# Reward Modeling with Ordinal Feedback: Wisdom of the Crowd

**Shang Liu** [*1] **Yu Pan** [*2] **Guanting Chen** [3] **Xiaocheng Li** [1]

## Abstract

The canonical setup of learning a reward model (RM) from human preferences with binary feedback discards potentially useful samples (such as "tied" between the two responses) and loses fine-grained information (such as "slightly better"). This paper proposes a framework for learning RMs under *ordinal feedback*, generalizing the binary feedback to arbitrary granularity. We first identify a marginal unbiasedness condition, which generalizes the existing assumption of the binary feedback. The condition is validated via the sociological concept called "wisdom of the crowd". Under this condition, we develop a natural probability model and prove the benefits of fine-grained feedback in terms of reducing the Rademacher complexity, which may be of independent interest to another problem: the bias-variance trade-off in knowledge distillation. The framework also sheds light on designing guidelines for human annotators. Our numerical experiments validate that: (1) fine-grained feedback leads to better RM learning for both in- and out-of-distribution settings; (2) incorporating a certain proportion of tied samples boosts RM learning.

## 1. Introduction

Reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2019; Askell et al., 2021; Ouyang et al., 2022) is vital to aligning large language models (LLMs) with human preferences. The RLHF involves either explicitly training a reward model (RM) from human preferences data (Ouyang et al., 2022) or implicitly using the LLM itself as one (Rafailov et al., 2024).

[*]Equal contribution [1]Imperial College Business School, Imperial College London, UK [2]The University of Sydney Business School, The University of Sydney, Australia [3]Department of Statistics and Operations Research, UNC at Chapel Hill, USA. Correspondence to: Guanting Chen <guanting@unc.edu>, Xiaocheng Li <xiaocheng.li@imperial.ac.uk>.

*Proceedings of the $42^{nd}$ International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. 

However, there is an inconsistency between current ways of collecting human preference data and the training of reward models. For example, the Llama team collects fine-grained human feedback: the annotations are made to 4 levels named "significantly better", "better", "slightly better", and "marginally better" (Llama Team, 2024), while the post-training of Llama 3 treats "significantly better" and "better" as the same and discard all the others. Such a process wastes the potentially useful samples that cost the human annotators additional time and discards the useful information hidden in the preference level.

In this paper, we study the problem of reward modeling under ordinal feedback. We relate the annotator's preference feedback with the probability that one response is better than the other on a population level. We introduce a marginal unbiasedness condition that validates the probability setup of the ordinal feedback system. The assumption is justified in two ways: (1) its restricted version in the binary feedback case is already widely adopted; (2) the assumption corresponds to the "wisdom of the crowd" concept in sociology. Under the assumption, we build a probability model of ordinal feedback. We also propose a learning objective as a generalized version of the binary feedback by, for example, replacing the binary-valued empirical probability with more fine-grained values in the cross-entropy loss. Theoretically, we establish the advantage of ordinal feedback, which draws an interesting connection with the literature on soft labeling and knowledge distillation.

Our paper is organized as follows:

In Section 2, we model the feedback by the probability that one response is better than the other on a population level. The binary feedback ($\mathcal{Z} = \{0, 1\}$) is extended to the general *ordinal feedback* ($\mathcal{Z} = \{z_j\}_{j=1}^m$ for $0 \leq z_1 < \cdots < z_m \leq 1$), providing a way to transform the qualitative label into quantitative ones.

In Section 3, we build up the probability model of ordinal feedback. We present the assumption (Assumption 3.1) that the annotators are labeling the human preference data in a marginally unbiased sense to the population-wise oracle. Such an assumption (which we call "wisdom of the crowd") is only a generalization of the Bernoulli feedback. In this light, we suggest revising the annotation guideline by providing a direct quantitative description of the qualitative opinions.

In Section 4, we prove the statistical benefits of the ordinal feedback. For example, the Rademacher complexity is reduced if the cross-entropy loss is adopted. The result is proved via a special coupling argument that we call *hierarchical expectation*, which also provides a new bias-variance trade-off in knowledge distillation and soft labeling.

In Section 5, we conduct two numerical experiments. The first experiment sets up four different ordinal feedback systems (oracle, 5-level, 3-level, and binary) and validates the theoretical findings that fine-grained ordinal feedback achieves higher accuracies in both in- and out-of-distribution settings. The second experiment mixes the training data with different proportions of tied and untied samples. With the same number of training samples, we find out that a certain level of tied samples boosts RM learning.

We defer more related works to Appendix A.

## 2. Problem Setup

Consider the task of reward modeling based on the pairwise preference data. Each sample consists of a tuple $(x, y_1, y_2, z)$ where $x \in \mathcal{X}$ denotes a prompt, $y_1, y_2 \in \mathcal{Y}$ are two candidate responses to the prompt $x$, and $Z$ is a random variable (taking values in $\mathcal{Z}$) that denotes the feedback (by either human annotators or advanced AI models) indicating the preference between $y_1$ and $y_2$. The feedback $Z$ is a proxy of the probability that $y_1$ is better than $y_2$ for the prompt $x$, denoted by $\mathbb{P}(y_1 \succ y_2|x)$.

Reward modeling is the learning of a reward function $r_\theta(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with $\theta \in \Theta$ from an annotated dataset $\mathcal{D}_\mathcal{Z} \coloneqq \{(x_i, y_{i,1}, y_{i,2}, Z_i)\}_{i=1}^n$ and approximates the preference accordingly. For example, the prevalent Bradley-Terry model (Bradley & Terry, 1952) says:

$$\hat{\mathbb{P}}_\theta(y_1 \succ y_2|x) = \frac{\exp\big(r_\theta(x, y_1)\big)}{\exp\big(r_\theta(x, y_1)\big) + \exp\big(r_\theta(x, y_2)\big)},$$

where the probabilities are approximated by the softmax reward values.

**Binary feedback**: In the canonical setup ((Bai et al., 2022; Ouyang et al., 2022) among others), the feedback $Z$ takes binary values, i.e., $\mathcal{Z} = \{0, 1\}$. Here, one assumes $Z_i$ is a Bernoulli random variable such that

$$\mathbb{P}(Z_i = 1) = 1 - \mathbb{P}(Z_i = 0) = \mathbb{P}\left(y_{i,1} \succ y_{i,2}|x_i\right). \quad (1)$$

This assumption is the backbone of the training of many reward models.

**Ordinal feedback**: In this paper, we consider the setting of *ordinal feedback*, which gives a richer feedback structure than the binary feedback above and is defined as follows.

**Definition 2.1** (Ordinal Feedback)**.** Suppose the feedback $Z$ takes values in $\mathcal{Z} \coloneqq \{z_1, \ldots, z_m\}$ where $0 \le z_1 < \cdots < z_m \le 1$, we call $Z$ an ordinal feedback and $\mathcal{Z}$ the ordinal feedback set.

The binary feedback is a special case of the ordinal feedback by letting $m = 2$, $z_1 = 0$, and $z_2 = 1$. The motivation to introduce the ordinal feedback definition is to capture the richer options in practice, where the annotator is allowed to choose from $\mathcal{Z}_{\text{text}} = \{\text{better/worse than, same as}\}$ or $\mathcal{Z}_{\text{text}} = \{\text{better/worse than, slightly better/worse}\}$ to describe the preference between responses $y_1$ and $y_2$. We defer to the next section the questions of how to match $\mathcal{Z}_{\text{text}}$ with $\mathcal{Z}$.

For now, suppose we have the dataset $\mathcal{D}_\mathcal{Z}$ taking values in $\mathcal{Z}$ translated from $\mathcal{Z}_{\text{text}}$. We propose to learn the reward model by minimizing the following cross-entropy loss (for other loss functions, see Appendix C.2)

$$\begin{aligned}\min_\theta \sum_{i=1}^n &- Z_i \cdot \log\left(\sigma\left(r_\theta(x_i, y_{i,1}) - r_\theta(x_i, y_{i,2})\right)\right) \\ &- (1 - Z_i) \cdot \log\left(\sigma\left(r_\theta(x_i, y_{i,2}) - r_\theta(x_i, y_{i,1})\right)\right),\end{aligned} \quad (2)$$

where $\sigma(\cdot)$ is the sigmoid function such that $\sigma(x) = \exp(x)/(1 + \exp(x))$.

When the feedback is binary $Z_i \in \{0, 1\}$, the above objective function reduces to exactly how people learn the reward models under the Bradley-Terry assumption. We generalize the binary feedback to avoid: (1) discarding potentially useful samples (for example, neglecting "same as"); or (2) losing subtle information by merging different options (for example, treating "better" and "significantly better" as the same).

This generalization raises three questions: (1) How to translate between qualitative $\mathcal{Z}_{\text{text}}$ and quantitative $\mathcal{Z}$? (2) How to relate the quantities in $\mathcal{Z}$ to $\mathbb{P}\left(y_1 \succ y_2|x_i\right)$? (3) How does this setup differ from the canonical binary feedback? The first two questions are addressed in Section 3, and the last is addressed theoretically in Section 4 and numerically in Section 5.

## 3. Probability Model of Ordinal Feedback

For a population of human preferences, we first define the oracle feedback as

$$z_{\text{oracle}}(x, y_1, y_2) \coloneqq \mathbb{P}\left(y_1 \succ y_2|x\right).$$

This is *the* preference model that one aims to learn, regardless of whatever model is assumed (for example, BT model or others). Here we should think of a human annotator as a random draw from the population.

**Assumption 3.1** (Ordinal feedback probability model – wisdom of the crowd)**.** We assume the ordinal feedback $Z$ defined in Definition 2.1 satisfies

$$\mathbb{E}[Z|(x, y_1, y_2)] = z_{\text{oracle}}(x, y_1, y_2) \\ \text{for any } (x, y_1, y_2) \in \mathcal{X} \times \mathcal{Y}^2.$$

As shown later, this assumption alone is sufficient to define the probability model of the ordinal feedback setting and validate the learning of the reward model. To interpret the assumption, we note that it is not stricter than the existing assumption (1) that people impose for the binary feedback model. Under the binary feedback model where $Z$ takes values in $\mathcal{Z} = \{0, 1\}$, Assumption 3.1 is equivalent to (1).

To see another example, consider the set $\mathcal{Z} = \{0, 0.5, 1\}$ where the labels 0 and 1 denote "better" and "worse" respectively, and the label 0.5 denotes "same as". The assumption requires that with $Z$ taking values in this new $\mathcal{Z}$ with an additional 0.5 option, its expectation matches the oracle value $z_{\text{oracle}}$ on the sample $(x, y_1, y_2)$. Under the general ordinal feedback setting, human annotators label their preferences on different scales, i.e.,

$$\mathcal{Z} = \{0, 1\},\ \{0, 0.5, 1\},\ \text{or}\ \{0, 0.25, 0.5, 0.75, 1\}.$$

The assumption says that the change of granularity does not introduce bias on the population level.

**Sociological interpretation**: We name the assumption by "wisdom of the crowd"; the concept was first coined by the article *Vox Populi* (Galton, 1907) for a social experiment under the title "the voice of the people". The social experiment is about a weight-judging competition conducted in England for random people to guess the weight of an ox. The average of all 787 guesses was 1,197 pounds, while the actual weight was 1,198 pounds, as shown in Figure 1. Each individual's guess can be far off the target, yet the population average tends to be very accurate. For the context of preference annotation, Assumption 3.1 and the current practice of human annotation exercise the wisdom of the crowd in two folds: First, each individual annotator has no access to the population preference $z_{\text{oracle}}$, but their annotation can be viewed as an unbiased random realization of $z_{\text{oracle}}$. Second, such unbiasedness does not change if a difference annotation scale $\mathcal{Z}$ (feedback set) is used.

### 3.1. Implications on annotation guidance

In practice, annotators label in the set $\mathcal{Z}_{\text{text}}$ (e.g. {better, same as, worse}), and this results in a gap towards the set $\mathcal{Z} = \{z_1, ..., z_m\}$ used in the learning of the reward model (2). For the binary or 3-level feedback setting, the following conversion is natural and reflects the thinking process of the annotator:

$$\mathcal{Z}_{\text{text}} = \{\text{better than, worse than}\} \Leftrightarrow \mathcal{Z} = \{0, 1\},$$

$$\mathcal{Z}_{\text{text}} = \begin{Bmatrix} \text{better than} \\ \text{same as} \\ \text{worse than} \end{Bmatrix} \Leftrightarrow \mathcal{Z} = \begin{Bmatrix} 0 \\ 0.5 \\ 1 \end{Bmatrix}.$$

For a more fine-grained 5-level feedback setting, it is less clear how one can convert as follows,

$$\mathcal{Z}_{\text{text}} = \begin{Bmatrix} \text{better than} \\ \text{slightly better} \\ \text{same as} \\ \text{slightly worse} \\ \text{worse than} \end{Bmatrix} \Leftrightarrow \mathcal{Z} = \begin{Bmatrix} 0 \\ ?\% \\ 0.5 \\ 1-?\% \\ 1 \end{Bmatrix}.$$

In this light, our ordinal feedback model and Assumption 3.1 provide the following insights in guiding the annotations. We suggest that rather than providing vague wording of "better than" or "sightly better", one can write the following in the guidance to the human annotators:

> **Example annotation guideline.** The label "slightly better" represents that 75% of the population think response $y_1$ is better than response $y_2$. The label "slightly worse" represents that 25% of the population think response $y_1$ is better than response $y_2$.

Such additional guidance endows the narrative labels in $\mathcal{Z}_{\text{text}}$ with a numerical meaning. The corresponding numerical values can be directly used in the learning objective (2).

### 3.2. Existence of feedback probability model

In this subsection, we detour from the discussions of reward modeling and show the existence of probability models that satisfy Assumption 3.1, which ensures the assumption is well-defined. The following theorem states that for any ordinal feedback set $\mathcal{Z}$ and any oracle model $z_{\text{oracle}}$, there exists a probability model satisfying Assumption 3.1; furthermore, the model is uniquely determined by the oracle $z_{\text{oracle}}$ (up to convex combinations).

