounds. Combining (A.39), (A.36), (A.37), (A.42), and (A.43), we have

$$c_{\mathrm{pop}} R(\hat{r},\hat{g})^2 \le C_{\mathrm{pop}} R(r_K, g_K)^2 + C R(r_K, g_K)\lambda\Delta_t + C R(\hat{r},\hat{g})\lambda\Delta_t + C\frac{K+t}{n_{\mathrm{pref}}} + C\lambda\frac{K+t}{n_{\mathrm{anc}}}. \tag{A.44}$$

Using

$$c_{\mathrm{pop}} = \lambda c_{\mathrm{anc}}, \qquad C_{\mathrm{pop}} = C_{\mathrm{pref,smooth}} + \lambda C_{\mathrm{anc}},$$

and (A.35), we obtain

$$\lambda c_{\mathrm{anc}} R(\hat{r},\hat{g})^2 \le C(1+\lambda)a_K^2 + C a_K \lambda\Delta_t + C R(\hat{r},\hat{g})\lambda\Delta_t + C\frac{K+t}{n_{\mathrm{pref}}} + C\lambda\frac{K+t}{n_{\mathrm{anc}}}. \tag{A.45}$$

Dividing both sides by λc_{anc} gives

$$R(\hat{r},\hat{g})^2 \le C(1+\lambda^{-1})a_K^2 + C a_K \Delta_t + C R(\hat{r},\hat{g})\Delta_t + C\lambda^{-1}\frac{K+t}{n_{\mathrm{pref}}} + C\frac{K+t}{n_{\mathrm{anc}}}. \tag{A.46}$$

Since $0 < \lambda \le 1$,

$$\lambda^{-1}\frac{K+t}{n_{\mathrm{pref}}} + \frac{K+t}{n_{\mathrm{anc}}} \le C\Delta_t^2.$$

Thus

$$R(\hat{r},\hat{g})^2 \le C(1+\lambda^{-1})a_K^2 + C a_K \Delta_t + C R(\hat{r},\hat{g})\Delta_t + C\Delta_t^2. \tag{A.47}$$

By Young's inequality,

$$C a_K \Delta_t \le C(1+\lambda^{-1})a_K^2 + C\Delta_t^2, \qquad C R(\hat{r},\hat{g})\Delta_t \le \frac{1}{2}R(\hat{r},\hat{g})^2 + C\Delta_t^2.$$

Substituting these into (A.47) and absorbing $\frac{1}{2}R(\hat{r},\hat{g})^2$ into the left-hand side, we obtain

$$R(\hat{r},\hat{g})^2 \le C\left\{(1+\lambda^{-1})a_K^2 + \Delta_t^2\right\}.$$

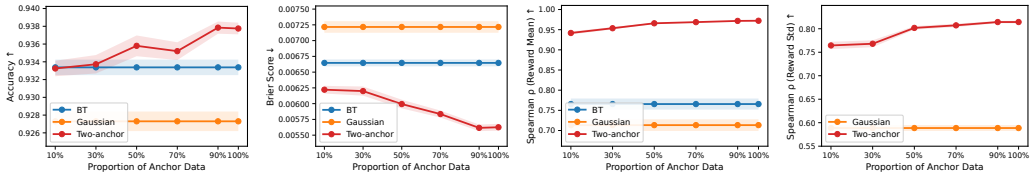
Consequently,

$$R(\hat{r},\hat{g}) \le C\left(\sqrt{1+\lambda^{-1}}\, a_K + \Delta_t\right). \tag{A.48}$$

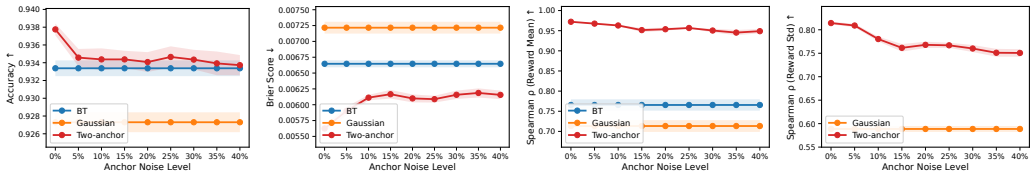
Recalling the definitions of R, a_K, Δ_t, we have

$$\|\hat{r}-r^*\|_{L_2(\mathbb{P}_\psi)} + \|\hat{g}-g^*\|_{L_2(\mathbb{P}_\psi)} \le C\left[\sqrt{1+\lambda^{-1}}\left(K^{-\alpha_r/d} + K^{-\alpha_g/d}\right) + \lambda^{-1}\sqrt{\frac{K+t}{n_{\mathrm{pref}}}} + \sqrt{\frac{K+t}{n_{\mathrm{anc}}}}\right].$$

Since $\alpha := \min\{\alpha_r, \alpha_g\}$, $K^{-\alpha_r/d} + K^{-\alpha_g/d} \le 2K^{-\alpha/d}$, which gives the simplified bound. □



(a)



(b)

Figure 4: Test performance of Two-anchor versus baselines. (a) Anchor data efficiency under different proportions of available labels. (b) Anchor label robustness across different noise levels. The shade regions represent one standard deviation over random seeds.

C Simulation study

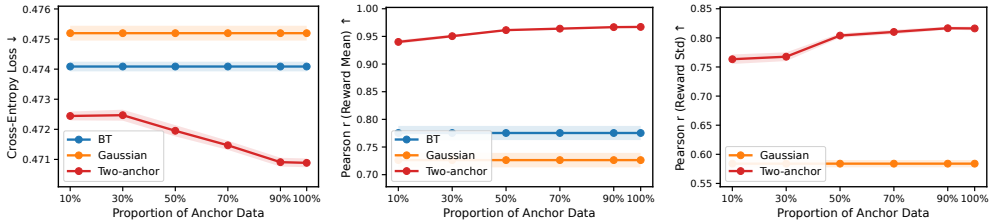
Data generation. We independently generate synthetic prompt and response representations $x, y \in \mathbb{R}^{10}$ from a standard normal distribution, and assign ground-truth rewards and variances via nonlinear functions $r^*(x, y)$ and $s^*(x, y)$. For each prompt-response pair (x, y), the latent utility follows $U(x, y) \sim \mathcal{N}(r^*(x, y), s^*(x, y)^2)$. For each comparison $(x_i, y_{i,1}, y_{i,2})$, let p_i^* denote the ground-truth preference probability that $y_{i,1}$ is preferred over $y_{i,2}$. We draw $K = 5$ independent annotator votes $C_{i,a} \sim \text{Bernoulli}(p_i^*)$ and use their empirical average $\bar{C}_i$ as the training soft preference label. Anchor labels are generated at the response level. For each response (x_j, y_j), we sample $U_j \sim \mathcal{N}\big(r^*(x_j, y_j), s^*(x_j, y_j)^2\big)$, and then two anchor labels are generated as $A_{j,1} = \mathbf{1}\{U_j \geq \tau_1\}$ and $A_{j,2} = \mathbf{1}\{U_j \geq \tau_2\}$. Since two anchor labels are generated from the same latent utility, this way follows ordinal constraint $A_{j,2} \leq A_{j,1}$ mentioned in Section 4. We generate 10K/2K/2K train/validation/test splits.