**Theorem 3.2.** *For any ordinal feedback set $\mathcal{Z} = \{z_1, \ldots, z_m\}$ and any oracle model $z_{oracle}(x, y_1, y_2)$, if $z_{oracle}(x, y_1, y_2) \in [z_j, z_k]$ for some $j, k \in [m]$, then one can set the marginal probability measure $\mu_{j,k}(z) := \mathbb{P}(Z = z|(x, y_1, y_2))$ to be*

$$\mu_{j,k}(z) = \begin{cases} (z_k - z_{oracle})/(z_k - z_j), & \text{if } z = z_j, \\ (z_{oracle} - z_j)/(z_k - z_j), & \text{if } z = z_k, \\ 0, & \text{otherwise}. \end{cases}$$

*The ordinal feedback $Z$ under $\mu_{j,k}$ fulfills Assumption 3.1 (where the dependence on $(x, y_1, y_2)$ is omitted when there is no confusion).*

*On the other hand, for any ordinal feedback $Z$ with marginal probability measure $\mu(z) := \mathbb{P}(Z = z|(x, y_1, y_2))$ satisfying Assumption 3.1, there must exist non-negative real*

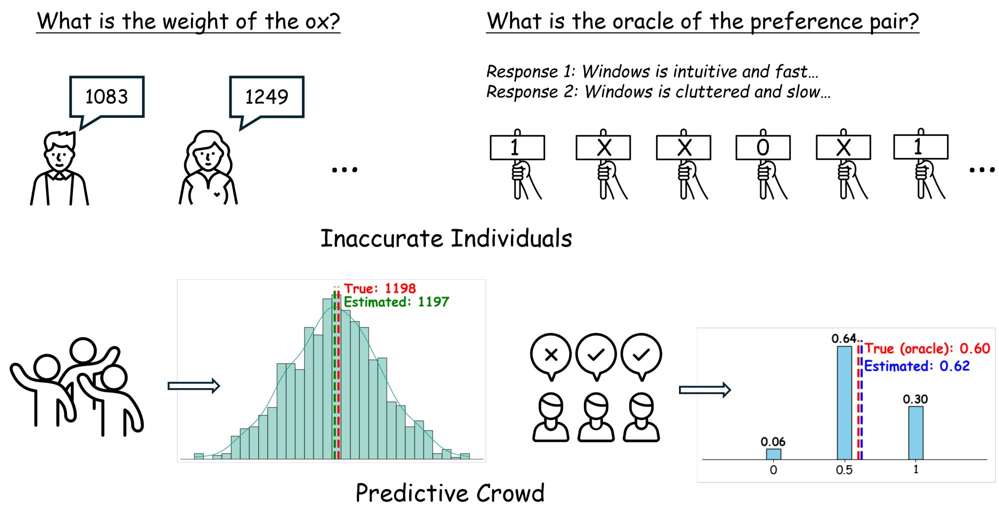



*Figure 1.* Wisdom of the crowd. Left: Each individual guess can be far off the target for an ox-weight-guessing social experiment, but the average tends to be very accurate. Each human annotator has no access to the population oracle preference model $z_{\text{oracle}}$, but their annotation constitutes an unbiased realization of $z_{\text{oracle}}$.

*numbers* $\sum_{j,k} \alpha_{j,k} = 1$ *such that* $\mu = \sum_{j,k} \alpha_{j,k}\ \mu_{j,k}$, *where the summation is made for all* $(j,k)$ *pairs such that* $z_{oracle} \in [z_j, z_k]$.

The theorem says that all such distributions should be (convex combinations of) some two-point distributions, where the weights are assigned to interpolate the linear system

$$\begin{cases} \mu_{j,k}(z_j) \cdot z_j + \mu_{j,k}(z_k) \cdot z_k = z_{\text{oracle}}, \\ \mu_{j,k}(z_j) + \mu_{j,k}(z_k) = 1. \end{cases}$$

For example, consider a reward model with ties where $\mathcal{Z} = \{0, 0.5, 1\}$. If the oracle feedback is, say, 0.8, then we can construct feedback $Z$ with probability masses $\mu(0.5) = 0.4$ and $\mu(1) = 0.6$ such that the unbiased assumption is fulfilled. In this way, if we are (in an ideal world) given an oracle model $z_{\text{oracle}}$, we can generate different unbiased ordinal feedback problem models accordingly, which is applied in our numerical experiments (Section 5).

## 4. Statistical Benefits of Ordinal Feedback

In this section, we provide a theoretical explanation of the benefits of ordinal feedback. The theory not only captures the benefits of the ordinal feedback model but also provides insights into the technique of *soft labeling* in knowledge distillation (Ba & Caruana, 2014; Hinton, 2015; Müller et al., 2019; Phuong & Lampert, 2019; Yuan et al., 2020; Zhou et al., 2021), which can be of independent interest. In short, the theoretical result says that any feedback model satisfying Assumption 3.1 reduces the Rademacher complexity compared to the canonical binary feedback model.

Before we proceed, we introduce the following affinity condition of the loss function that is satisfied by the cross-entropy loss (2) (and generalized hinge loss in Appendix B.2, for example).

$$\mathbb{E}_Z\left[\ell(Z, z)\right] = \ell\big(\mathbb{E}[Z], z\big) \text{ for any } z \in [0,1]. \tag{3}$$

Introducing the affinity condition (3) enables us to compare two ordinal feedback systems (say, $Z$ and $Z'$) if their conditional expectation is the same under Assumption 3.1. Specifically, if $\ell$ satisfies (3), then the population loss of $Z$ and $Z'$ are the same for arbitrary same $h$ (can be taken as the reward function $r$, for example),

$$\mathbb{E}_{x,y,Z}\left[\ell\big(Z, h(x, y_1, y_2)\big)\right] = \mathbb{E}_{x,y,Z'}\left[\ell\big(Z', h(x, y_1, y_2)\big)\right]. \tag{4}$$

However, (4) only states that two population losses are the same under the *same* function $h$; what if we get two different learned reward models $\hat{h}$ and $\hat{h}'$ from $Z$ and $Z'$ correspondingly? In the following, we shall see that from the perspective of the generalization bound (specifically, Rademacher complexity), a more fine-grained feedback system induces a smaller population loss.

### 4.1. Finite-sample benefits

We first introduce the concepts of *coupling* and *hierarchical expectation.*

**Definition 4.1** (Coupling)**.** For any two random variables $\xi$ and $\xi'$, if there exist two random variables $\zeta$ and $\zeta'$ over one probability space such that $\zeta$ has the same distribution as $\xi$ and $\zeta'$ the same as $\xi'$, we call them a coupling of $\xi$ and $\xi'$.

**Definition 4.2** (Hierarchical Expectation)**.** For any two ordinal feedback systems $Z$ and $Z'$ taking values in $\mathcal{Z}$ and $\mathcal{Z}'$

over the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we say that $Z$ is a *hierarchical expectation* of $Z'$ if there exists a combination of random variables $(W, W')$ over a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ such that: (a) $(W, W')$ forms a coupling between $Z$ and $Z'$; and (b) $W = \mathbb{E}[W'|W]$ holds almost surely.

The concept of hierarchical expectation defines the relative granularity of the feedback system. In general, if $Z$ is a hierarchical expectation of $Z'$, then we say $Z$ is more fine-grained than $Z'$ since $Z$ allows annotators to give more subtle responses.

**Proposition 4.3** (Existence of Hierarchical Expectation)**.** *For any two ordinal feedback systems $Z$ and $Z'$ taking values in $\mathcal{Z}$ and $\mathcal{Z}'$ over the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, suppose the marginal distribution of $Z$ is of measure $\mu = \sum_{z_i \in \mathcal{Z}} \alpha_i \delta_{z_i}$ and that of $Z'$ is $\mu' = \sum_{z'_i \in \mathcal{Z}'} \alpha'_i \delta_{z'_i}$, where $\delta(\cdot)$ is the Dirac delta distribution. If there exist real numbers $\beta_{j,k} \in [0, 1]$ such that*

*(a) $z_j$ is a convex combination of $z'_k$'s with coefficients $\beta_{j,k}$'s. That is, $\sum_{k, z'_k \in \mathcal{Z}'} \beta_{j,k} = 1$ and $\sum_{k, z'_k \in \mathcal{Z}'} \beta_{j,k} z'_k = z_j$ for any $z_j \in \mathcal{Z}$;*

*(b) $\sum_{j, z_j \in \mathcal{Z}} \beta_{j,k} \alpha_j = \alpha'_k$ for any $z'_k \in \mathcal{Z}'$;*

*Then $Z$ must be a hierarchical expectation of $Z'$. On the other hand, if $Z$ is a hierarchical expectation of $Z'$ according to coupling $(W, W')$ on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, then there must exist real numbers $\beta_{j,k} \in [0, 1]$ satisfying the above requirements by setting $\beta_{j,k} = \tilde{\mathbb{P}}(W' = z'_k | W = z_j)$.*

The following corollary gives some concrete examples (perhaps the most concerning ones).

**Corollary 4.4.** *Suppose $\mathcal{Z}$ and $\mathcal{Z}'$ are two ordinal feedback sets satisfying Assumption 3.1. Then $Z$ is always a hierarchical expectation of $Z'$ if any one of the following two conditions holds:*

*(a) $Z$ is the oracle model such that $Z = z_{oracle}$;*

*(b) $Z'$ is the binary feedback such that $\mathcal{Z}' = \{0, 1\}$.*

With the above definitions, we are ready to characterize the finite sample properties of different ordinal feedback systems using Rademacher complexity.

**Definition 4.5** (Rademacher Complexity)**.** Let $\mathcal{D}_{\mathcal{Z}} = \{(x_i, y_{i,1}, y_{i,2}, Z_i)\}_{i=1}^n$ be a dataset with $Z$ taking values in $\mathcal{Z}$. Consider a hypothesis class $\mathcal{H}$ of real-valued functions over $\mathcal{X} \times \mathcal{Y}^2$ and a loss function $\ell$. Then, the empirical Rademacher complexity is defined as

$$\mathrm{Rad}_{\mathcal{D}_{\mathcal{Z}}}(\ell \circ \mathcal{H}) := \frac{1}{n} \mathbb{E}_{\varepsilon} \left[ \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \varepsilon_i \ell\big(Z_i, h(x_i, y_{i,1}, y_{i,2})\big) \right],$$

where $\varepsilon_i$'s are independent random variables all taking values in $\{+1, -1\}$ with equal chances. By assuming that $(x_i, y_{i,1}, y_{i,2}, Z_i)$'s are i.i.d. and taking expectations over the entire distribution $\mathbb{P}$, we have the Rademacher complexity as

$$\mathrm{Rad}_{\mathcal{Z},n}(\ell \circ \mathcal{H}) := \mathbb{E}_{\mathcal{D}_{\mathcal{Z}} \sim \mathbb{P}^n} \big[ \mathrm{Rad}_{\mathcal{D}_{\mathcal{Z}}}(\ell \circ \mathcal{H}) \big].$$

Rademacher complexity is a popular tool for describing generalizing behavior. The following theorem says that a more fine-grained feedback system leads to a smaller (which is better) Rademacher complexity for any hypothesis class $\mathcal{H}$, where setting $\mathcal{H} = \{r_\theta | \theta \in \Theta\}$ yields results in RM learning. We include how the Rademacher complexity consequently affects the generalization bound in Appendix B.10 for completeness.

**Theorem 4.6.** *Suppose the loss function $\ell$ satisfies the affinity to feedback condition (3) and $\mathcal{H}$ is a hypothesis class of real-valued functions over $\mathcal{X} \times \mathcal{Y}^2$. For any two ordinal feedback systems $Z$ and $Z'$ taking values in $\mathcal{Z}$ and $\mathcal{Z}'$ such that $Z$ is a hierarchical expectation of $Z'$, we have*

$$\mathrm{Rad}_{\mathcal{Z},n}(\ell \circ \mathcal{H}) \le \mathrm{Rad}_{\mathcal{Z}',n}(\ell \circ \mathcal{H}).$$

As a consequence, the following corollary states that any feedback system that satisfies Assumption 3.1 can be viewed as in the middle of the binary feedback system and the ideal oracle feedback system $z_{\text{oracle}}$. It is impossible to access the population preference model $z_{\text{oracle}}$ through human annotators in practice, but any ordinal feedback system provides more fine-grained information and leads us towards $z_{\text{oracle}}$.

**Corollary 4.7** (Ordinal feedback better than binary)**.** *Suppose $Z$ is an ordinal feedback taking values in $\mathcal{Z}$ and $Z'$ is a binary feedback in $\mathcal{Z}' = \{0, 1\}$ over the same probability space. If they both satisfy Assumption 3.1 and the loss function $\ell$ satisfies condition (3), then*

$$\mathrm{Rad}_{\mathcal{Z}_{oracle},n}(\ell \circ \mathcal{H}) \le \mathrm{Rad}_{\mathcal{Z},n}(\ell \circ \mathcal{H}) \le \mathrm{Rad}_{\mathcal{Z}',n}(\ell \circ \mathcal{H}).$$

The above discussions show that a more fine-grained feedback system can help reduce learning difficulties. The problem of how large the benefit is has been discussed in Appendix B.11. In short, the reduction of Radmacher complexity results from Jensen's inequality gap; the gap is approximately the same order as the reduction of variance.

### 4.2. Implications on soft labeling

As noted earlier, the results developed above also have implications for soft labeling/knowledge distillation. We will show how the analysis can be applied to the context of soft labeling and induce a novel bias-variance trade-off for knowledge distillation. For general $k$-nary classification problems, the standard feedback (labeling of the target variable) is a $k$-dimensional one-hot vector. However, these all-zero-but-one labels might make the model overfit. Knowledge distillation (Hinton, 2015) is a well-known technique to regularize the model from fitting the noise. The original data is used to train a teacher model whose predictions are named *soft labels*. Then, a student model is trained to mimic the predictions of the teacher model, that is, minimizing the training loss against the soft labels generated by the teacher model rather than the original ones.

Existing theoretical works (Phuong & Lampert, 2019; Zhou et al., 2021; Pareek et al., 2024) are developed to understand the benefits of knowledge distillation and soft labeling. Our theoretical perspective in the preceding subsection provides a new perspective on that problem: the trained teacher model could be viewed as one (possibly biased) oracle feedback, and learning from the oracle eases the overfitting via reducing the labeling variance (and hence a smaller Rademacher complexity). Consider the following four learning paradigms:
(a) Oracle(ideal): learn from oracle labeled samples (denoted as $(x, y_{\text{oracle}})$'s in this subsection). $y_{\text{oracle}} = \mathbb{E}[y|x]$.
(b) Original: learn from the original samples (denoted as $(x, y)$'s in this subsection), where $y \sim y_{\text{oracle}}$ is a random one-hot vector.
(c) Knowledge distillation: learn from the teacher model $\mathcal{T}$'s output (denoted as $(x, \overline{y_{\mathcal{T}}})$'s in this subsection). The labels $\overline{y_{\mathcal{T}}}$'s are random vectors where the randomness comes from the teacher model $\mathcal{T}$.
(d) Sampling from teacher: learn from the teacher model, but not directly from $\overline{y_{\mathcal{T}}}$; instead, we use a sampled label $y_{\mathcal{T}} \sim \overline{y_{\mathcal{T}}}$.

What is the "bias-variance" tradeoff in the learning paradigm (c) compared to (b)? First, $y_{\text{oracle}}$ shares the same conditional expectation with $y$, and so do $\overline{y_{\mathcal{T}}}$ and $y_{\mathcal{T}}$, where their population cross-entropy losses are the same due to (4). The "bias" comes from $\overline{y_{\mathcal{T}}}$ as an imperfect estimation of $y_{\text{oracle}}$, while training according to (c) or (d) introduces an additional loss in the original population loss. That is the "bias" term

$$\text{Bias} := \mathbb{E}_{\mathcal{T}}\Big[\mathbb{E}_{x,y'}\left[\ell(y', h^*_{\mathcal{T}}(x))\right] - \mathbb{E}_{x,y'}\left[\ell(y', h^*(x))\right]\Big] \geq 0,$$

where we set $y'$ to be i.i.d. as $y$ to prevent the dependence of $\mathcal{T}$ on $y$, and

$$h^*_{\mathcal{T}} := \underset{h \in \mathcal{H}}{\arg\min}\, \mathbb{E}_{x,\overline{y_{\mathcal{T}}}}\left[\ell(\overline{y_{\mathcal{T}}}, h(x))\right],$$
$$h^* := \underset{h \in \mathcal{H}}{\arg\min}\, \mathbb{E}_{x,y}\left[\ell(y, h(x))\right],$$

which are the corresponding hypotheses minimizing population losses.

To explicitly show that the "variance" is reduced, we make the following assumption that the marginal output of the teacher model is unbiased:

**Assumption 4.8** (Soft-labeling version of Assumption 3.1)**.** We assume that the teacher model is marginally unbiased. That is,

$$\mathbb{E}_{\mathcal{T}}[\overline{y_{\mathcal{T}}}|x] = y_{\text{oracle}}(x).$$

Here $\mathcal{T}$ is the teacher model trained from the original labels $y$'s. The randomness of $\mathcal{T}$ comes from the randomness of $y$'s and the training procedure (e.g. random seeds).

In general, the learned teacher model outputs $\overline{y_{\mathcal{T}}}$'s are biased estimations of $y_{\text{oracle}}$'s, and the biases are towards the original labels $y$'s (in the extreme case where $\mathcal{T}$ interpolates all the labels, we have $\overline{y_{\mathcal{T}}} = y$). We note that the assumption only requires an unbiasedness in a marginal sense, that those biases cancel out each other in the marginal sense. We have the following result (analogous to Theorem 4.6).

**Theorem 4.9.** *Under Assumption 4.8 and with the cross-entropy loss, we have*

$$\textit{Reduced variance} := \mathrm{Rad}_y(\ell \circ \mathcal{H}) - \mathrm{Rad}_{\overline{y_{\mathcal{T}}}}(\ell \circ \mathcal{H}) \geq 0.$$

In this light, the notion of hierarchical expectation and the reduced Rademacher complexity in Theorem 4.9 render a new bias-variance tradeoff for the knowledge distillation methods. Compared to that of Zhou et al. (2021), our approach theoretically shows that the variance is always reduced by introducing the soft labels, while Zhou et al. (2021) makes the reduction an assumption and verifies it empirically.

## 5. Numerical Experiments

We perform numerical experiments to answer two questions: (1) How do different granularities of the feedback model affect the learning of the reward model? (2) Does the inclusion of these ordinal feedback data with the objective (2) benefit the learning of the reward model?

### 5.1. Experiment Settings

**Datasets.** In the following numerical experiments, we leverage the Skywork-Reward-Preference-80K-v0.2 dataset (Liu et al., 2024a) as our base training dataset. We perform multiple runs and report the average performance (along with the confidence intervals); for each run, we randomly sample a 1024-sized subset as the hold-out evaluation dataset. In addition, we use the RewardBench dataset (Lambert et al., 2024) for the out-of-distribution evaluation task to comprehensively assess the performance of different trained models. More details can be found in Appendix C.3.‘

**Base Models.** Our base models for the following experiments include llama-3.2-1b-instruct (Llama Team, 2024), gemma-2-2b-it (Gemma Team, 2024), and qwen2.5-1.5b-instruct (Yang et al., 2024). All models are trained under

under full-parameter fine-tuning. For the training parameters and other details, we refer to Appendix C.4.

**Ordinal Feedback.** The original Skywork-Reward-Preference-80K-v0.2 dataset only contains a binary feedback for each prompt $x$ and a response pair $y_1$ and $y_2$. To generate feedback labels with different levels of granularities, we adopt a well-trained reward model, Skywork-Reward-Gemma-2-27B-v0.2 (Liu et al., 2024a), as the oracle scoring model $r_{\text{oracle}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ in this case. We chose this model because (1) it was exclusively trained on the to-be-labeled base training dataset, hence there is hardly a risk of out-of-distribution mislabeling; (2) the model ranks first on the RewardBench online leaderboard up to the time of this paper, making its output oracle scores more reliable. Accordingly, the induced oracle model being $z_{\text{oracle}}(x, y_1, y_2) = \sigma((r_{\text{oracle}}(x, y_1) - r_{\text{oracle}}(x, y_2))/T)$ where $T$ is a temperature parameter and $\sigma(\cdot)$ denotes the sigmoid function.

We consider the following four types of feedback systems:

- Oracle: $z_{\text{oracle}}$ is directly used as the feedback label and $\mathcal{Z}_{\text{oracle}} \subset [0, 1]$.
- Binary: $Z_{\text{binary}} \sim \text{Bernoulli}(z_{\text{oracle}})$ and $\mathcal{Z}_2 = \{0, 1\}$.
- 3-level: the label is sampled as the process in Theorem 3.2 considering only the smallest interval containing $z_{\text{oracle}}$ and $\mathcal{Z}_3 = \{0, 0.5, 1\}$.
- 5-level: the label is sampled as the process in Theorem 3.2 considering only the smallest interval containing $z_{\text{oracle}}$ and $\mathcal{Z}_5 = \{0, 0.2, 0.5, 0.8, 1\}$.

We provide a label histogram in Appendix C.3 and adopt the cross-entropy objective function (2) to train the model.

### 5.2. Fine-grained feedback benefits

As discussed earlier, a more fine-grained feedback system should lead to better reward learning. For the four feedback models listed above, they should have the following orders in terms of performance:

$$\text{Oracle “} \geq \text{” 5-level “} \geq \text{” 3-level “} \geq \text{” Binary}$$

where “ $\geq$ ” represents an advantage in model performance.

Here we perform numerical experiments to verify such intuitions and for each combination of the reward model and feedback system, we conduct 5 independent training runs and report the average results. For the setting of learning with oracle feedback, we set $Z_i = \mathbb{P}(y_{i,1} \succ y_{i,2} | x_i) = z_{\text{oracle}}(x_i, y_{i,1}, y_{i,2})$ in the learning objective (2). For more fine-grained ordinal feedback, we sample the feedback according to Section 5.1.

Table 1 (and Figure 2 in Appendix C.1) summarizes the experiment results, which are aligned with the findings in the previous sections. Three take-away messages are: First, a more fine-grained feedback structure leads to better reward learning for both in-distribution (ID) and out-of-distribution (OOD) performance. Second, though we do not have access to the oracle model in practice, the 5-level feedback system provides a good proxy for that. Third, the learning objective (2), as a generalization of the canonical cross-entropy loss for binary feedback, is an effective one to handle the ordinal feedback data.

### 5.3. Tied data benefits

Now we restrict our attention to the 3-level feedback setting and investigate the effect of the proportion of the tied data (samples with labels of $y_1$ “same as” $y_2$). Specifically, we limit the training samples to 32,768 and consider 5 different proportions of the tied data:

- 0%-tied: All the data samples are binary-labeled.
- 25%, 50%, 75% of the data samples are tied.
- 100%-tied: All the data samples are tied.

More details of data generation are deferred to C.3.

Table 2 (and Figure 3 in Appendix C.1) summarizes the experiment results. We make the following observations. First, the 100%-tied setting fails in that it results in a significantly worse performance than the other settings. This is natural as it leads to a reward collapse, as also observed in other semi-supervised learning algorithms; to see this, if we are given only the tied data, one way to learn the reward model is to have all the rewards equal to a constant. Second, mixing a proportion of the tied data and using the learning objective function (2) leads to a better performance than the case of 0%-tied data. One subtle point here is that, in practice, if we do not employ the learning objective (2) and simply drop the tied samples, this will result in a smaller sample size for learning the reward model, and an even worse performance than the 0%-tied setting here.

**Discussions on other ways to deal with tied data.** Recent works (Chen et al., 2024; Liu et al., 2024b) have noticed the importance of incorporating tied samples and employed the Rao-Kupper model (Rao & Kupper, 1967), or the Bradley-Terry model with Ties (BTT) abbreviated by Liu et al. (2024b), for preference modeling and to explore the benefits of leveraging ties. The BTT model represents preference probabilities as follows: $\hat{\mathbb{P}}(y_1 \succ y_2 \mid x) = \frac{e_1}{e_1 + \lambda e_2}$, $\hat{\mathbb{P}}(y_1 \sim y_2 \mid x) = \frac{(\lambda^2 - 1) e_1 e_2}{(e_1 + \lambda e_2)(\lambda e_1 + e_2)}$, where $e_1$ and $e_2$ are, respectively, abbreviations for $\exp\left(r_\theta(x, y_1)\right)$ and $\exp\left(r_\theta(x, y_2)\right)$, $y_1 \sim y_2$ denotes a tie between $y_1$ and $y_2$

*Table 1.* Model convergence statistics under different feedback models. ID stands for in-distribution. OOD stands for out-of-distribution. The ID and OOD datasets are in Section 5.1. The oracle CE loss is computed by adopting $z_{\text{oracle}}$ as $Z_i$ regardless of the feedback type.

| Model | Feedback | Oracle CE Loss | | ID Accuracy | | OOD Accuracy | |
|---|---|---|---|---|---|---|---|
| | | Mean | Std | Mean | Std | Mean | Std |
| Llama | Oracle | 0.5711 | 0.0020 | 0.9382 | 0.0037 | 0.8193 | 0.0016 |
| | 5-level | 0.5714 | 0.0019 | 0.9372 | 0.0040 | 0.8100 | 0.0013 |
| | 3-level | 0.5715 | 0.0021 | 0.9359 | 0.0044 | 0.8016 | 0.0034 |
| | Binary | 0.5736 | 0.0024 | 0.9329 | 0.0044 | 0.7667 | 0.0012 |
| Gemma | Oracle | 0.5698 | 0.0018 | 0.9401 | 0.0031 | 0.8697 | 0.0072 |
| | 5-level | 0.5704 | 0.0016 | 0.9371 | 0.0082 | 0.8584 | 0.0107 |
| | 3-level | 0.5704 | 0.0018 | 0.9381 | 0.0083 | 0.8580 | 0.0016 |
| | Binary | 0.5709 | 0.0021 | 0.9368 | 0.0074 | 0.8237 | 0.0101 |
| Qwen | Oracle | 0.5740 | 0.0026 | 0.9411 | 0.0031 | 0.8543 | 0.0074 |
| | 5-level | 0.5746 | 0.0028 | 0.9394 | 0.0019 | 0.8422 | 0.0080 |
| | 3-level | 0.5749 | 0.0025 | 0.9388 | 0.0039 | 0.8352 | 0.0092 |
| | Binary | 0.5756 | 0.0026 | 0.9378 | 0.0034 | 0.8232 | 0.0102 |

*Table 2.* Model convergence statistics under different tied data ratios. The evaluation dataset remains fixed across all ratio settings and is directly sampled from the original dataset, ensuring its distribution closely matches that of the whole dataset.

| Model | Tied Ratio | Oracle CE Loss | | ID Accuracy | | OOD Accuracy | |
|---|---|---|---|---|---|---|---|
| | | Mean | Std | Mean | Std | Mean | Std |
| Llama | 0% | 1.0421 | 0.0363 | 0.9224 | 0.0080 | 0.7661 | 0.0182 |
| | 25% | 0.3327 | 0.0051 | 0.9341 | 0.0173 | 0.7672 | 0.0093 |
| | 50% | 0.4187 | 0.0043 | 0.9336 | 0.0014 | 0.7545 | 0.0082 |
| | 75% | 0.5339 | 0.0052 | 0.9268 | 0.0180 | 0.7749 | 0.0008 |
| | 100% | 0.6931 | 0.0017 | 0.3428 | 0.0677 | 0.4424 | 0.0393 |
| Gemma | 0% | 6.4762 | 0.3392 | 0.9355 | 0.0041 | 0.8319 | 0.0080 |
| | 25% | 0.6031 | 0.0019 | 0.9467 | 0.0117 | 0.8487 | 0.0100 |
| | 50% | 0.5775 | 0.0001 | 0.9526 | 0.0075 | 0.8277 | 0.0006 |
| | 75% | 0.6122 | 0.0006 | 0.9521 | 0.0069 | 0.8236 | 0.0084 |
| | 100% | 0.6931 | 0.0001 | 0.4814 | 0.0055 | 0.4928 | 0.0158 |
| Qwen | 0% | 4.4457 | 0.1296 | 0.9313 | 0.0010 | 0.8215 | 0.0068 |
| | 25% | 0.6110 | 0.0037 | 0.9339 | 0.0060 | 0.8354 | 0.0046 |
| | 50% | 0.5796 | 0.0031 | 0.9336 | 0.0017 | 0.8267 | 0.0032 |
| | 75% | 0.6136 | 0.0030 | 0.9228 | 0.0093 | 0.8193 | 0.0092 |
| | 100% | 0.6933 | 0.0007 | 0.4342 | 0.0256 | 0.4616 | 0.0356 |

and $\lambda \geq 1$ is a threshold hyperparameter that controls the likelihood of assigning ties. Notably, when $\lambda = 1$, the BTT model reduces to the canonical BT model.

The BTT model may not suit preference learning perfectly as it introduces unnecessary complexity. First, the BTT model introduces an additional hyperparameter $\lambda$ designed for other purposes (for example, predicting a football match). The BTT model introduces a tunable $\lambda$ to output an additional tied probability, which is useful when ties have real-world implications. However, the ultimate goal of the RM training is to provide a reference feedback function for the following RLHF step rather than separately predicting each probability. Second, even if the BTT model is suited for the 3-level feedback case, it is not enough for the richer feedback system that has already been adopted by those LLM companies. Generalizing the BTT model to the more fine-grained case is challenging if the only solution is to introduce more hyperparameters.

We defer more numerical experiments to Appendix C.

## 6. Conclusion

In this paper, we propose reward modeling with ordinal feedback as a generalization of binary feedback. Such a framework fully utilizes the potentially useful samples and the fine-grained information discarded by the binary feedback practice. We generalize the assumption of the BT model to the general marginal unbiasedness assumption named by a sociological concept "wisdom of the crowd". Under the assumption, we build a natural probability model for ordinal feedback. We also show that the Rademacher complexity is reduced by adopting ordinal feedback. Numerical results validate the theoretical findings. Further experiments imply that mixing some tied preference samples benefits RM learning, which may be worth future exploration. Our results suggest that the annotation guideline should encourage the quantitative description (for example, 70%) of the qualitative option (for example, "slightly better"). Our theoretical analysis based on hierarchical expectation may be of independent interest to the field of knowledge distillation, providing a novel bias-variance trade-off perspective.

## Impact Statement

This paper contributes to advancing machine learning by improving reward modeling with ordinal feedback. A key societal implication is the potential to mitigate bias in AI systems. Traditional binary feedback often oversimplifies human preferences, leading to biased or unrepresentative learning outcomes. By incorporating fine-grained ordinal feedback, our approach captures more nuanced human judgments, promoting fairer and more equitable AI decision-making. While our method enhances preference modeling, its effectiveness depends on the diversity and representativeness of annotators, an aspect that warrants further study. Beyond this, our work may have additional societal impacts, though none require specific emphasis at this time.

## References

Akrour, R., Schoenauer, M., and Sebag, M. Preference-based policy learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11*, pp. 12–27. Springer, 2011.

Amini, A., Vieira, T., and Cotterell, R. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Ba, J. and Caruana, R. Do deep nets really need to be deep? *Advances in neural information processing systems*, 27, 2014.