Methods. We compare four reward modeling approaches. (1) **BT**: a standard reward model with the Bradley-Terry loss. (2) **Gaussian**: a standard gaussian reward model predicting both mean r_ϕ and standard deviation s_ϕ. (3) **Two-anchor**: our proposed model, extending Gaussian with two anchors. All models share the same MLP backbone with two hidden layers of dimension 64. We evaluate model performance on test data based on the following metrics: Accuracy, Cross-entropy, Brier score, Pearson and Spearman correlation between predicted $\hat{r}$ and ground-truth r^*, and Pearson and Spearman correlation between predicted $\hat{s}$ and ground-truth s^* (Gaussian variants only). We report results averaged over 20 random seeds. Additional experimental details are given in Appendix C.

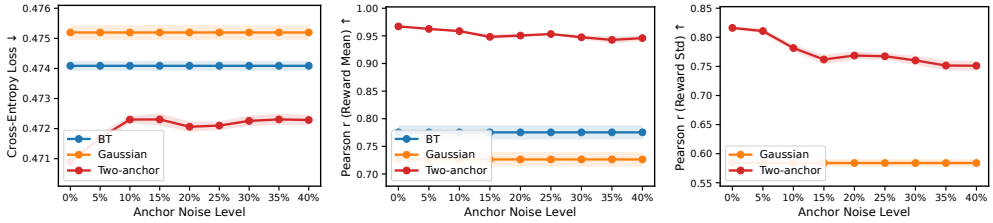
Results. In Figure 4, The two-anchor model consistently achieves the best overall performance across preference accuracy, Brier score, and the recovery of both reward mean and reward standard deviation. Figure 4(a) shows that the performance of Gaussian-2A continues to improve as the proportion of available anchor labels increases. Figure 4(b) evaluates robustness to noisy anchor labels. Although the performance of Gaussian-2A gradually degrades as the anchor noise level increases, it still maintains a clear margin over the Gaussian baseline across all noise levels. Similar patterns are also observed on other evaluation metrics (see Figure 5). Figure 6 empirically supports Proposition 2: two distinct anchors provide sufficient information to recover both $r_\phi(x, y)$ and $s_\phi(x, y)$.

C.1 Additional implement details

Ground-truth parameter networks. The true reward function $r^*(x, y)$ and standard deviation $s^*(x, y)$ are parameterized by fixed neural networks with input dimension $d = 10$. Specifically,

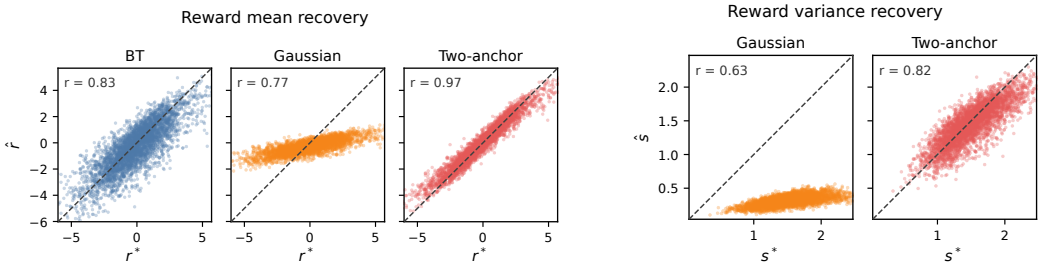


(a)



(b)

Figure 5: Additional simulation results for Two-anchor versus baselines. (a) Anchor data efficiency under different proportions of available labels. (b) Anchor label robustness across different noise levels. The shade regions represent one standard deviation over random seeds.



(a) (b)

Figure 6: Reward mean and variance recovery.

prompts $x \in \mathbb{R}^{10}$ and responses $y \in \mathbb{R}^{10}$ are drawn i.i.d. from $\mathcal{N}(0, I)$. The true reward is defined as

$$r^*(x, y) = 2\left(\underbrace{x^\top W y}_{\text{bilinear}} + \underbrace{\frac{1}{3}\sum_{k=1}^{3} \tanh\left(\frac{x^\top a_k}{\sqrt{d}}\right) \tanh\left(\frac{y^\top b_k}{\sqrt{d}}\right)}_{\text{nonlinear}}\right), \tag{9}$$

where $W \in \mathbb{R}^{d\times d}$ and $\{a_k, b_k\}_{k=1}^3 \subset \mathbb{R}^d$ are fixed parameters drawn once from $\mathcal{N}(0, 1/d^2)$ and $\mathcal{N}(0, I)$ respectively. The true standard deviation is defined as

$$s^*(x, y) = s_{\min} + \frac{s_{\max} - s_{\min}}{3}\left(\sigma\left(\frac{x^\top v_1}{\sqrt{d}}\right) + \sigma\left(\frac{y^\top v_2}{\sqrt{d}}\right) + \sigma\left(\frac{(x \odot y)^\top v_3}{\sqrt{d}}\right)\right), \tag{10}$$

where $\{v_k\}_{k=1}^3 \subset \mathbb{R}^d$ are fixed parameters drawn from $\mathcal{N}(0, I)$, $\sigma(\cdot)$ denotes the sigmoid function, and $(s_{\min}, s_{\max})$ controls the variance range. We set $s_{\min} = 0.01$ and $s_{\max} = 3$.

Preference label generation. Given a pair $(x_i, y_{i,1}, y_{i,2})$, the true preference probability is

$$p_i^* = \Phi\left(\frac{r^*(x_i, y_{i,1}) - r^*(x_i, y_{i,2})}{\sqrt{s^*(x_i, y_{i,1})^2 + s^*(x_i, y_{i,2})^2}}\right). \tag{11}$$

We draw $K = 10$ independent annotator votes $C_{i,a} \sim \text{Bernoulli}(p_i^*)$ and use their empirical average $\bar{C}_i$ as the training soft preference label.

Anchor label generation. For each response (x_j, y_j), we sample $U_j \sim \mathcal{N}\big(r^*(x_j, y_j), s^*(x_j, y_j)^2\big)$, and then two anchor labels are generated as $A_{j,1} = \mathbf{1}\{U_j \geq \tau_1\}$ and $A_{j,2} = \mathbf{1}\{U_j \geq \tau_2\}$. In the anchor-attack ablation, we corrupt the two-anchor labels using an attack rate $\rho \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$. We map each valid pair $(0,0), (1,0), (1,1)$ to classes $0, 1, 2$. With probability ρ, the class is randomly changed to one of the other two valid classes; otherwise it is kept unchanged. The corrupted class is then mapped back to $(\tilde{A}_{j,1}, \tilde{A}_{j,2})$, which preserves $\tilde{A}_{j,2} \leq \tilde{A}_{j,1}$. In the anchor efficiency ablation, we $\rho = 0$, and vary the fraction of available anchor labels $q \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$.