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

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

Burges, C., Ragno, R., and Le, Q. Learning to rank with nonsmooth cost functions. *Advances in neural information processing systems*, 19, 2006.

Chakraborty, S., Qiu, J., Yuan, H., Koppel, A., Huang, F., Manocha, D., Bedi, A. S., and Wang, M. Maxmin-RLHF: Towards equitable alignment of large language models with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

Chen, J., Yang, G., Lin, W., Mei, J., and Byrne, B. On extending direct preference optimization to accommodate ties. *arXiv preprint arXiv:2409.17431*, 2024.

Cheng, W., Fürnkranz, J., Hüllermeier, E., and Park, S.-H. Preference-based policy iteration: Leveraging preference learning for reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11*, pp. 312–327. Springer, 2011.

Choshen, L., Fox, L., Aizenbud, Z., and Abend, O. On the weaknesses of reinforcement learning for neural machine translation. *arXiv preprint arXiv:1907.01752*, 2019.

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

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., and Sun, M. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Davidson, R. R. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329): 317–328, 1970.

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

Dong, Z., Yuan, Y., Hao, J., Ni, F., Mu, Y., Zheng, Y., Hu, Y., Lv, T., Fan, C., and Hu, Z. Aligndiff: Aligning diverse human preferences via behavior-customisable diffusion model. *arXiv preprint arXiv:2310.02054*, 2023b.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Galton, F. Vox populi. *Nature*, 75(1949):450–451, 1907. URL https://doi.org/10.1038/075450a0.

Gemma Team, G. D. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Hinton, G. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., Smith, N. A., and Hajishirzi, H. Rewardbench: Evaluating reward models for language modeling, 2024.

Liu, C. Y., Zeng, L., Liu, J., Yan, R., He, J., Wang, C., Yan, S., Liu, Y., and Zhou, Y. Skywork-reward: Bag of tricks for reward modeling in LLMs. *arXiv preprint arXiv:2410.18451*, 2024a.

Liu, J., Ge, D., and Zhu, R. Reward learning from preference with ties. *arXiv preprint arXiv:2410.05328*, 2024b.

Liu, T., Zhao, Y., Joshi, R., Khalman, M., Saleh, M., Liu, P. J., and Liu, J. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023.

Liu, T., Qin, Z., Wu, J., Shen, J., Khalman, M., Joshi, R., Zhao, Y., Saleh, M., Baumgartner, S., Liu, J., et al. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*, 2024c.

Llama Team, A. . M. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

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

Pareek, D., Du, S. S., and Oh, S. Understanding the gains from repeated self-distillation. *arXiv preprint arXiv:2407.04600*, 2024.

Phuong, M. and Lampert, C. Towards understanding knowledge distillation. In *International conference on machine learning*, pp. 5142–5151. PMLR, 2019.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization:

Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Rao, P. and Kupper, L. L. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317): 194–204, 1967.

Schölkopf, B., Tsuda, K., and Vert, J.-P. *Kernel methods in computational biology*. MIT press, 2004.

Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., and Wang, H. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38(17), pp. 18990–18998, 2024.

Wang, C., Jiang, Y., Yang, C., Liu, H., and Chen, Y. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*, 2023.

Wang, H., Lin, Y., Xiong, W., Yang, R., Diao, S., Qiu, S., Zhao, H., and Zhang, T. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*, 2024a.

Wang, H., Xiong, W., Xie, T., Zhao, H., and Zhang, T. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024b.

Wang, Z., Dong, Y., Delalleau, O., Zeng, J., Shen, G., Egert, D., Zhang, J. J., Sreedhar, M. N., and Kuchaiev, O. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024c.

Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Van Durme, B., Murray, K., and Kim, Y. J. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Yuan, H., Yuan, Z., Tan, C., Wang, W., Huang, S., and Huang, F. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Yuan, L., Tay, F. E., Li, G., Wang, T., and Feng, J. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3903–3911, 2020.

Zeng, D., Dai, Y., Cheng, P., Wang, L., Hu, T., Chen, W., Du, N., and Xu, Z. On diversified preferences of large language model alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9194–9210, 2024a.

Zeng, Y., Liu, G., Ma, W., Yang, N., Zhang, H., and Wang, J. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024b.

Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M., and Liu, P. J. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Zhou, H., Song, L., Chen, J., Zhou, Y., Wang, G., Yuan, J., and Zhang, Q. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. *arXiv preprint arXiv:2102.00650*, 2021.

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

# A. Related Works

Reinforcement learning from human feedback (RLHF) originates from the idea of preference-based reinforcement learning (Cheng et al., 2011; Akrour et al., 2011). The term RLHF is proposed by the large language model (LLM) community and has been a mainstream framework for aligning LLMs with human preferences (Askell et al., 2021; Ouyang et al., 2022). For a more detailed survey of the history of RLHF, we refer to Kaufmann et al. (2023). While many people have incorporated the supervised fine-tuning (SFT) stage as a part of the RLHF pipeline (Ziegler et al., 2019; Ouyang et al., 2022; Ji et al., 2023), our discussion would focus on the reward modeling and policy optimization via reinforcement learning (RL), which takes a supervised fine-tuned model as the starting point.

Although proven effective for aligning LLMs with human preferences, the canonical RLHF suffers in several aspects, including complicated implementation, difficult hyperparameter tuning, low sampling efficiency (Choshen et al., 2019), and computational overhead (Yuan et al., 2024), which promotes the studies to optimize with the relative preferences without depending on RL. One major alternative to the RLHF is the direct policy optimization (DPO) proposed by Rafailov et al. (2024), which directly trains the language model to increase the probability of the preferred response and decrease the other. Wang et al. (2023) discusses the influence of $f$-divergence as the constraint term in PPO and proposes $f$-DPO. Azar et al. (2024) considers a general learning objective under pair-wise preferences, i.e., $\Psi$PO, points out the potential overfitting issue in RLHF and PPO, and further mitigates the problem with a specific instance of $\Psi$PO, i.e., IPO. Zeng et al. (2024b) investigates the optimization at a finer level and employs forward KL divergence constraints for each token to improve the alignment. There are also some methods (Xu et al., 2024; Ethayarajh et al., 2024) that try to skip the SFT stage to lower the costs and mitigate the issues while directly imitating the reference data. Liu et al. (2023) uses statistical rejection sampling to address the mismatch between the training data and optimal policy data, hence enhancing the preference data collection. Amini et al. (2024) proposes DPO with an offset (ODPO) that takes the preference strength into consideration, which forces the language model to separate the probabilities of two responses in the training dataset by an offset. Compared to our work, the authors do not theoretically prove the benefits of considering the reference strength, and their proposal requires tuning the offset.

Parallel to previously mentioned efforts to overcome the dependence on RL, researchers also directly optimize the policy under different loss functions with different types of preference data. We list some efforts non-exhaustively here, including RAFT (Dong et al., 2023a), SLiC (Zhao et al., 2023), LiPO (Liu et al., 2024c), RRHF (Yuan et al., 2024), and PRO (Song et al., 2024). For pairwise comparison data, Zhao et al. (2023) proposes a hinge-type loss to encourage the LLM to output the chosen sequence more likely than the rejected one. For the list data that compares many different responses, Dong et al. (2023a) selects the best response and fine-tunes the LLM on the best-of-$K$ data. Yuan et al. (2024) adopts the ranking loss in the same spirit of increasing the preferred sequence's likelihood and summing up all the pairwise comparisons. Song et al. (2024) replaces the pairwise comparison with the preferred one against the remaining responses and recursively iterates from the most preferred one to the second-least preferred. All those ranking-based methods are purely based on the relative positions without considering the preference strength, and forcing a rank among nearly tied responses introduces additional noise. To mitigate this issue, Liu et al. (2024c) applies the LambdaLoss (Burges et al., 2006) to build a LiPO-$\lambda$ objective that takes the preference strength into the weights of the ranking loss. However, the proposed method has two drawbacks. First, the loss is heuristically defined and lacks theoretical guarantees. Second, the requirement of labeling each response with a quantitative reward is not easy for human annotators: on the contrary, the initial motivation behind the RLHF method is to train a reward model by pure *comparison* data to avoid asking human annotators to quantify a *reward* precisely. In contrast, our ordinal feedback only requires the human annotators to calibrate the qualitative *comparison* with a quantity.

With the rapid development of optimization frameworks, some scholars notice the issue of diverse preferences in reward modeling, not only in the field of LLM (Dong et al., 2023b). Zeng et al. (2024a) proposes a multi-objective reward learning method (MORE) to calibrate the reward models with the shared preferences and enhance alignment performance. Chakraborty et al. (2024) introduces a MaxMin alignment objective to learn a mixture of diverse preference mixtures and greatly improve the overall performance. Wang et al. (2024a) employs the multi-objective reward modeling and models the preferences of users, implementing better objective control. Wang et al. (2024b) enhances the interpretability and performance of reward models by combining multi-objective reward modeling and mixture-of-experts techniques. Different from these works, our method considers a more practical scenario without requiring multiple labels for preference pairs and constructs the probability model for the single-objective preference learning.

Two recent works (Chen et al., 2024; Liu et al., 2024b) incorporate the tied samples into the learning of the reward model by considering some generalized versions (Rao & Kupper, 1967; Davidson, 1970) of the Bradley-Terry model (Bradley

& Terry, 1952). By introducing a more complicated model, they enable the reward model to predict three probabilities: better, worse, and tied. The authors directly use the cross-entropy loss for the trinary classification for the RM learning. As a comparison, our ordinal feedback is more general and not limited to the Bradley-Terry model and the 3-level feedback. We also avoid introducing additional hyperparameters and keep the original training paradigm.

## B. Proofs and Theoretical Discussions

### B.1. Proof of Theorem 3.2

*Proof.* We first prove that the constructed ordinal feedback $Z \sim \mu_{j,k}$ satisfies Assumption 3.1, where

$$\mu_{j,k}(z) = \begin{cases} (z_k - z_{\text{oracle}})/(z_k - z_j), & \text{if } z = z_j, \\ (z_{\text{oracle}} - z_j)/(z_k - z_j), & \text{if } z = z_k, \\ 0, & \text{otherwise.} \end{cases}$$

If $z_{\text{oracle}} = z_i$ for any $z_i \in \mathcal{Z}$, then the constructed measure is the Dirac measure for $z_{\text{oracle}}$, which automatically fulfills the requirement.

We consider the cases where $z_{\text{oracle}} \in (z_j, z_k)$ for some $z_j, z_k \in \mathcal{Z}$. Then the expectation of $Z$ w.r.t. $\mu_{j,k}$ is

$$\begin{aligned} \mathbb{E}_{Z \sim \mu_{j,k}}[Z] &= \mu_{j,k}(z_j) \cdot z_j + \mu_{j,k}(z_k) \cdot z_k \\ &= (z_j z_k - z_{\text{oracle}} z_j + z_{\text{oracle}} z_k - z_j z_k)/(z_k - z_j) \\ &= z_{\text{oracle}}. \end{aligned}$$

For the second part of Theorem 3.2, we prove the conclusion case by case. Suppose we have an ordinal feedback $Z \colon \Omega \to \mathcal{Z}$ satisfying Assumption 3.1. If $z_{\text{oracle}} \in \mathcal{Z}$, for example, $z_{\text{oracle}} = z_{i_0}$, then

$$\mu_{j,i_0} = \mu_{i_0,k} = \delta_{z_{i_0}},$$

implying we can set those coefficients $\alpha_{j,i_0}$'s and $\alpha_{i_0,k}$'s to be arbitrary non-negative real numbers such that $\sum_j \alpha_{j,i_0} + \sum_k \alpha_{i_0,k} = \mu(z_{i_0})$ without affecting the measure on the other points.

Therefore, we can (without loss of generality) assume $z_{\text{oracle}} \notin \mathcal{Z}$. Assume $z_1 < \cdots < z_{m_1} < z_{\text{oracle}} < z_{m_1+1} < \cdots < z_m$.

Case (1): when there is only one point on each side of $z_{\text{oracle}}$, that is, $m_1 = 1$ and $m_1 + 1 = m$.
Then by Assumption 3.1, we have

$$\mu(z_1) \cdot z_1 + \mu(z_m) \cdot z_m = z_{\text{oracle}}.$$

Combining it with the constraint such that

$$\mu(z_1) + \mu(z_m) = 1,$$

we have

$$\mu(z_1) = (z_m - z_{\text{oracle}})/(z_m - z_1), \quad \mu(z_m) = (z_{\text{oracle}} - z_1)/(z_m - z_1),$$

which is exactly the same as $\mu_{1,m}$.

Case (2): when there is only one point larger than $z_{\text{oracle}}$, that is, $m_1 + 1 = m$.
Then we will prove that there exist non-negative real numbers $\alpha_{j,m}$ such that

$$\mu = \sum_j \alpha_{j,m}\, \mu_{j,m}.$$

In fact, we can construct (for any $j \neq m$)

$$\alpha_{j,m} := \frac{\mu(z_j)}{\mu_{j,m}(z_j)}.$$

Then each $\alpha_{j,m}$ is non-negative. Furthermore, by Assumption 3.1,

$$\begin{aligned}\sum_{j\neq m}\alpha_{j,m}\mu_{j,m}(z_m) &= \sum_{j\neq m}\frac{\mu(z_j)}{\mu_{j,m}(z_j)}\cdot\mu_{j,m}(z_m)\\ &= \frac{\sum_{j\neq m}\mu(z_j)\cdot z_{\text{oracle}}-\sum_{j\neq m}\mu(z_j)\cdot z_j}{z_m - z_{\text{oracle}}}\\ &= \frac{(1-\mu(z_m))\cdot z_{\text{oracle}}-(z_{\text{oracle}}-\mu(z_m)\cdot z_m)}{z_m-z_{\text{oracle}}}\\ &= \mu(z_m),\end{aligned}$$

indicating that $\mu=\sum_{j\neq m}\alpha_{j,m}\mu_{j,m}$.
By the property of probability measures, we can easily see that

$$\sum_{j\neq m}\alpha_{j,m}=1.$$

Case (3): when there is only one point smaller than $z_{\text{oracle}}$, that is, $m_1=1$. This case can be proved similarly to Case (2).