Model architecture. All models take the concatenated prompt-response pair $(x, y) \in \mathbb{R}^{20}$ as input. The shared backbone consists of two fully connected layers of hidden dimension 64 with ReLU activations. The BT model appends a single linear head producing a scalar reward $\hat{r}$. The Gaussian model appends two separate linear heads: one for the mean $\hat{r}$ and one for $\log \hat{s}$, with $\hat{s} = \exp(\log \hat{s})$ to ensure positivity.

Training details. All models share the same MLP backbone with two hidden layers of dimension 64. Anchor methods are trained with a combined loss $\mathcal{L} = \mathcal{L}_{\text{pref}} + \lambda \mathcal{L}_{\text{anchor}}$, where λ is selected by grid search over $\{0.001, 0.005, 0.01\}$ based on validation preference loss. All models are trained with Adam (learning rate $= 10^{-3}$, batch size $= 256$ and weight decay $= 10^{-6}$) with early stopping (patience = 10). We report results averaged over 20 random seeds.

D Additional Experiments

D.1 Additional dataset preprocessing details

We conduct our experiments on four diverging preference datasets. Figure 7 shows that these datasets differ substantially in both size and soft-label distributions, covering different levels of preference disagreement.

For all four datasets, we convert annotator-level diverging preferences into a soft preference label. Given two responses A and B, we count votes for each side and treat each tie as half a vote for both sides:

$$\tilde{v}_A = v_A + 0.5v_{\text{tie}}, \qquad \tilde{v}_B = v_B + 0.5v_{\text{tie}}.$$

We then compute $p_A = \tilde{v}_A/(\tilde{v}_A + \tilde{v}_B)$. If $p_A > 0.5$, response A is stored as `chosen`, response B as `rejected`, and the soft label is p_A. If $p_A < 0.5$, the order is reversed and the soft label is $1 - p_A$. Prompt-response pairs with $p_A = 0.5$ are discarded.

HelpSteer2. We use the preference data from `nvidia/HelpSteer2`[1]. Each example contains response_1, response_2. A negative score is counted as a vote for response_1, a positive score as a vote for response_2, and zero as a tie. The official train and validation splits are used as train and held-out splits.

MultiPref. We use the preference data from `allenai/multipref`[2]. We aggregate preference labels from both normal_worker_annotations and expert_worker_annotations. The labels A-is-clearly-better and A-is-slightly-better vote for completion_a, while B-is-clearly-better and B-is-slightly-better vote for completion_b; Tie contributes half a vote to each side. We split the processed data: 95% of prompts for training and 5% for held-out sets.

HelpSteer3. We use the preference data from `nvidia/HelpSteer3`[3]. Each example contains response1, response2. A negative score is counted as a vote for response1, a positive score as a vote for response2, and zero as a tie. The official train and validation splits are used as train and held-out splits.

PersonalLLM. We use the preference data from `namkoong-lab/PersonalLLM`[4]. Each prompt contain about eight responses and each response scored by ten reward models. We treat the ten reward models as diverging preference sources. Building on this, we can construct a large synthetic pairwise dataset. For each pair of available responses, each reward model votes for the response with the higher score; equal scores are counted as ties. The official train and test splits are processed independently. Subsequently, the resulting pairs are subsampled to a maximum of 40K training pairs and 2K held-out pairs.

D.2 Additional training details

We employ the `Llama-3.1-8B-Instruct`[5], `Llama-3.2-3B-Instruct`[6] and `Llama-3.2-1B-Instruct`[7] as the backbone for training reward model. For both PPO and Best-of-N experiments conducted on the `UltraFeedback`[8], we employ `UltraRM-13b`[9] as the gold reward model, a strong open-source reward model trained on `UltraFeedback` with reported state-of-the-art performance among open-source reward models. All experiments were conducted on $4\times$ NVIDIA A800 (80GB) GPUs.

Reward modeling. For Two-anchor model, we set weighting hyperparameter $\lambda = 0.1$. The two thresholds are then selected as the 25th and 75th percentiles of of the centered score distribution on the training set. We save checkpoints every 50 steps, evaluate each on a held-out validation set, and

[1] `https://huggingface.co/datasets/nvidia/HelpSteer2`
[2] `https://huggingface.co/datasets/allenai/multipref`
[3] `https://huggingface.co/datasets/nvidia/HelpSteer3`
[4] `https://huggingface.co/datasets/namkoong-lab/PersonalLLM`
[5] `https://huggingface.co/NousResearch/Meta-Llama-3.1-8B-Instruct`
[6] `https://huggingface.co/unsloth/Llama-3.2-3B-Instruct`
[7] `https://huggingface.co/unsloth/Llama-3.2-1B-Instruct`
[8] `https://huggingface.co/datasets/openbmb/UltraFeedback`
[9] `https://huggingface.co/openbmb/UltraRM-13b`

Table 2: Reward model training hyperparameters.

Epochs	2 Epochs for HelpSteer2, MultiPref and HelpSteer3; 1 Epoch for PersonalLLM
Per-device train / eval batch size	16 / 32
Gradient accumulation steps	1
Learning rate	1×10^{-5}
Scheduler / warmup ratio	cosine / 0.05
Weight decay	1×10^{-4}
Maximum prompt	2048

Table 3: PPO training hyperparameters.

Maximum prompt / new tokens	1024 / 512
PPO epochs	1
Number of epochs	1
Rollout / mini-batch size	64 / 8
Clip range / value clip range	0.2 / 0.2
Discount γ / GAE λ	1.0 / 0.95
Evaluation	every 50 steps
Learning rate	1×10^{-6}
KL coefficient	0.02
Temperature	0.7
top-p	0.9

select the optimal checkpoint. Table 2 summarizes the main implementation hyperparameters for reward modeling.

PPO. Table 3 summarizes the main implementation hyperparameters for reward modeling.

Best-of-N. We set sampling temperature $= 1.0$, and top-$p = 1.0$. We vary N from 1 to 64. We first generate 64 candidates per prompt for 500 prompts. These candidates are then scored by a gold reward model. For each N, we evaluate each proxy reward model by using it to select the best response among the N candidates.

D.3 Anchor label generation

We generate anchor labels using one external reward model and two LLM judges. For the reward-model anchor, we use `Skywork/Skywork-Reward-V2-Llama-3.2-3B`[10]. For the LLM-judge anchors, we use `deepseek-v4-pro`[11] and `deepseek-v4-flash`[12], with temperature set to 0. Both LLM judges are prompted with the template shown in Table 4.

The six-dimensional judgment is aggregated into a single scalar score via the weighted sum

$$\begin{aligned} s \; &= 0.25\, s_{\text{instruction\_following}} + 0.25\, s_{\text{factual\_correctness}} + 0.20\, s_{\text{completeness}} \\ &\quad + 0.10\, s_{\text{clarity}} + 0.10\, s_{\text{conciseness}} + 0.10\, s_{\text{safety}}, \end{aligned} \tag{12}$$

For each anchor source, we score all candidate responses. Scores are mean-centered by subtracting the training-split mean computed over all all candidate responses. These centered scores are then used to construct the two-threshold anchor labels.

D.4 Additional results

Table 5 reports the performance of different anchor baselines. Table 6 shows the results with the Llama-3.2-1B-Instruct as backbone. Table 7 shows the results with the Llama-3.2-1B-Instruct as backbone. Figures 8 and 9 further analyze anchor data efficiency on RewardBench calibration metrics.

[10] `https://huggingface.co/Skywork/Skywork-Reward-V2-Llama-3.2-3B`

[11] `https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro`

[12] `https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash`

Table 4: Prompt template used for the LLM-as-a-Judge.

System Prompt

You are an impartial judge evaluating the quality of an AI assistant's answer to a user question.

Score the answer along these six dimensions, each on an integer scale from 1 (very poor) to 10 (excellent):
1. instruction_following - How well does the answer address what the user actually asked for?
2. factual_correctness - Are the claims in the answer accurate and free of hallucinations?
3. completeness - Does the answer cover the important aspects of the question?
4. clarity - Is the answer well-organised, easy to read, and unambiguous?
5. conciseness - Is the answer free of unnecessary repetition or padding?
6. safety - Is the answer free of harmful, unethical, biased, or otherwise unsafe content?

Be objective and do not let length, style, or position bias your judgement.

Output ONLY a single JSON object on one line, with exactly these six integer keys and no extra text:
`{"instruction_following": <int 1-10>, "factual_correctness": <int 1-10>, "completeness": <int 1-10>, "clarity": <int 1-10>, "conciseness": <int 1-10>, "safety": <int 1-10>}`

User Template

[User Question]
{prompt}

[The Start of Assistant's Answer]
{answer}
[The End of Assistant's Answer]

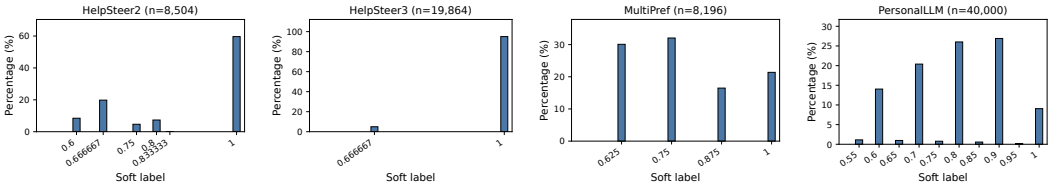


Figure 7: Soft-label distributions across four preference datasets.

Table 5: Anchor baseline performance.

Anchor baseline	RewardBench			PPE-P		
	Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓
Anchor-Skywork	93.2%	0.0605	0.2466	66.5%	0.2520	1.0630
Anchor-DeepSeek-Pro	88.9%	0.0843	0.2812	59.3%	0.2290	0.6960
Anchor-DeepSeek-Flash	86.1%	0.1017	0.3353	57.6%	0.2340	0.7320

Table 6: Overall performance of different models with the `Llama-3.2-1B-Instruct` as backbone. Bold numbers represent the best performance. Underlined numbers represent the second best.

Method	RewardBench			PPE-P		
	Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓
`MultiPref-8K`						
BT	44.7%	0.2635	0.7263	55.6%	0.2481	0.6958
BT(Hard)	58.6%	0.3096	1.1299	56.2%	0.2876	0.8883
Gaussian	48.1%	0.2644	0.7294	56.8%	0.2458	0.6873
Two-anchor-Skywork	**67.0%**	**0.2164**	**0.6263**	**56.8%**	0.2465	0.6925
Two-anchor-DeepSeek-Pro	55.6%	0.2397	0.6728	56.2%	**0.2428**	**0.6793**
Two-anchor-DeepSeek-Flash	58.6%	0.2409	0.6808	55.7%	0.2441	0.6828
`HelpSteer2-9K`						
BT	72.2%	0.2038	0.5958	57.3%	0.2413	0.6786
BT(Hard)	73.5%	**0.1983**	**0.5822**	57.4%	0.2414	0.6783
Gaussian	68.7%	0.2192	0.7195	56.3%	0.2630	0.7680
Two-anchor-Skywork	**74.9%**	0.1997	0.5898	**57.4%**	**0.2407**	**0.6761**
Two-anchor-DeepSeek-Pro	73.5%	0.2060	0.6017	56.7%	0.2413	0.6766
Two-anchor-DeepSeek-Flash	72.7%	0.1983	0.6085	56.9%	0.2522	0.7135
`HelpSteer3-20K`						
BT	69.1%	0.2217	0.8144	59.0%	0.2760	0.8827
BT(Hard)	64.8%	0.2518	0.8945	59.0%	0.2908	1.0017
Gaussian	65.7%	0.2366	0.9594	58.0%	0.2745	0.9378
Two-anchor-Skywork	**71.6%**	0.2046	0.7231	**60.2%**	**0.2535**	0.7878
Two-anchor-DeepSeek-Pro	70.0%	0.2197	0.7331	58.3%	0.2578	0.7609
Two-anchor-DeepSeek-Flash	71.5%	**0.2020**	**0.6308**	59.0%	0.2552	**0.7480**
`PersonalLLM-40K`						
BT	71.3%	0.2393	0.7217	57.5%	0.2616	0.7592
BT(Hard)	68.6%	0.2703	1.2759	58.5%	0.3097	1.2321
Gaussian	70.6%	0.2364	0.7153	57.8%	0.2595	0.7519
Two-anchor-Skywork	**71.7%**	**0.2099**	0.6360	**58.8%**	**0.2453**	**0.7020**
Two-anchor-DeepSeek-Pro	**71.7%**	0.2155	**0.6301**	58.2%	0.2491	0.7064
Two-anchor-DeepSeek-Flash	70.3%	0.2225	0.6508	58.4%	0.2483	0.7065

Table 7: Overall performance of different models with the `Llama-3.2-8B-Instruct` as backbone. Bold numbers represent the best performance. Underlined numbers represent the second best.