Case (4): general cases where there are (possibly) multiple points on each side of $z_{\text{oracle}}$. We prove it by induction. Suppose there are $a$ elements in $\mathcal{Z}$ smaller than $z_{\text{oracle}}$ and $b$ elements larger than $z_{\text{oracle}}$. Denote the case by $(a,b)$. Suppose that the conclusion has been proved for all the cases $(a,b)$ if $a<m_1$ or $b<m-m_1$. Now we prove it for the case $(m_1,m-m_1)$. We first define the corresponding index (if there are multiple elements in a tie, select arbitrarily)

$$i_1 \coloneqq \arg\min_i |z_i - z_{\text{oracle}}|\cdot\mu(z_i). \tag{5}$$

Without loss of generality, we assume $z_{i_1}<z_{\text{oracle}}$. Then we have $m_1>1$ due to Assumption 3.1. We now construct a coefficient $\alpha_{i_1,m_1+1}$ such that

$$\alpha_{i_1,m_1+1}\coloneqq\frac{\mu(z_{i_1})}{\mu_{i_1,m_1+1}(z_{i_1})}.$$

By the definition (5), we have

$$\mu(z_{m_1+1})\geq\alpha_{i_1,m_1+1}\cdot\mu_{i_1,m_1+1}(z_{m_1+1}).$$

Hence we can construct a new measure

$$\mu'(z_i)=\begin{cases}0, & \text{if } i=i_1;\\ \big(\mu(z_{m_1+1})-\alpha_{i_1,m_1+1}\cdot\mu_{i_1,m_1+1}(z_{m_1+1})\big)\big/(1-\alpha_{i_1,m_1+1}), & \text{if } i=m_1+1;\\ \mu(z_i)\big/(1-\alpha_{i_1,m_1+1}), & \text{otherwise.}\end{cases}$$

This measure can be easily verified as a probability measure with $m_1-1$ elements smaller than $z_{\text{oracle}}$. By induction hypothesis, we can construct non-negative real numbers $\alpha_{j,k}$'s summing up to 1 such that

$$\mu'=\sum_{j,k}\alpha'_{j,k}\cdot\mu_{j,k}.$$

Then we have

$$\mu=\alpha_{i_1,m_1+1}\cdot\mu_{i_1,m_1+1}+\sum_{j,k}\frac{\alpha'_{j,k}}{1-\alpha_{i_1,m_1+1}}\cdot\mu_{j,k},$$

of which the coefficients are non-negative and sum to 1. $\square$

### B.2. Validating (3)

**Cross-entropy loss:**

The cross-entropy loss is

$$\ell_{\text{ce}}(Z, z) = -\left[Z \log(z) + (1 - Z)\log(1 - z)\right],$$

where $z \in [0, 1]$, which is the probability under the BT model. The cross-entropy loss is affine to $Z$:

$$\begin{aligned}
\mathbb{E}_Z\big[\ell_{\text{ce}}(Z, z)\big] &= -\sum_{j, z_j \in \mathcal{Z}} \mathbb{P}(Z = z_j) \cdot [z_j \cdot \log(z) + (1 - z_j) \cdot \log(1 - z)] \\
&= -\left[\Big(\sum_{j, z_j \in \mathcal{Z}} \mathbb{P}(Z = z_j) \cdot z_j\Big) \cdot \log(z) + \bigg(1 - \Big(\sum_{j, z_j \in \mathcal{Z}} \mathbb{P}(Z = z_j) \cdot z_j\Big)\bigg) \cdot \log(1 - z)\right] \\
&= -\big[\mathbb{E}[Z] \cdot \log(z) + \big(1 - \mathbb{E}[Z]\big) \cdot \log(1 - z)\big] \\
&= \ell_{\text{ce}}\big(\mathbb{E}[Z], z\big).
\end{aligned}$$

**Generalized hinge loss:**

The hinge loss sets the loss to be

$$\ell_{\text{hinge}}(Z, z) = \mathbf{1}\{Z = 1\} \cdot \max\big(C - z, 0\big) + \mathbf{1}\{Z = 0\} \cdot \max\big(C + z, 0\big),$$

where $z \in \mathbb{R}$, which is the difference of the reward functions $r_\theta(x, y_1) - r_\theta(x, y_2)$. We can generalize the hinge loss as

$$\ell_{\text{hinge}}(Z, z) \coloneqq Z \cdot \max\big(C - z, 0\big) + (1 - Z) \cdot \max\big(C + z, 0\big)$$

The (generalized) hinge loss is affine to $Z$ by a similar argument to the cross-entropy loss:

$$\begin{aligned}
&\mathbb{E}_Z\big[\ell_{\text{hinge}}(Z, z)\big] \\
&= \sum_{j, z_j \in \mathcal{Z}} \mathbb{P}(Z = z_j) \cdot \big[z_j \cdot \max\big(C + z, 0\big) + (1 - z_j) \cdot \max\big(C - z, 0\big)\big] \\
&= \Big(\sum_{j, z_j \in \mathcal{Z}} \mathbb{P}(Z = z_j) \cdot z_j\Big) \cdot \max\big(C + z, 0\big) + \bigg(1 - \Big(\sum_{j, z_j \in \mathcal{Z}} \mathbb{P}(Z = z_j) \cdot z_j\Big)\bigg) \cdot \max\big(C - z, 0\big) \\
&= \mathbb{E}[Z] \cdot \max\big(C + z, 0\big) + \big(1 - \mathbb{E}[Z]\big) \cdot \max\big(C - z, 0\big) \\
&= \ell_{\text{hinge}}\big(\mathbb{E}[Z], z\big).
\end{aligned}$$

### B.3. Validating (4)

We can prove a stronger conclusion such that for any $(x, y_1, y_2) \in \mathcal{X} \times \mathcal{Y}^2$,

$$\mathbb{E}_Z\Big[\ell\big(Z, h(x, y_1, y_2)\big)\Big|(x, y_1, y_2)\Big] = \mathbb{E}_{Z'}\Big[\ell\big(Z', h(x, y_1, y_2)\big)\Big|(x, y_1, y_2)\Big].$$

Since both $Z$ and $Z'$ satisfy Assumption 3.1, we have for any $(x, y_1, y_2) \in \mathcal{X} \times \mathcal{Y}^2$,

$$\mathbb{E}[Z|(x, y_1, y_2)] = z_{\text{oracle}}(x, y_1, y_2) = \mathbb{E}[Z'|(x, y_1, y_2)].$$

Then by the affinity condition (3), we have for any $h \in \mathcal{H}$,

$$\begin{aligned}
&\mathbb{E}_Z\Big[\ell\big(Z, h(x, y_1, y_2)\big)\Big|(x, y_1, y_2)\Big] \\
&= \ell\big(\mathbb{E}[Z|(x, y_1, y_2)], h(x, y_1, y_2)\big) \\
&= \ell\big(z_{\text{oracle}}(x, y_1, y_2), h(x, y_1, y_2)\big).
\end{aligned}$$

The same arguments also lead to that

$$\mathbb{E}_{Z'}\Big[\ell\big(Z', h(x, y_1, y_2)\big)\Big|(x, y_1, y_2)\Big] = \ell\big(z_{\text{oracle}}(x, y_1, y_2), h(x, y_1, y_2)\big),$$

which concludes the proof.

### B.4. Proof of Proposition 4.3

*Proof.* We construct $(W, W')$ as follows: we set $W$ to be identically distributed as $Z$, and

$$\tilde{\mathbb{P}}(W' = z'_k | W = z_j) \coloneqq \beta_{j,k}.$$

Then by property (a) of $\beta_{j,k}$'s, we have

$$\mathbb{E}[W'|W = z_j] = z_j,$$

indicating

$$W = \mathbb{E}[W'|W].$$

By property (b) of $\beta_{j,k}$'s, we have that the constructed $W'$ has a marginal distribution identical to $Z'$. Hence, $(W, W')$ is a coupling satisfying the hierarchical expectation requirements.

On the other hand, if we have a coupling $(W, W')$ satisfying the hierarchical expectation condition, we can easily verify that the conditional probabilities satisfy the requirements in Proposition 4.3. $\square$

### B.5. Proof of Corollary 4.4

*Proof.* For case (a) such that $Z = z_{\text{oracle}} =: z_1$ almost surely, we have $\mu = \delta_{z_{\text{oracle}}}$. By Assumption 3.1, the probability measure $\mu'$ of $Z'$ satisfies

$$\sum_{k, z'_k \in \mathcal{Z}'} \mu'(z'_k)\, z'_k = z_{\text{oracle}}$$

almost surely. Setting $\beta_{1,k} \coloneqq \mu'(z'_k)$ fulfills the properties in Proposition 4.3.

For case (b) such that $\mathcal{Z}' = \{z'_1 \coloneqq 0, z'_2 \coloneqq 1\}$, we can construct $\beta_{j,k}$'s as

$$\beta_{j,1} = 1 - z_j, \quad \beta_{j,2} = z_j,$$

for any $z_j \in \mathcal{Z}$. Then one can easily verify that the construction satisfies the requirement (a) in Proposition 4.3. For part (b), by Assumption 3.1, we have

$$\sum_{j, z_j \in \mathcal{Z}} \alpha_j z_j = z_{\text{oracle}},$$

and

$$\alpha'_2 = z_{\text{oracle}}.$$

Combining the above two equalities, we have

$$\sum_{j, z_j \in \mathcal{Z}} \alpha_j \beta_{j,2} = \alpha'_2.$$

On the other hand,

$$\sum_{j, z_j \in \mathcal{Z}} \alpha_j (1 - z_j) = 1 - z_{\text{oracle}},$$

and

$$\alpha'_1 = 1 - z_{\text{oracle}},$$

which implies that

$$\sum_{j, z_j \in \mathcal{Z}} \alpha_j \beta_{j,1} = \alpha'_1.$$

$\square$

### B.6. Proof of Theorem 4.6

**Lemma B.1.** *Any affine function is also convex.*

**Lemma B.2.** *The pointwise supremum of a family of convex functions is still convex. In other words, for any family of convex functions $f_s(\cdot)$ where $s \in S$, we have $\sup_{s\in S} f_s(\cdot)$ still being convex.*

We provide those two lemmas without proof since the proof can be found in any convex analysis textbook. A corollary is the following.

**Lemma B.3.** *If the loss function satisfies the affinity condition (3), then for any hypothesis class $\mathcal{H}$, the following function*

$$\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \varepsilon_i \ell\big(\cdot, h(x_i, y_{i,1}, y_{i,2})\big)$$

*is convex for any realization of $\varepsilon_i$ taking values in $\{+1, -1\}$.*

*Proof of Lemma B.3.* The argument breaks up into three pieces: first, any linear combination of affine functions is still affine (which is straightforward), hence

$$\sum_{i=1}^{n} \varepsilon_i \ell\big(\cdot, h(x_i, y_{i,1}, y_{i,2})\big)$$

is affine as long as condition (3) holds.

Second, any affine function is also convex (Lemma B.1), therefore,

$$\sum_{i=1}^{n} \varepsilon_i \ell\big(\cdot, h(x_i, y_{i,1}, y_{i,2})\big)$$

is also convex for any $h \in \mathcal{H}$.

Third, the supremum of any class of convex functions is still convex (Lemma B.2), which means taking the supremum over the hypothesis class $\mathcal{H}$ suffices. □

*Proof of Theorem 4.6.* We denote the hierarchical expectation coupling of $Z$ and $Z'$ by $(W, W')$. By direct inspection, we have

$$\begin{aligned}
\mathrm{Rad}_{\mathcal{Z}',n}(\ell \circ \mathcal{H}) &= \frac{1}{n}\mathbb{E}_{x,y,Z',\varepsilon}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \varepsilon_i \ell\big(Z_i', h(x_i, y_{i,1}, y_{i,2})\big)\right] \\
&= \frac{1}{n}\mathbb{E}_{x,y,W,W',\varepsilon}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \varepsilon_i \ell\big(W_i', h(x_i, y_{i,1}, y_{i,2})\big)\right] \\
&= \frac{1}{n}\mathbb{E}_{x,y,\varepsilon}\left[\mathbb{E}_{W,W'}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \varepsilon_i \ell\big(W_i', h(x_i, y_{i,1}, y_{i,2})\big)\right]\right] \\
&= \frac{1}{n}\mathbb{E}_{x,y,\varepsilon}\left[\mathbb{E}_{W}\left[\mathbb{E}_{W'}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \varepsilon_i \ell\big(W_i', h(x_i, y_{i,1}, y_{i,2})\big)\middle| W\right]\right]\right] \\
&\geq \frac{1}{n}\mathbb{E}_{x,y,\varepsilon}\left[\mathbb{E}_{W}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \varepsilon_i \ell\big(\mathbb{E}_{W'}\left[W_i'|W\right], h(x_i, y_{i,1}, y_{i,2})\big)\right]\right] \\
&= \frac{1}{n}\mathbb{E}_{x,y,\varepsilon}\left[\mathbb{E}_{W}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \varepsilon_i \ell\big(W, h(x_i, y_{i,1}, y_{i,2})\big)\right]\right] \\
&= \frac{1}{n}\mathbb{E}_{x,y,\varepsilon}\left[\mathbb{E}_{Z}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \varepsilon_i \ell\big(Z, h(x_i, y_{i,1}, y_{i,2})\big)\right]\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\mathbb{E}_{x,y,Z,\varepsilon}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\varepsilon_i \ell\big(Z, h(x_i, y_{i,1}, y_{i,2})\big)\right] \\
&= \mathrm{Rad}_{\mathcal{Z},n}(\ell\circ\mathcal{H}),
\end{aligned}
$$

where the first equality is by definition, the second equality is because the coupling's marginal distribution on $W'$ is equal to that of $Z'$, the third equality is due to the exchangeability of the order of integration (by Fubini's Theorem), the fourth equality is because of the tower property of the conditional expectation, the first inequality is the result of Lemma B.3 and Jensen's inequality, the fifth equality is owing to the definition of the hierarchical expectation (Definition 4.2), the sixth is on account of the property of the coupling again, the seventh thanks to Fubini's Theorem again, and the last equality is the definition of Rademacher complexity again. $\square$

### B.7. Proof of Corollary 4.7

*Proof.* A straightforward conclusion of Theorem 4.6 and Corollary 4.4. $\square$

### B.8. Proof of Theorem 4.9

*Proof.* Denote the one-hot vector at dimension $j$ as $e_j$. The following arguments are made for any $x \in \mathcal{X}$, and we omit the dependence on $x$ for notation simplicity.

We only need to show that $\overline{y_{\mathcal{T}}}$ is a hierarchical expectation of $y$ and the theorem is the result of Theorem 4.6. We construct the coupling $(w, w')$ as follows:

$$w \coloneqq \overline{y_{\mathcal{T}}},$$

and

$$\mathbb{P}(w' = e_j | w) \coloneqq w_j,$$

where $w_j$ denotes the $j$-th entry of $w$. We now verify that $(w, w')$ is a coupling of $(\overline{y_{\mathcal{T}}}, y)$. The fact that $w$ and $\overline{y_{\mathcal{T}}}$ have the same distribution is easy. For $w'$ and $y$, we have

$$
\begin{aligned}
\mathbb{P}(w' = e_j) &= \mathbb{E}[\mathbb{P}(w' = e_j|w)] \\
&= \mathbb{E}[w_j] \\
&= \mathbb{E}[\bar{y}_{\mathcal{T},j}] \\
&= y_{\text{oracle},j} \\
&= \mathbb{P}(y = e_j),
\end{aligned}
$$

where $\bar{y}_{\mathcal{T},j}$ (or $y_{\text{oracle},j}$) denotes the $j$-th entry of $\overline{y_{\mathcal{T}}}$ (or $y_{\text{oracle}}$), and the dependence on $x$ has been omitted. Here, the first equality is the tower property of the conditional expectation, the second due to the definition of $w'$, the third because of the construction of $w$, the fourth on account of Assumption 4.8, and the last is the definition of $y_{\text{oracle}}$.

Hence $(w, w')$ is a coupling of $(\overline{y_{\mathcal{T}}}, y)$.