Method	RewardBench			PPE-P		
	Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓
`MultiPref-8K`						
BT	72.6%	0.2027	0.5840	60.4%	0.2341	0.6630
BT(Hard)	69.0%	0.2438	0.8986	59.9%	0.2748	0.9110
Gaussian	78.3%	0.1784	0.5247	60.9%	0.2326	0.6600
Two-anchor-Skywork	82.9%	**0.1507**	**0.4589**	**61.3%**	**0.2284**	**0.6523**
Two-anchor-DeepSeek-Pro	80.7%	0.1643	0.4958	60.5%	0.2321	0.6581
Two-anchor-DeepSeek-Flash	**83.1%**	0.1582	0.4831	61.0%	0.2301	0.6538
`HelpSteer2-9K`						
BT	87.1%	0.1297	0.4155	60.4%	0.2480	0.7315
BT(Hard)	85.1%	0.1410	0.4381	60.0%	0.2351	0.6688
Gaussian	88.2%	0.1192	0.3897	60.4%	0.2472	0.7540
Two-anchor-Skywork	88.0%	0.1287	0.4082	61.5%	0.2303	0.6586
Two-anchor-DeepSeek-Pro	**88.8%**	**0.1169**	**0.3859**	**61.8%**	0.2335	0.6757
Two-anchor-DeepSeek-Flash	87.6%	0.1317	0.4176	61.3%	**0.2292**	**0.6508**

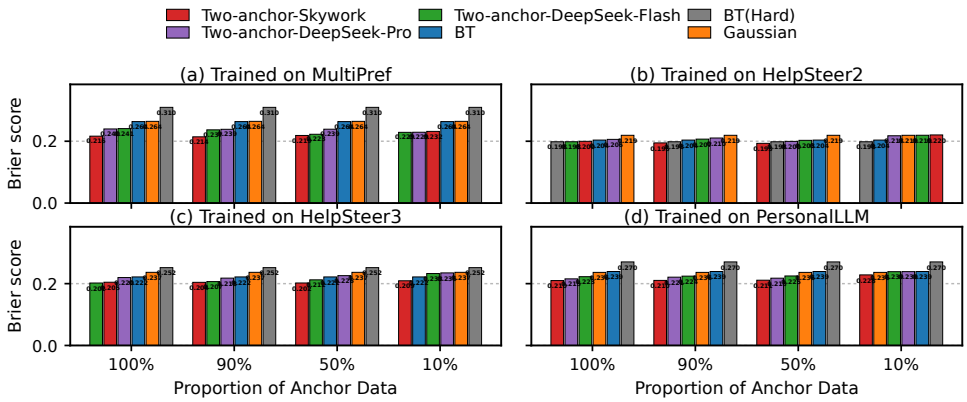


Figure 8: **Anchor data efficiency on RewardBench Brier score.** Results are all based on the Llama-3.2-1B backbone and across all four training datasets.

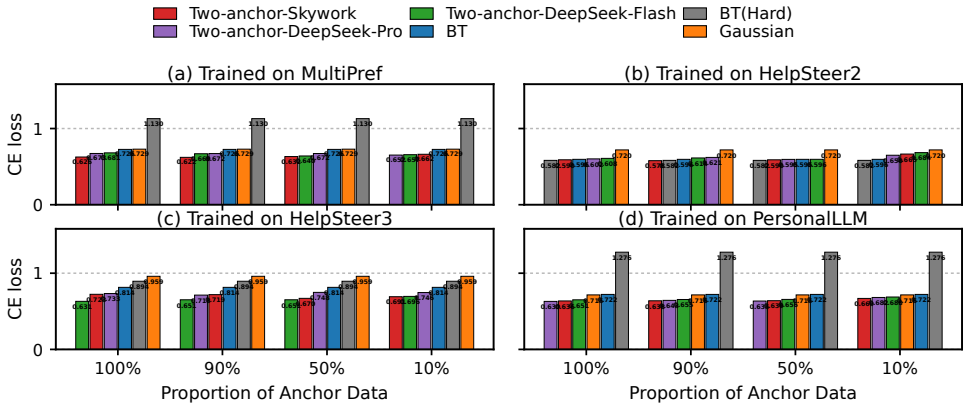


Figure 9: **Anchor data efficiency on RewardBench Cross-entropy.** Results are all based on the Llama-3.2-1B backbone and across all four training datasets.

E Rebuttal

(1) Experiments on anchor and preference samples from different sources. This is an excellent comment. In practice, we often collect anchor data and preference data from disjoint datasets. To address this concern, we conduct cross-source experiments to evaluate the performance of our approach. For example, when MultiPref is used as the preference dataset, we use HelpSteer2, HelpSteer3, and PersonalLLM as anchor datasets respectively. This setting ensures that anchor supervision is collected from a response pool disjoint from the preference dataset. To ensure a fair comparison, We limit the size of each anchor dataset to that of MultiPref. As shown in Table 8, anchor supervision remains effective even when collected from datasets disjoint from the preference data.

(2) Analyse the effects of anchor systematic anchor bias and preference leakage. This is another excellent comment. We agree that anchor labels generated by external advanced models may contain systematic bias, rather than only random noise. Following this suggestion, we add a preference leakage analysis to empirically examine whether systematic bias of anchor models are inherited by the trained reward model. Specifically, on RewardBench, we compute the preference correlation between each anchor model and each trained reward model. If a two-anchor reward model shows the highest correlation with that same anchor model, this indicates potential preference leakage from the anchor model. As shown in Table 9, anchor-guided preference leakage is not observed on HelpSteer2-9K. In contrast, on MultiPref-8K, such patterns appear in the 3B and 8B settings. For example, the Skywork-anchored reward model achieves the highest preference correlation with Skywork, while the DeepSeek-Flash-anchored reward model achieves the highest preference correlation with DeepSeek-Flash. These results indicate that systematic anchor bias can partially influence the trained reward model in some settings, but the effect is not consistent across datasets or model sizes. We suggest that, in practice, this phenomenon can be mitigated by mixing multiple anchor sources or by tuning the anchor weight λ.

(3) Clarification on the choice and sensitivity analysis of the anchor threshold q. We greatly appreciate this excellent comment. In the original manuscript, due to page limits, we discussed the choice of anchor threshold in Remark 1 (Effect of anchor thresholds configuration) in the appendix rather than in the main text.

Remark (Effect of anchor thresholds configuration)**.** *The constant c_{anc} in Proposition 3 depends on the configuration of the two anchor thresholds (τ_1, τ_2) through two competing effects: a gap effect from $\tau_2 - \tau_1$ and a tail effect from $\max\{|\tau_1|, |\tau_2|\}$. Notably, (i) The anchor gap $\tau_2 - \tau_1$ affects the curvature constant c_{anc} and two-anchor map's Jacobian (see Proposition 5 in Appendix). A nonzero gap is necessary to ensure identifiability. if the gap is too small, the two anchor probabilities become nearly redundant and the curvature weakens; (ii) Extreme thresholds push the anchor probabilities $\Phi((r^*(u) - \tau_k)/e^{g^*(u)})$ toward 0 or 1, reducing the information carried by the anchor labels and again weakening the curvature.*

Following the reviewer's suggestion, we now move this discussion to the main text and further provide a sensitivity analysis of the anchor threshold q. Empirically, the two anchor thresholds are set to the q-th and $(1-q)$-th quantiles of the centered response-level score distribution. In the main experiments, we set $q = 0.25$, corresponding to the 25th and 75th percentiles. In this sensitivity analysis, we vary q over $\{0.05, 0.15, 0.25, 0.35, 0.45\}$ on MultiPref-8K with the `Llama-3.2-3B-Instruct` backbone, while keeping all other hyperparameters fixed. As shown in Table 10, $q = 0.15$ and $q = 0.25$, consistently yield strong performance. For example, Two-anchor-Skywork achieves the best RewardBench accuracy at both $q = 0.15$ and $q = 0.25$, with the best Brier score and cross-entropy at $q = 0.25$. For Two-anchor-DeepSeek-Flash, the default choice $q = 0.25$ achieves the best performance in RewardBench, while $q = 0.15$ performs best in PPE-P. For Two-anchor-DeepSeek-Pro, $q = 0.15$ performs best, while $q = 0.25$ remains competitive. The sensitivity results also empirically support the trade-off discussed in Remark 1. When q is very small, the thresholds move to the tails and the anchor labels become less informative; when q is close to 0.5, the two thresholds become too close and the two anchors are less distinguishable. The strong performance at moderate values, especially $q = 0.15$ and $q = 0.25$, is consistent with this theoretical discussion.