We combine that conclusion with the fact that

$$\mathbb{E}[w'|w] = w,$$

leading to the conclusion that $\overline{y_{\mathcal{T}}}$ is a hierarchical expectation of $y$. $\square$

### B.9. Generalizations to DPO

Our results naturally extend to the direct policy optimization (DPO) (Rafailov et al., 2024) case. We first give a quick introduction to the DPO training objective. Suppose the ground-truth reward function for any prompt-response pair $(x, y)$ is $r^*(x, y)$. Then, under the reinforcement learning objective of maximizing the reward (with a Kullback-Leibler divergence regularization of strength $\beta$ from the original policy $\pi_{\text{ref}}$), the optimal policy under the ground truth reward function $r^*$ should be

$$\pi^*(y|x) \propto \pi_{\text{ref}}(y|x)\cdot\exp\left(\frac{1}{\beta}r^*(x, y)\right).$$

Under the Bradley-Terry model, we have

$$\mathbb{P}(y_1 \succ y_2|x) = \sigma\left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} - \beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right),$$

where the $\sigma(\cdot)$ is the sigmoid function $\sigma(x) = \exp(x)/(1+\exp(x))$. Then the DPO training objective is to minimize the cross-entropy loss under the binary feedback setting

$$\begin{aligned}\min_\theta \sum_{i=1}^n - Z_i \cdot \log\Bigg(\sigma\Bigg(\beta \log \frac{\pi_\theta(y_{i,1}|x_i)}{\pi_{\text{ref}}(y_{i,1}|x_i)} \\ - \beta \log \frac{\pi_\theta(y_{i,2}|x_i)}{\pi_{\text{ref}}(y_{i,2}|x_i)}\Bigg)\Bigg) \\ - (1-Z_i) \cdot \log\Bigg(\sigma\Bigg(\beta \log \frac{\pi_\theta(y_{i,2}|x_i)}{\pi_{\text{ref}}(y_{i,2}|x_i)} \\ - \beta \log \frac{\pi_\theta(y_{i,1}|x_i)}{\pi_{\text{ref}}(y_{i,1}|x_i)}\Bigg)\Bigg). \end{aligned} \tag{6}$$

By considering a richer feedback than the binary case $\mathcal{Z} = \{0, 1\}$, we can use (6) to train the LLM $\pi_\theta$ directly under the ordinal feedback. The affinity condition (3) is fulfilled since the loss is still the cross-entropy loss. Thus, applying Theorem 4.6 for $\Pi = \{\pi_\theta, \theta \in \Theta\}$ yields a similar result. We present here without repeating the proof.

**Corollary B.4.** *Consider the corresponding loss function $\ell$ in the training objective (6) and a policy class $\Pi = \{\pi_\theta, \theta \in \Theta\}$. For any two ordinal feedback systems $Z$ and $Z'$ taking values in $\mathcal{Z}$ and $\mathcal{Z}'$ such that $Z$ is a hierarchical expectation of $Z'$, we have*

$$\mathrm{Rad}_{\mathcal{Z},n}(\ell \circ \Pi) \le \mathrm{Rad}_{\mathcal{Z}',n}(\ell \circ \Pi).$$

### B.10. Generalization bound under Rademacher complexity

The following proposition is a well-known generalization bound. We present it here only for the completeness of our argument as the proof can be found in any statistical learning lecture notes.

**Proposition B.5** (Generalization Bound)**.** *Suppose we have $\mathcal{D}_{\mathcal{Z}}$ as a dataset consisting of $n$ i.i.d. samples. For any hypothesis class $\mathcal{H}$, we have with probability at least $1-\delta$ that for every function $h \in \mathcal{H}$,*

$$\mathbb{E}\Big[\ell\big(Z_i, h(x_i, y_{i,1}, y_{i,2})\big)\Big] \le \hat{\mathbb{E}}_{\mathcal{D}_{\mathcal{Z}}}\Big[\ell\big(Z_i, h(x_i, y_{i,1}, y_{i,2})\big)\Big] + 2\mathrm{Rad}_{\mathcal{Z},n}(\ell \circ \mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{n}}.$$

### B.11. Lower bounds on reduction of Rademacher complexity

We take the cross-entropy loss as an example. Suppose we have two ordinal feedbacks $Z$ and $Z'$, taking values in $\mathcal{Z}$ and $\mathcal{Z}'$, respectively. Suppose $Z$ is a hierarchical expectation of $Z'$, of which the coupling is formulated as $(W, W')$. We make further assumptions here besides the "wisdom of the crowd" assumption (Assumption 3.1).

**Assumption B.6.** Assume that each $Z_i$ (or $Z_i'$) is independent. Furthermore, assume that each $z_j \in \mathcal{Z}$ is bounded away from 0 and 1 by

$$\gamma < z_j < 1 - \gamma.$$

The above assumption is not restrictive: the independence naturally holds if each annotation is independent. The boundedness can be satisfied by clipping the feedback options, and is assumed to prevent those infinite values attained in the supremum step due to logarithms (if $\varepsilon = +1$) while making sure those maxima can be obtained if $\varepsilon = -1$.

**Assumption B.7.** Assume that the function class $\mathcal{H}$ is rich enough (w.r.t. $\gamma$) with high probability ($\ge 1-\delta$). That is, $\exists \delta \ge 0$, s.t.

$$E \coloneqq \big\{[\gamma, 1-\gamma]^n \in \big\{\big(h(x_1, y_{1,1}, y_{1,2}), \dots, h(x_n, y_{n,1}, y_{n,2})\big)\big| h \in \mathcal{H}\big\}\big\}$$

holds with probability at least $1-\delta$.

We denote the number of positive Rademacher variables (that is, those $\varepsilon_i = +1$) by $\tau$ (which is a random variable itself). From the definition of $\tau$, we know

$$\mathbb{E}_\varepsilon[\tau] = \frac{n}{2}. \tag{7}$$

Without loss of generality, we assume that $\varepsilon_1 = \cdots = \varepsilon_\tau = +1$, and $\varepsilon_{\tau+1} = \cdots = \varepsilon_n = -1$. Due to the assumption that each $(x_i, y_{i,1}, y_{i,2}, Z_i)$ is i.i.d. (or $Z_i'$), such a $\tau$ summarizes all the dependence on $\varepsilon$. The two Rademacher complexities are (omitting the dependence on $x, y_1, y_2$ for notational simplicity and denoting $h_i \coloneqq h(x_i, y_{i,1}, y_{i,2})$)

$$\begin{aligned}\mathrm{Rad}_{\mathcal{Z},n}(\ell \circ \mathcal{H}) = \frac{1}{n}\mathbb{E}_{\tau,W}\Bigg[\sup_{h\in\mathcal{H}} \ &\sum_{i=1}^{\tau} -W_i \cdot \log(h_i) - (1-W_i)\cdot\log(1-h_i) \\ &+ \sum_{i=\tau+1}^{n} W_i \cdot \log(h_i) + (1-W_i)\cdot\log(1-h_i)\Bigg],\end{aligned}$$

and

$$\begin{aligned}\mathrm{Rad}_{\mathcal{Z}',n}(\ell \circ \mathcal{H}) = \frac{1}{n}\mathbb{E}_{\tau,W}\Bigg[\mathbb{E}_{W'}\Bigg[\sup_{h\in\mathcal{H}} \ &\sum_{i=1}^{\tau} -W_i' \cdot \log(h_i) - (1-W_i')\cdot\log(1-h_i) \\ &+ \sum_{i=\tau+1}^{n} W_i' \cdot \log(h_i) + (1-W_i')\cdot\log(1-h_i)\Bigg| W\Bigg]\Bigg].\end{aligned}$$

**Lemma B.8.** *Under Assumptions 4.8 and B.6, if $E$ holds, then the following holds (here, $q_i$'s are real numbers independent of $h_i$'s):*

$$\begin{aligned}&\sup_{h\in\mathcal{H}} \sum_{i=1}^{\tau} -W_i \cdot \log(h_i) - (1-W_i)\cdot\log(1-h_i) + \sum_{i=\tau+1}^{n} W_i \cdot \log(h_i) + (1-W_i)\cdot\log(1-h_i) \\ &= \sup_{h\in\mathcal{H}} \sum_{i=1}^{\tau} -W_i \cdot \log(h_i) - (1-W_i)\cdot\log(1-h_i) + \sup_{q_i\in[\gamma,1-\gamma]} \sum_{i=\tau+1}^{n} W_i \cdot \log(q_i) + (1-W_i)\cdot\log(1-q_i),\end{aligned}$$

*and*

$$\begin{aligned}&\sup_{h\in\mathcal{H}} \sum_{i=1}^{\tau} -W_i' \cdot \log(h_i) - (1-W_i')\cdot\log(1-h_i) + \sum_{i=\tau+1}^{n} W_i' \cdot \log(h_i) + (1-W_i')\cdot\log(1-h_i) \\ &\geq \sup_{h\in\mathcal{H}} \sum_{i=1}^{\tau} -W_i' \cdot \log(h_i) - (1-W_i')\cdot\log(1-h_i) + \sup_{q_i\in[\gamma,1-\gamma]} \sum_{i=\tau+1}^{n} W_i' \cdot \log(q_i) + (1-W_i')\cdot\log(1-q_i).\end{aligned}$$

*Proof.* From the fact that $\{(h_1,\ldots,h_\tau,q_{\tau+1},\ldots,q_n) | h\in\mathcal{H}\} \subset \{(h_1,\ldots,h_n)|h\in\mathcal{H}\}$ under $E$, we know that

$$\begin{aligned}&\sup_{h\in\mathcal{H}} \sum_{i=1}^{\tau} -W_i \cdot \log(h_i) - (1-W_i)\cdot\log(1-h_i) + \sum_{i=\tau+1}^{n} W_i \cdot \log(h_i) + (1-W_i)\cdot\log(1-h_i) \\ &\geq \sup_{h\in\mathcal{H}, q_i\in[\gamma,1-\gamma]} \sum_{i=1}^{\tau} -W_i \cdot \log(h_i) - (1-W_i)\cdot\log(1-h_i) + \sum_{i=\tau+1}^{n} W_i \cdot \log(q_i) + (1-W_i)\cdot\log(1-q_i) \\ &= \sup_{h\in\mathcal{H}} \sum_{i=1}^{\tau} -W_i \cdot \log(h_i) - (1-W_i)\cdot\log(1-h_i) + \sup_{q_i\in[\gamma,1-\gamma]} \sum_{i=\tau+1}^{n} W_i \cdot \log(q_i) + (1-W_i)\cdot\log(1-q_i),\end{aligned}$$

and

$$\sup_{h\in\mathcal{H}} \sum_{i=1}^{\tau} -W_i' \cdot \log(h_i) - (1-W_i')\cdot\log(1-h_i) + \sum_{i=\tau+1}^{n} W_i' \cdot \log(h_i) + (1-W_i')\cdot\log(1-h_i)$$

$$
\begin{aligned}
&\geq \sup_{h\in\mathcal{H}, q_i\in[\gamma,1-\gamma]} \sum_{i=1}^{\tau} -W_i'\cdot\log(h_i) - (1-W_i')\cdot\log(1-h_i) + \sum_{i=\tau+1}^{n} W_i'\cdot\log(q_i) + (1-W_i')\cdot\log(1-q_i)\\
&= \sup_{h\in\mathcal{H}} \sum_{i=1}^{\tau} -W_i'\cdot\log(h_i) - (1-W_i')\cdot\log(1-h_i) + \sup_{q_i\in[\gamma,1-\gamma]} \sum_{i=\tau+1}^{n} W_i'\cdot\log(q_i) + (1-W_i')\cdot\log(1-q_i).
\end{aligned}
$$

From Assumption B.6, we know that $\forall h \in \mathcal{H}$,

$$
\sum_{i=\tau+1}^{n} W_i\cdot\log(h_i) + (1-W_i)\cdot\log(1-h_i) \leq \sum_{i=\tau+1}^{n} -\mathrm{Ent}(W_i) = \sup_{q_i\in[\gamma,1-\gamma]} \sum_{i=\tau+1}^{n} W_i\cdot\log(q_i) + (1-W_i)\cdot\log(1-q_i),
$$

which implies that

$$
\begin{aligned}
&\sup_{h\in\mathcal{H}} \sum_{i=1}^{\tau} -W_i\cdot\log(h_i) - (1-W_i)\cdot\log(1-h_i) + \sum_{i=\tau+1}^{n} W_i\cdot\log(h_i) + (1-W_i)\cdot\log(1-h_i)\\
&\leq \sup_{h\in\mathcal{H}} \sum_{i=1}^{\tau} -W_i\cdot\log(h_i) - (1-W_i)\cdot\log(1-h_i) + \sup_{h\in\mathcal{H}} \sum_{i=\tau+1}^{n} W_i\cdot\log(h_i) + (1-W_i)\cdot\log(1-h_i)\\
&= \sup_{h\in\mathcal{H}} \sum_{i=1}^{\tau} -W_i\cdot\log(h_i) - (1-W_i)\cdot\log(1-h_i) + \sup_{q_i\in[\gamma,1-\gamma]} \sum_{i=\tau+1}^{n} W_i\cdot\log(q_i) + (1-W_i)\cdot\log(1-q_i).
\end{aligned}
$$

Combining the above, we conclude the proof. $\square$

**Lemma B.9.** *Denote the supremum over the cross-entropy as*

$$
f(z) \coloneqq \sup_{q\in[\gamma,1-\gamma]} z\cdot\log(q) + (1-z)\cdot\log(1-q).
$$

*Then*

$$
\mathrm{Rad}_{\mathcal{Z}',n}(\ell\circ\mathcal{H}) - \mathrm{Rad}_{\mathcal{Z},n}(\ell\circ\mathcal{H}) \geq \frac{1}{n}\mathbb{E}_{\tau,W}\left[\sum_{i=\tau+1}^{n} \mathbb{E}_{W'}\left[f(W_i')\middle|W_i\right]\right] - \frac{1}{n}\mathbb{E}_{\tau,W}\left[\sum_{i=\tau+1}^{n} f(W_i)\right].
$$

*Proof.* Due to the property of supremum, the supremum taken over the entire $(q_{\tau+1},\ldots,q_n) \in [\gamma, 1-\gamma]^{n-\tau}$ is equivalent to the supremum taken step by step (that is, by taking each $q_i$'s supremum one by one). Therefore,

$$
\sup_{q_i\in[\gamma,1-\gamma]} \sum_{i=\tau+1}^{n} W_i\cdot\log(q_i) + (1-W_i)\cdot\log(1-q_i) = \sum_{i=\tau+1}^{n} \sup_{q_i\in[\gamma,1-\gamma]} W_i\cdot\log(q_i) + (1-W_i)\cdot\log(1-q_i) = \sum_{i=\tau+1}^{n} f(W_i).
$$

Similar conclusions also hold for $W_i'$'s. We conclude the proof by combining the facts with Lemma B.8. $\square$

**Lemma B.10.** *Suppose* $w \in [\gamma, 1-\gamma]$ *and* $w' \in [0, 1]$. *Denote the clipped* $w'$ *by*

$$
w_\gamma' \coloneqq \max\{\gamma, \min\{w', 1-\gamma\}\}.
$$

*Then*

$$
f(w') \geq f(w) + \left.\frac{\mathrm{d}f}{\mathrm{d}z}\right|_{z=w} \cdot (w'-w) + 2(w_\gamma' - w)^2.
$$

*Proof.* For those $w' \in [\gamma, 1-\gamma]$, $w_\gamma' = w'$. The function $f$ of $w'$ is the entropy, of which the second-order derivative is

$$
\left.\frac{\mathrm{d}^2 f}{\mathrm{d}z^2}\right|_{z=w'} = \frac{1}{w'(1-w')} \geq 4.
$$

Then by Taylor's formula, we know that $\exists\theta \in [0, 1]$, s.t.

$$
f(w') = f(w) + \left.\frac{\mathrm{d}f}{\mathrm{d}z}\right|_{z=w} \cdot (w'-w) + \frac{1}{2}\cdot\left.\frac{\mathrm{d}^2 f}{\mathrm{d}z^2}\right|_{z=\theta w+(1-\theta)w'} \cdot (w'-w)^2
$$

$$\geq f(w) + \frac{\mathrm{d}f}{\mathrm{d}z}\bigg|_{z=w} \cdot (w' - w) + 2(w' - w)^2.$$

For those $w' > 1 - \gamma$, $w'_\gamma = 1 - \gamma$. The function $f$ is linear in the interval $[1 - \gamma, 1]$, of which the first-order derivative is $\frac{\mathrm{d}f}{\mathrm{d}z}|_{z=1-\gamma}$ (easy to check $f$ is continuously differentiable at $1 - \gamma$). Thus,

$$\begin{aligned} f(w') &= f(w'_\gamma) + \frac{\mathrm{d}f}{\mathrm{d}z}\bigg|_{z=w'_\gamma} \cdot (w' - w'_\gamma) \\ &\geq f(w'_\gamma) + \frac{\mathrm{d}f}{\mathrm{d}z}\bigg|_{z=w} \cdot (w' - w'_\gamma) \\ &\geq f(w) + \frac{\mathrm{d}f}{\mathrm{d}z}\bigg|_{z=w} \cdot (w' - w'_\gamma + w'_\gamma - w) + (w'_\gamma - w)^2, \end{aligned}$$

which verifies the proof.