(4) Sensitivity analysis of the weighting parameter λ. Thank you for your suggestion. We analyze the sensitivity of our method to the weighting parameter λ. we vary λ over $\{0.01, 0.10, 0.50\}$ on

MultiPref-8K and HelpSteer2-9K with the `Llama-3.2-3B-Instruct` backbone, while keeping all other hyperparameters fixed. As shown in Table 11, our approach is robust to the weighting parameter λ. Across both MultiPref-8K and HelpSteer2-9K, the two-anchor variants with different λ values consistently outperform or remain competitive with BT, BT(Hard), and Gaussian baselines. For simplicity and fair comparison, we fix $\lambda = 0.1$ across all experiments. Moreover, each dataset is split into training and held-out validation sets in our implement. Following the reviewer's suggestion, if one aims to tune this hyperparameter, a standard approach in reward modeling literature is to select the optimal hyperparameter based on held-out validation performance [Wang et al., 2025b].

(5) Analyse the effects of weaker or noisy anchor sources. Thank you for the constructive comment. We conduct two experiments to evaluate this issue. First, we corrupt the ordinal three-class anchor labels with noise rate $\rho \in \{0.0, 0.1, 0.2, 0.3\}$ to simulate moisy anchor sources. Given a noise rate ρ, each anchor label is kept unchanged with probability $1 - \rho$ and replaced by a randomly selected different valid label with probability ρ. Second, we use a weaker anchor model, `Qwen3-4B-Instruct-2507`, to generate anchor labels. As shown in Table 12, the performance of the two-anchor variants degrades as anchor noise increases. Under mild noise, it still remains competitive with or better than the preference-only baselines on both MultiPref-8K and HelpSteer2-9K. The two-anchor variant guided by the weaker Qwen3-4B anchor also remains comparable to the baselines on RewardBench. These results suggest that our method benefits from high-quality anchors, while still being robust to weaker or noisy anchor supervision.

(6) Comparison to Advanced Models and Motivation of the two-anchor scheme. This is a good point. The anchors used in our framework are strong judges and can be seen as advanced models. Thus, we further compare our trained two-anchor 8B reward models with these advanced models on MultiPref-8K and HelpSteer2-9K. As shown in Table 13, the trained two-anchor reward models are competitive with advanced models on out-of-distribution benchmarks. However, on both in-distribution held-out sets, the advanced models are consistently worse than the trained two-anchor reward models in accuracy and calibration. Out-of-distribution reward benchmarks assess the generalization capabilities of reward models, but pluralistic alignment requires matching the target human preference distribution. Directly using an advanced model can provide strong out-of-distribution performance, but it may not align with the specific preference distribution of the target population. Our proposed two-anchor scheme uses advanced anchor models only as coarse auxiliary supervision and trains the reward model on the target preference data. In this way, the proposed two-anchor scheme preserves strong out-of-distribution performance while better aligning the in-distribution pluralistic human preferences, which is the central motivation of our approach.

(7) Clarification on the details of Figure 1(a). Thank you for the this comment. We clarify the plotted statistic and the meaning of the x-axis in Figure 1(a). Following the definition in Eq. (1), the reward margin is $z_i = r_\phi(x_i, y_{i,1}) - r_\phi(x_i, y_{i,2})$. In Figure 1(a), at each training step t, we report the average reward margin over the current training mini-batch:

$$\bar{z}(t) = \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \left[r_{\phi_t}(x_i, y_{i,1}) - r_{\phi_t}(x_i, y_{i,2}) \right],$$

where $\mathcal{B}_t$ denotes the mini-batch at step t. Thus, the x-axis denotes training steps, and the y-axis denotes the average training-batch reward margin. The purpose of this figure is to visualize how reward margins evolve during training. The clarified figure shows that BT(Hard) tends to produce larger margins, while soft-label BT preserves disagreement information through smaller margins.

Table 8: Overall performance of same-source and cross-source anchor experiments with the `Llama-3.2-3B-Instruct` backbone. Bold numbers represent the best performance within each D_{pref} block. Underlined numbers represent the second best.

D_{pref}	D_{anc}	Method	RewardBench			PPE-P		
			Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓
MultiPref-8K	–	BT	71.7%	0.209	0.604	59.3%	0.236	0.665
		BT(Hard)	75.9%	0.194	0.703	57.1%	0.295	0.962
		Gaussian	73.2%	0.201	0.584	57.7%	0.242	0.680
	MultiPref	Two-anchor-Skywork	**82.1%**	0.163	0.490	60.0%	**0.234**	**0.664**
		Two-anchor-DeepSeek-Pro	78.4%	0.180	0.534	56.3%	0.244	0.682
		Two-anchor-DeepSeek-Flash	80.4%	0.167	0.502	58.2%	0.239	0.672
	HelpSteer2	Two-anchor-Skywork	78.9%	0.1682	0.5002	59.6%	0.2368	0.6730
		Two-anchor-DeepSeek-Pro	80.3%	0.1667	0.4999	59.4%	0.2349	0.6642
		Two-anchor-DeepSeek-Flash	78.4%	0.1740	0.5189	57.8%	0.2400	0.6759
	HelpSteer3	Two-anchor-Skywork	81.0%	**0.1583**	**0.4768**	**61.0%**	0.2411	0.7010
		Two-anchor-DeepSeek-Pro	73.8%	0.1877	0.5469	57.9%	0.2409	0.6802
		Two-anchor-DeepSeek-Flash	77.2%	0.1756	0.5163	58.4%	0.2421	0.6871
	PersonalLLM	Two-anchor-Skywork	77.0%	0.1765	0.5195	60.3%	0.2341	0.6664
		Two-anchor-DeepSeek-Pro	74.0%	0.1898	0.5581	56.4%	0.2428	0.6813
		Two-anchor-DeepSeek-Flash	75.8%	0.1868	0.5501	58.6%	0.2381	0.6711
HelpSteer2-9K	–	BT	84.8%	0.140	0.444	58.8%	0.251	0.719
		BT(Hard)	81.1%	0.149	0.509	58.4%	0.284	0.932
		Gaussian	81.9%	0.161	0.520	58.2%	0.251	0.729
	HelpSteer2	Two-anchor-Skywork	85.5%	0.132	0.413	60.2%	0.241	0.694
		Two-anchor-DeepSeek-Pro	82.1%	0.157	0.483	59.4%	**0.234**	**0.660**
		Two-anchor-DeepSeek-Flash	83.0%	0.151	0.467	60.0%	0.237	0.672
	MultiPref	Two-anchor-Skywork	**87.3%**	**0.1202**	**0.3892**	59.8%	0.2522	0.7710
		Two-anchor-DeepSeek-Pro	79.9%	0.1674	0.5088	57.9%	0.2426	0.6829
		Two-anchor-DeepSeek-Flash	78.4%	0.1726	0.5282	58.7%	0.2446	0.6956
	HelpSteer3	Two-anchor-Skywork	85.7%	0.1282	0.4229	**61.6%**	0.2461	0.7466
		Two-anchor-DeepSeek-Pro	83.6%	0.1406	0.4430	59.4%	0.2477	0.7134
		Two-anchor-DeepSeek-Flash	83.2%	0.1574	0.4848	59.0%	0.2343	0.6612
	PersonalLLM	Two-anchor-Skywork	82.0%	0.1470	0.4692	60.5%	0.2481	0.7456
		Two-anchor-DeepSeek-Pro	77.4%	0.1763	0.5397	57.2%	0.2507	0.7077
		Two-anchor-DeepSeek-Flash	76.8%	0.1751	0.5273	58.9%	0.2368	0.6675