For those $w' < \gamma$, the conclusion holds similarly. □

**Proposition B.11.** *Under Assumptions 3.1, B.6, and B.7, we have*

$$\mathrm{Rad}_{\mathcal{Z}',n}(\ell \circ \mathcal{H}) - \mathrm{Rad}_{\mathcal{Z},n}(\ell \circ \mathcal{H}) = \Omega\left(\mathbb{E}_W[\mathrm{Var}(W'|W)]\right).$$

*Proof.* Denote the conditional distribution of $W'_i$ on observing $W_i$ by $P_{W'_i|W_i}$. From Lemmas B.9 and B.10, we obtain that

$$\begin{aligned} &\mathrm{Rad}_{\mathcal{Z}',n}(\ell \circ \mathcal{H}) - \mathrm{Rad}_{\mathcal{Z},n}(\ell \circ \mathcal{H}) \\ \geq& \frac{1}{n}\mathbb{E}_{\tau,W}\left[\sum_{i=\tau+1}^{n} \mathbb{E}_{W'}\left[f(W'_i)\middle|W_i\right]\right] - \frac{1}{n}\mathbb{E}_{\tau,W}\left[\sum_{i=\tau+1}^{n} f(W_i)\right] \\ =& \frac{1}{n}\mathbb{E}_{\tau,W}\left[\sum_{i=\tau+1}^{n} \int f(W'_i) - f(W_i)\,\mathrm{d}P_{W'_i|W_i}\right] \\ \geq& \frac{1}{n}\mathbb{E}_{\tau,W}\left[\sum_{i=\tau+1}^{n} \int f(W_i) + \frac{\mathrm{d}f}{\mathrm{d}z}\bigg|_{z=W_i} \cdot (W'_i - W_i) + 2(W'_{i,\gamma} - W_i)^2 - f(W_i)\,\mathrm{d}P_{W'_i|W_i}\right] \\ =& \frac{1}{n}\mathbb{E}_{\tau,W}\left[\sum_{i=\tau+1}^{n} \int 2(W'_{i,\gamma} - W_i)^2\,\mathrm{d}P_{W'_i|W_i}\right]. \end{aligned}$$

By Assumption B.6, we know that

$$\frac{(W'_{i,\gamma} - W_i)^2}{(W'_i - W_i)^2} = \Omega(1),$$

which implies that

$$\int 2(W'_{i,\gamma} - W_i)^2\,\mathrm{d}P_{W'_i|W_i} \geq \Omega(1) \cdot \int (W'_i - W_i)^2\,\mathrm{d}P_{W'_i|W_i} = \Omega\big(\mathrm{Var}(W'_i|W_i)\big).$$

Since $\mathbb{E}[\tau] = \frac{n}{2}$, we know that

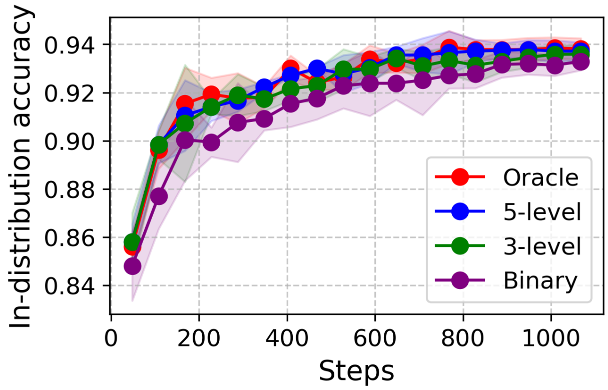
$$\mathrm{Rad}_{\mathcal{Z}',n}(\ell \circ \mathcal{H}) - \mathrm{Rad}_{\mathcal{Z},n}(\ell \circ \mathcal{H}) = \Omega\big(\mathrm{Var}(W'_i|W_i)\big).$$

□

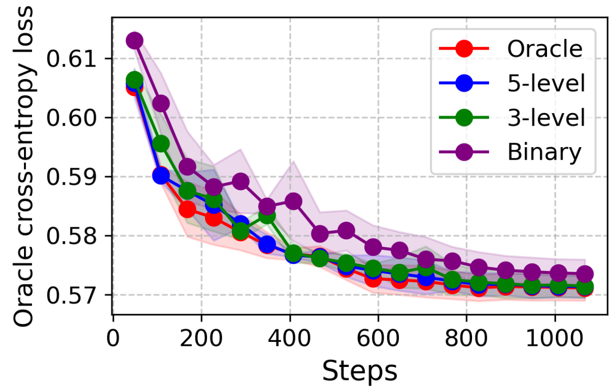
## C. Complementary Numerical Results

In this section, we present some more experimental results that are not listed in the main paper due to space limitations.

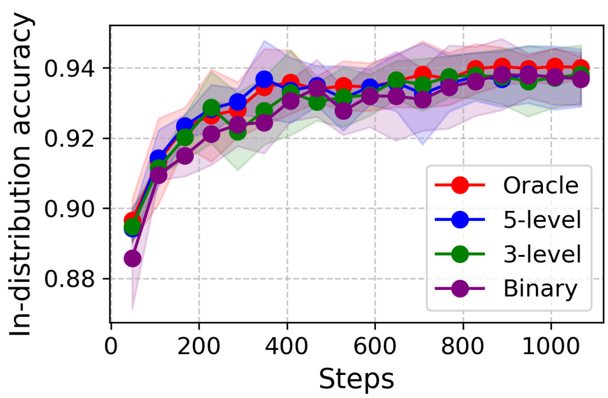
### C.1. Complementary figures



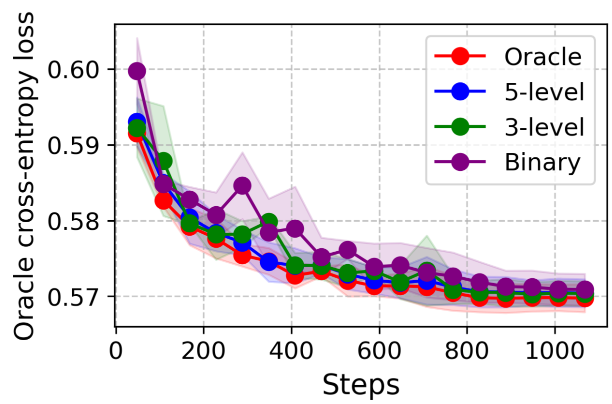

(a) In-distribution (ID) accuracy/llama models



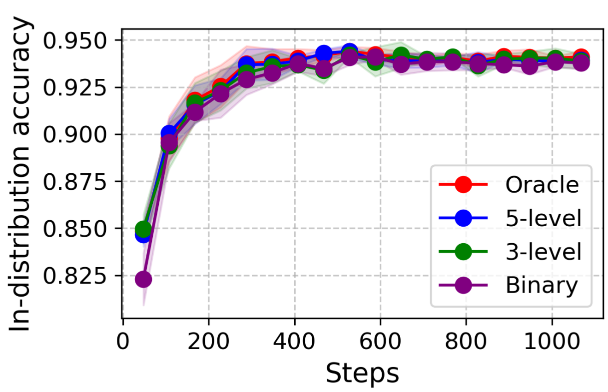

(b) Oracle CE loss/llama models



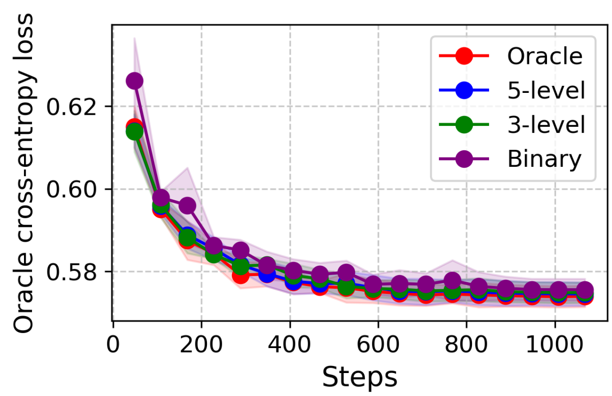

(c) In-distribution (ID) accuracy/gemma models




(d) Oracle CE loss/gemma models

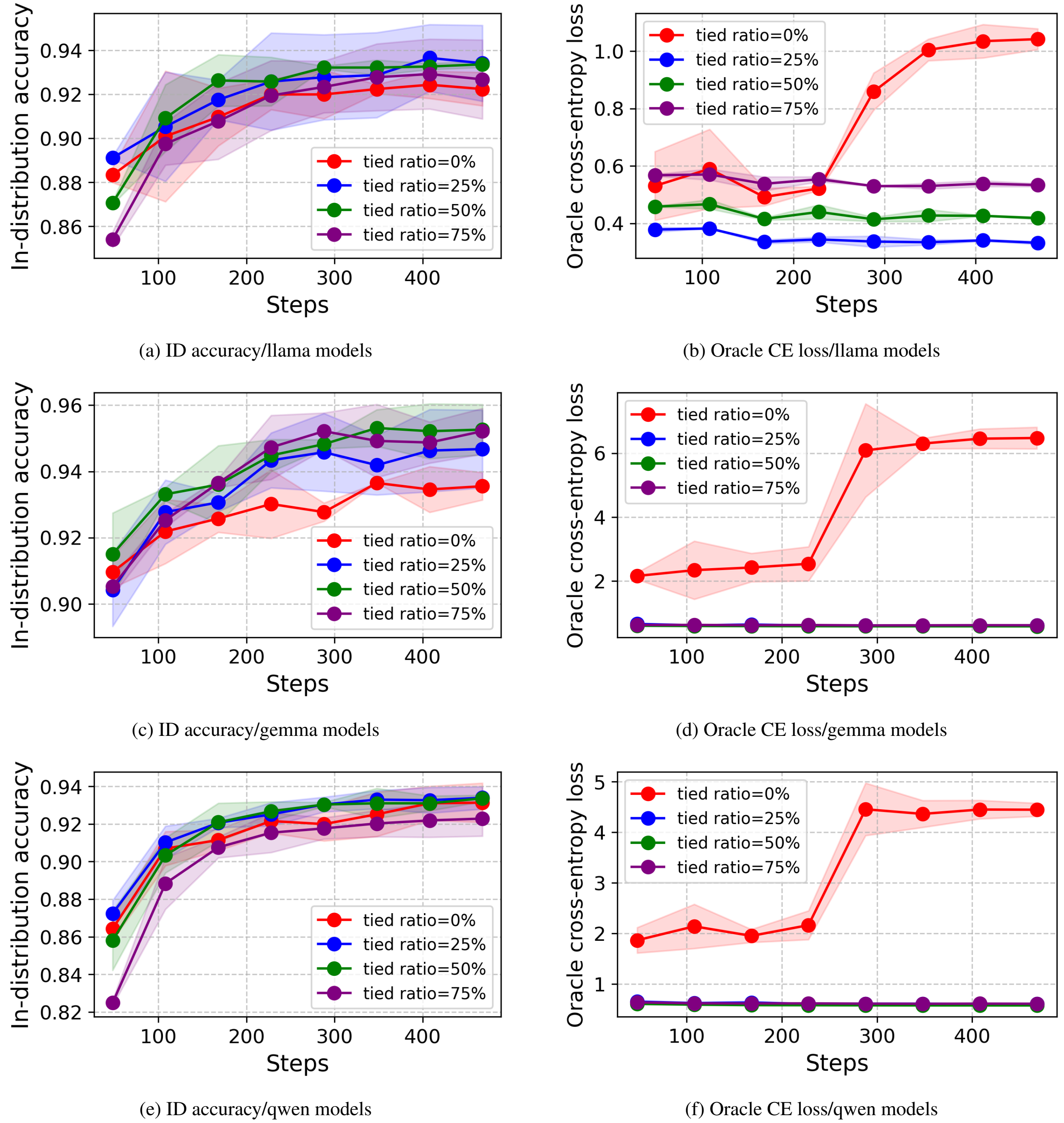



(e) In-distribution (ID) accuracy/qwen models




(f) Oracle CE loss/qwen models

*Figure 2.* The evaluation dynamics of the three base models for different ordinal feedback labels. The horizontal axis "Steps" represents the number of training samples; the rightward direction means a larger training set. The vertical axis represents the "accuracy" or the "loss" separately. For the accuracies, the larger, the better. For the losses, the smaller, the better. We can see from the figure that more fine-grained feedback leads to better RM learning.




(a) ID accuracy/llama models

(b) Oracle CE loss/llama models

(c) ID accuracy/gemma models

(d) Oracle CE loss/gemma models

(e) ID accuracy/qwen models

(f) Oracle CE loss/qwen models

*Figure 3.* The evaluation dynamics of the three base models for different tied data ratios. The horizontal axis "Steps" represents the number of training samples; the rightward direction means a larger training set. The vertical axis represents the "accuracy" or the "loss" separately. For the accuracies, the larger, the better. For the losses, the smaller, the better. The 100%-tied case is not plotted as it would detract from the clarity and readability of the plot due to its failure. We can see from the figure that mixing a certain ratio of tied data benefits the RM learning compared to 0% performance.

### C.2. Reward modeling with ordinal feedback under hinge loss

In the main paper, we focus on reward modeling with ordinal feedback under the cross-entropy loss. Here we extend the analysis to the case of hinge loss. Recall that the learning objective (2) in the main paper can be interpreted as an induction of the cross-entropy loss based on the Bradley-Terry model. Given the probabilistic model of ordinal feedback, the learning objective can naturally be extended to incorporate other types of loss functions. Among these, hinge loss (Schölkopf et al., 2004) is one of the most widely used, particularly in classification tasks, alongside cross-entropy loss. Hinge loss is commonly associated with Support Vector Machines (SVM) and is characterized by its core principle of enforcing a margin between distinct classes. Building on this observation, we define the learning objective under hinge loss as follows:

$$\min_{\theta} \sum_{i=1}^{n} Z_i \cdot \left[\max\left(0, C - (r_\theta(x_i, y_{i,1}) - r_\theta(x_i, y_{i,2}))\right)\right] + (1 - Z_i) \cdot \left[\max\left(0, C - (r_\theta(x_i, y_{i,2}) - r_\theta(x_i, y_{i,1}))\right)\right], \quad (8)$$

where $C$ is the margin hyperparameter that controls the separation between preference classes. When the feedback is binary, i.e., $Z_i \in \{0, 1\}$, the above objective simplifies to the hinge loss commonly used in reward modeling (Liu et al., 2024a). In our experiments, the margin parameter is tuned with grid search, where the search space is $\{0.5, 1, 2, 4\}$, and we select $C = 1$.

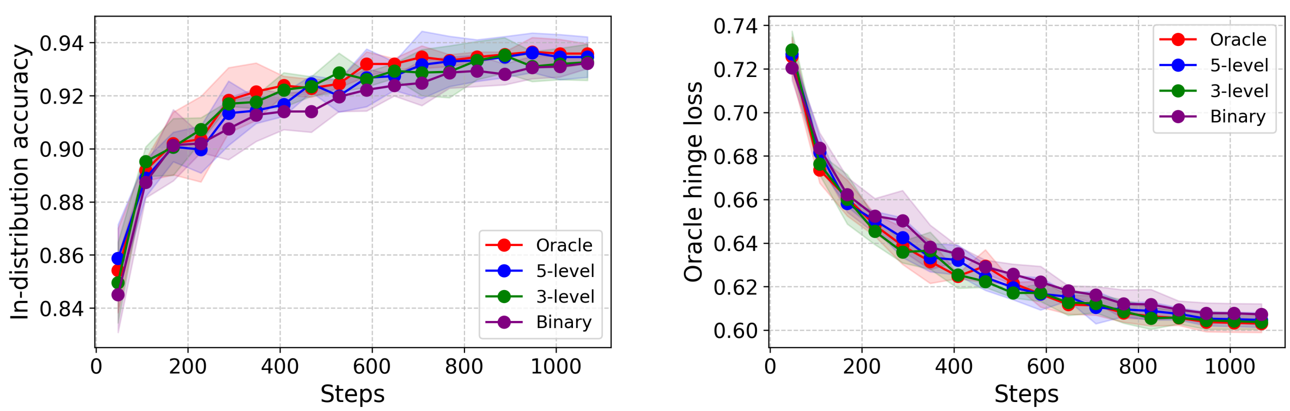
To further validate the conclusions presented in Section 5, we conduct analogous experiments under the learning objective (8). The feedback in the oracle feedback setting is still kept as

$$Z_i = \mathbb{P}\left(y_{i,1} \succ y_{i,2} | x_i\right) = z_{\text{oracle}}(x_i, y_{i,1}, y_{i,2})$$

just as the cross-entropy setting. The feedback in the 5-level/3-level/binary settings is sampled as the process in Theorem 3.2 by considering only the smallest interval $[z_j, z_{j+1}] \ni z_{\text{oracle}}$. We then replicate the experimental setup from Section 5.2, employing llama-3.2-1b-instruct as our base model. The experiment results are reported in Figure 4 and Table 3.

Experiments illustrate the same three conclusions in Section 5.2: (1) more fine-grained feedback structures result in better learning for both ID and OOD performance; (2) the fine-grained feedback (e.g. 5-level) may be a good proxy for the oracle; (3) the generalized hinge loss handle the feedback richer than binary well.

Apart from the observations analogous to Section 5.2, we observe that the hinge objective performs weaker than the cross-entropy objective. We attribute this outcome to two key factors. First, the margin hyperparameter $C$ significantly impacts the model's convergence speed and overall performance, which we only tune with a coarse grid due to computational constraints. Second, the inherent nature of the hinge objective means that only a specific subset of data points influences the final decision boundary. In contrast, the cross-entropy objective leverages the entire dataset during optimization. Given the complexity of language modeling and the intricacies of semantic space, we hypothesize that preference data embeddings are distributed in a noisy and overlapping manner such that the decision boundary may not be effectively established under the hinge objective.




(a) ID accuracy

(b) Oracle GH loss

*Figure 4.* The evaluation dynamics of llama models for different ordinal feedback labels under generalized hinge loss. The horizontal axis "Steps" represents the number of training samples; the rightward direction means a larger training set. The vertical axis represents the "accuracy" or the "loss" separately. For the accuracies, the larger, the better. For the losses, the smaller, the better.

*Table 3.* Model convergence statistics for llama under generalized hinge loss.

| Model | Feedback | Oracle CE Loss | | ID Accuracy | | OOD Accuracy | |
|---|---|---|---|---|---|---|---|
| | | Mean | Std | Mean | Std | Mean | Std |
| Llama | Oracle | 0.6030 | 0.004 | 0.9359 | 0.0036 | 0.7798 | 0.0076 |
| | 5-level | 0.6046 | 0.0003 | 0.9345 | 0.0076 | 0.7660 | 0.0068 |
| | 3-level | 0.6037 | 0.0003 | 0.9326 | 0.0068 | 0.7617 | 0.0163 |
| | Binary | 0.6072 | 0.0005 | 0.9322 | 0.0020 | 0.7580 | 0.0040 |

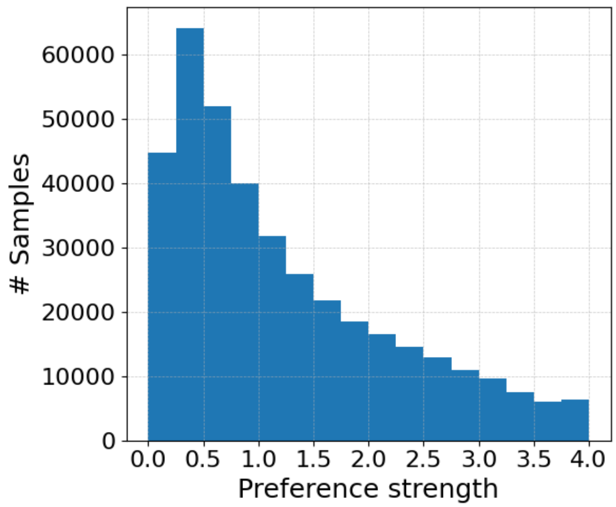
**Algorithm 1** 3-level Sampling Algorithm

**Input:** Oracle label $z_{\text{oracle}} \in [0, 1]$
**Output:** Sampled label $z \in \mathcal{Z}_3 = \{0, 0.5, 1\}$
**if** $z_{\text{oracle}} < 0.5$ **then**
  Sample $y \sim$ Bernoulli $\left(\frac{z_{\text{oracle}}}{0.5}\right)$
  $z \leftarrow 0.5 \cdot y$
**else if** $z_{\text{oracle}} > 0.5$ **then**
  Sample $y \sim$ Bernoulli $\left(\frac{z_{\text{oracle}}-0.5}{0.5}\right)$
  $z \leftarrow 0.5 \cdot y + 0.5$
**else**
  $z \leftarrow 0.5$
**end if**
**return** $z$

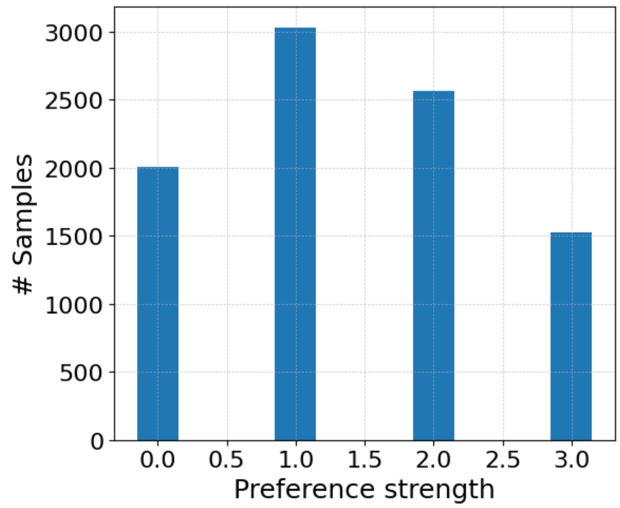
### C.3. Dataset details

**Skywork-Reward-Preference-80K-v0.2.** The Skywork-Reward-Preference-80K-v0.2 (SRP) dataset is a curated subset of publicly available preference data, spanning a wide range of knowledge domains. Reward models trained on this dataset have achieved top performance in the Reward Bench benchmark. The released version contains 77,016 samples, with approximately 5,000 overlapping samples removed compared to v0.1. In our experiments, we used Skywork-Reward-Gemma-2-27B-v0.2 to annotate data pairs with oracle scores.

**Rationale For Scaling.** Instead of directly using the sigmoid values of the oracle model's score differences, we introduce a scaling parameter $T$ because the raw output scores of the oracle model are highly concentrated, as shown in Figure 6a. However, it is generally understood that people often hold diverse opinions on preference samples, meaning real-world preference distributions should not be heavily concentrated near a probability of 1. To investigate, we consider two commonly used preference datasets with fine-grained preference scores, UltraFeedback (Cui et al., 2023) and HelpSteer2 (Wang et al., 2024c). A visualization of their preference ratings is provided in Figure 5, where many samples show no strong preference but only weak agreement. Based on these observations, we carefully adjusted the scaling parameter $T$ and chose $T = \frac{20}{3}$, ensuring it produces a peak within the slight agreement interval (approximately 0.6-0.7), as shown in Figure 6b. This choice is further justified in the explanation of the tied sample experiments discussed later.

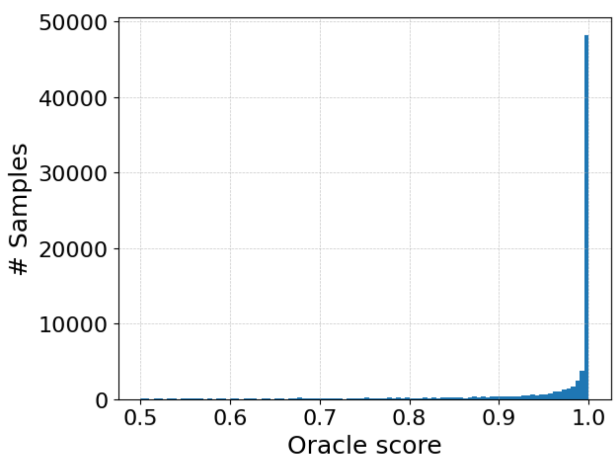



(a) Preference strength distribution in UltraFeedback

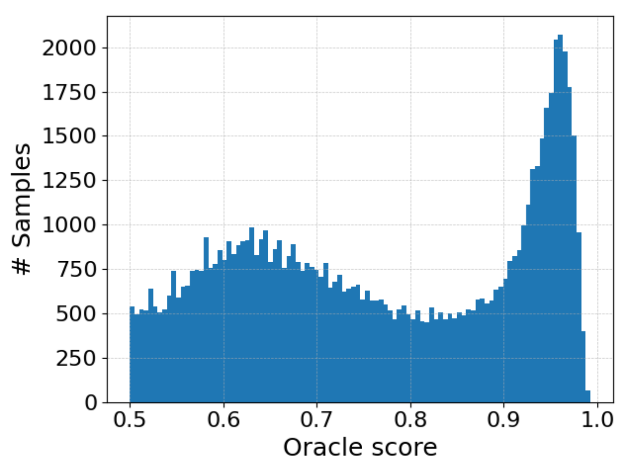



(b) Preference strength distribution in HelpSteer2

*Figure 5.* Distributions of preference strengths in the two datasets. For the UltraFeedback dataset, we compare the chosen and rejected scores pairwisely and use their differences as preference strengths. For the HelpSteer2 dataset, its latest version provides a preference strength label and we directly adopt it.




(a) Distribution of raw oracle scores




(b) Distribution of scaled oracle scores

*Figure 6.* Distribution of oracle labels before and after scaling with $T = \frac{20}{3}$. Note that the oracle score is always recorded for the chosen response relative to the rejected response, hence, all oracle scores are no less than 0.5.

**Tied Samples.** As presented in Section 5.1, we refer to preference pairs with a label $z = 0.5$ under the 3-level feedback system $\mathcal{Z}_3 = \{0, 0.5, 1\}$ as tied data or tied samples. We use the term "tied" because the labels $z$ are generated based on the oracle label $z_{\text{oracle}}$ using an interpolation paradigm, as described in Theorem 3.2. These labels represent preference samples where people perceive (almost) equal advantages. The detailed sampling process is summarized in Algorithm 1. And it can be easily extended to any ordinal feedback system.

Intuitively, the closer the oracle label of a preference pair is to 0.5, the more likely it is to be assigned a label of $0.5$. An important observation is that the sampled label distribution is dominated by $z_{\text{oracle}}$. To ensure sufficient binary and tied data for the tied ratio experiments, we aim for the binary and tied data to each constitute half of the dataset. After tuning, we found that selecting $T = \frac{20}{3}$ not only simulates a real-world preference data distribution but also satisfies the tied ratio experiment requirements. This further justifies the choice of $T$.

**RewardBench Evaluation.** The RewardBench evaluation dataset combines multiple datasets across four categories: Chat,

Chat Hard, Safety, and Reasoning. The scoring follows a standard reward modeling paradigm, where success is defined as the chosen response having a higher score than the rejected response for a given prompt. The evaluation score is computed as a weighted average across all prompts in the selected subset.

### C.4. Training details

*Table 4.* Hyperparameter Search Space

| Hyperparameter | Search Range/Values |
|---|---|
| Learning Rate | [1e-5, 5e-6, 2e-5] |
| Batch Size | [64, 128] |
| Warm-up Ratio | [0.03, 0.05, 0.10] |

*Table 5.* Shared Hyperparameters

| Hyperparameter | Value |
|---|---|
| Batch Size | 128 |
| Optimizer | paged_adamw_32bit |
| Weight Decay | 1e-3 |
| Epochs | 2 |
| Scheduler | Linear Warm-up + Cosine Decay |

*Table 6.* Model-specified Hyperparameters

| Hyperparameter | Llama-3.2-1b | Gemma-2-2b | Qwen-2.5-1.5b |
|---|---|---|---|
| Learning Rate | 1e-5 | 5e-6 | 5e-6 |
| Warm-up Ratio | 0.1 | 0.05 | 0.05 |

We choose some key training hyperparameters based on a grid search. The performance is assessed by in-distribution evaluation (CE) loss under oracle label settings. The grid search space is shown in Table 4.

The base models share most of the training parameters, as given in Table 5. The different parameters are listed in Table 6.