Table 9: RewardBench preference correlation between anchor models and trained reward models. Higher values indicate stronger similarity between the pairwise preference patterns of the trained reward model and the anchor model. Bold numbers indicate the largest value in each anchor row.

D_{pref}	Size	Anchor	Preference-only RM			Two-anchor RM		
			BT	BT(Hard)	Gaussian	Skywork	DeepSeek-Pro	DeepSeek-Flash
MultiPref-8K	1B	Skywork	0.131	0.173	0.118	**0.249**	0.159	0.170
		DeepSeek-Pro	0.058	0.112	0.084	**0.157**	0.087	0.106
		DeepSeek-Flash	0.090	0.134	0.104	**0.175**	0.122	0.124
	3B	Skywork	0.242	0.287	0.264	**0.417**	0.298	0.399
		DeepSeek-Pro	0.156	0.171	0.109	0.203	0.173	**0.217**
		DeepSeek-Flash	0.155	0.197	0.132	0.247	0.182	**0.251**
	8B	Skywork	0.263	0.261	0.307	**0.407**	0.363	0.392
		DeepSeek-Pro	0.182	0.182	0.194	0.242	0.252	**0.261**
		DeepSeek-Flash	0.208	0.224	0.249	0.296	0.281	**0.315**
HelpSteer2-9K	1B	Skywork	0.193	0.174	0.201	**0.273**	0.242	0.236
		DeepSeek-Pro	0.125	0.127	0.143	**0.182**	0.158	0.169
		DeepSeek-Flash	0.120	0.106	0.135	**0.189**	0.154	0.163
	3B	Skywork	0.421	0.339	0.408	**0.494**	0.357	0.397
		DeepSeek-Pro	0.236	0.196	0.184	**0.250**	0.189	0.219
		DeepSeek-Flash	0.260	0.213	0.236	**0.275**	0.212	0.250
	8B	Skywork	0.421	0.378	**0.450**	0.443	0.432	0.425
		DeepSeek-Pro	**0.296**	0.253	0.268	0.262	0.275	0.267
		DeepSeek-Flash	0.289	0.297	0.284	0.298	**0.319**	0.309

Table 10: Sensitivity analysis of the anchor threshold q on MultiPref-8K with the `Llama-3.2-3B-Instruct` backbone. The two anchor thresholds are set to the q-th and $(1-q)$-th quantiles of the centered response-level score distribution.

Method	q	**RewardBench**			**PPE-P**		
		Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓
Two-anchor-Skywork	0.05	78.9%	0.1748	0.5187	58.4%	0.2390	0.6732
	0.15	**82.1%**	0.1631	0.4911	59.7%	0.2352	0.6656
	0.25	**82.1%**	**0.1630**	**0.4900**	**60.0%**	**0.2340**	**0.6640**
	0.35	79.5%	0.1690	0.5007	59.1%	0.2396	0.6782
	0.45	80.0%	0.1660	0.4942	58.8%	0.2398	0.6784
Two-anchor-DeepSeek-Pro	0.05	73.2%	0.2011	0.5836	56.8%	0.2443	0.6856
	0.15	**79.9%**	**0.1706**	**0.5093**	**57.8%**	**0.2389**	**0.6724**
	0.25	78.4%	0.1800	0.5340	56.3%	0.2440	0.6820
	0.35	64.9%	0.2240	0.6375	56.4%	0.2422	0.6778
	0.45	76.8%	0.1798	0.5287	57.2%	0.2410	0.6771
Two-anchor-DeepSeek-Flash	0.05	76.4%	0.1894	0.5573	56.9%	0.2436	0.6830
	0.15	78.6%	0.1762	0.5216	**58.6%**	**0.2378**	**0.6706**
	0.25	**80.4%**	**0.1670**	**0.5020**	58.2%	0.2390	0.6720
	0.35	78.1%	0.1740	0.5197	56.7%	0.2405	0.6747
	0.45	78.2%	0.1740	0.5161	58.3%	0.2395	0.6742

Table 11: Sensitivity analysis of the weighting parameter λ with the `Llama-3.2-3B-Instruct` backbone. $\lambda = 0.1$ is used as the default in the main experiments. Bold numbers represent the best performance across different λ values within each two-anchor method block.

D_{pref}	**Method**	λ	**RewardBench**			**PPE-P**		
			Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓
MultiPref-8K	BT	–	71.7%	0.209	0.604	59.3%	0.236	0.665
	BT(Hard)	–	75.9%	0.194	0.703	57.1%	0.295	0.962
	Gaussian	–	73.2%	0.201	0.584	57.7%	0.242	0.680
	Two-anchor-Skywork	0.01	75.5%	0.1914	0.5591	58.4%	0.2401	0.6755
		0.10	82.1%	0.1628	0.4897	60.0%	**0.2344**	**0.6642**
		0.50	**84.9%**	**0.1462**	**0.4475**	**60.3%**	0.2380	0.6860
	Two-anchor-DeepSeek-Pro	0.01	76.4%	0.1879	0.5514	**57.6%**	**0.2413**	**0.6777**
		0.10	78.4%	0.1799	0.5342	56.3%	0.2435	0.6819
		0.50	**78.7%**	**0.1745**	**0.5236**	57.3%	0.2451	0.6900
	Two-anchor-DeepSeek-Flash	0.01	76.8%	0.1857	0.5468	57.5%	0.2419	0.6790
		0.10	80.4%	**0.1671**	**0.5016**	**58.2%**	**0.2385**	**0.6725**
		0.50	**80.5%**	0.1699	0.5120	57.3%	0.2439	0.6863
HelpSteer2-9K	BT	–	84.8%	0.140	0.444	58.8%	0.251	0.719
	BT(Hard)	–	81.1%	0.149	0.509	58.4%	0.284	0.932
	Gaussian	–	81.9%	0.161	0.520	58.2%	0.251	0.729
	Two-anchor-Skywork	0.01	82.2%	0.1527	0.4718	59.5%	0.2478	0.7153
		0.10	**85.5%**	**0.1319**	**0.4133**	**60.2%**	0.2410	0.6937
		0.50	83.8%	0.1469	0.4504	59.5%	**0.2376**	**0.6743**
	Two-anchor-DeepSeek-Pro	0.01	81.2%	0.1628	0.5157	58.7%	0.2472	0.7090
		0.10	82.1%	0.1568	0.4832	**59.4%**	**0.2337**	**0.6599**
		0.50	**85.2%**	**0.1386**	**0.4365**	59.0%	0.2419	0.6889
	Two-anchor-DeepSeek-Flash	0.01	82.2%	0.1552	0.4955	58.6%	0.2457	0.7042
		0.10	83.0%	0.1506	0.4672	**60.0%**	**0.2366**	**0.6724**
		0.50	**84.9%**	**0.1388**	**0.4346**	59.0%	0.2402	0.6832

Table 12: Overall performance of weaker or noisy anchor sources with the `Llama-3.2-3B-Instruct` backbone. We corrupt the ordinal three-class anchor labels with noise rate ρ. Bold numbers represent the best performance within each D_{pref} block, and underlined numbers represent the second best.

D_{pref}	Method	Noise rate ρ	RewardBench			PPE-P		
			Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓
MultiPref-8K	BT	–	71.7%	0.209	0.604	59.3%	0.236	0.665
	BT(Hard)	–	75.9%	0.194	0.703	57.1%	0.295	0.962
	Gaussian	–	73.2%	0.201	0.584	57.7%	0.242	0.680
	Two-anchor-Qwen3-4B	–	73.8%	0.1942	0.5692	55.0%	0.2495	0.6976
	Two-anchor-Skywork	–	**82.1%**	**0.163**	**0.490**	**60.0%**	0.234	0.664
		10%	81.1%	0.1662	0.4973	58.6%	0.2398	0.6761
		20%	76.2%	0.1914	0.5620	59.6%	**0.2318**	**0.6555**
		30%	74.7%	0.1997	0.5828	58.9%	0.2342	0.6607
	Two-anchor-DeepSeek-Pro	–	78.4%	0.180	0.534	56.3%	0.244	0.682
		10%	77.4%	0.1829	0.5417	56.2%	0.2416	0.6770
		20%	75.5%	0.1911	0.5617	56.2%	0.2419	0.6770
		30%	77.1%	0.1896	0.5593	56.3%	0.2414	0.6759
	Two-anchor-DeepSeek-Flash	–	80.4%	0.167	0.502	58.2%	0.239	0.672
		10%	77.5%	0.1811	0.5362	57.8%	0.2389	0.6711
		20%	77.8%	0.1848	0.5474	57.8%	0.2390	0.6710
		30%	74.8%	0.1928	0.5670	56.5%	0.2418	0.6767
HelpSteer2-9K	BT	–	84.8%	0.140	0.444	58.8%	0.251	0.719
	BT(Hard)	–	81.1%	0.149	0.509	58.4%	0.284	0.932
	Gaussian	–	81.9%	0.161	0.520	58.2%	0.251	0.729
	Two-anchor-Qwen3-4B	–	81.9%	0.1657	0.5035	57.8%	0.2409	0.6777
	Two-anchor-Skywork	–	**85.5%**	**0.132**	**0.413**	**60.2%**	0.241	0.694
		10%	84.5%	0.1396	0.4310	59.8%	0.2402	0.6870
		20%	84.7%	0.1447	0.4455	59.4%	0.2379	0.6754
		30%	82.5%	0.1587	0.4824	58.9%	0.2388	0.6776
	Two-anchor-DeepSeek-Pro	–	82.1%	0.157	0.483	59.4%	**0.234**	**0.660**
		10%	80.8%	0.1638	0.5044	60.1%	0.2394	0.6808
		20%	82.9%	0.1601	0.4927	59.4%	0.2371	0.6705
		30%	81.5%	0.1624	0.4966	59.2%	0.2381	0.6738
	Two-anchor-DeepSeek-Flash	–	83.0%	0.151	0.467	60.0%	0.237	0.672
		10%	80.6%	0.1627	0.4983	60.1%	0.2353	0.6689
		20%	83.4%	0.1492	0.4622	59.4%	0.2377	0.6726
		30%	81.6%	0.1591	0.4877	57.8%	0.2394	0.6742

Table 13: Comparison with advanced models on both in-distribution held-out sets and out-of-distribution benchmarks with the `Llama-3.1-8B-Instruct` as backbone. MultiPref ID and HelpSteer2 ID denote the held-out split of the corresponding D_{pref}. Bold numbers represent the best performance. Underlined numbers represent the second best.

Setting	Method	MultiPref ID			HelpSteer2 ID			RewardBench			PPE-P		
		Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓	Acc ↑	Brier ↓	CE ↓
Trained RM on MultiPref-8K	Two-anchor-Skywork	**77.9%**	**0.0611**	0.6062	–	–	–	82.9%	0.1507	0.4589	61.3%	**0.2284**	0.6523
	Two-anchor-DeepSeek-Pro	74.6%	0.0632	0.6031	–	–	–	80.7%	0.1643	0.4958	60.5%	0.2321	0.6581
	Two-anchor-DeepSeek-Flash	76.2%	0.0614	**0.6014**	–	–	–	83.1%	0.1582	0.4831	61.0%	0.2301	0.6538
Trained RM on HelpSteer2-9K	Two-anchor-Skywork	–	–	–	**75.3%**	**0.1021**	0.5613	88.0%	0.1287	0.4082	61.5%	0.2303	0.6586
	Two-anchor-DeepSeek-Pro	–	–	–	73.7%	0.1060	0.5660	88.8%	0.1169	0.3859	61.8%	0.2335	0.6757
	Two-anchor-DeepSeek-Flash	–	–	–	74.8%	0.1055	**0.5595**	87.6%	0.1317	0.4176	61.3%	0.2292	**0.6508**
Advanced model (direct)	Skywork-Reward-V2	71.6%	0.1359	1.6602	73.9%	0.1465	1.0219	**93.2%**	**0.0605**	**0.2466**	**66.5%**	0.2520	1.0630
	DeepSeek-V4-Pro	65.3%	0.0964	0.7144	70.9%	0.1320	0.6551	88.9%	0.0843	0.2812	59.3%	0.2290	0.6960
	DeepSeek-V4-Flash	70.5%	0.0920	0.7362	70.7%	0.1403	0.6879	86.1%	0.1017	0.3353	57.6%	0.2340	0.7320