Mult-DPO: Multinomial Direct Preference Optimization for Recommender Systems

Yaochen Zhu[1] **Harald Steck**[2] **James McInerney**[2] **Aditya Sinha**[2]
Yinhan He[1] **Nathan Kallus**[2,3] **Jundong Li**[1]

[1]University of Virginia [2]Netflix [3]Cornell University
{uqp4qh,nee7ne,jundong}@virginia.edu
{hsteck,jmcinerney,adityasinha,nkallus}@netflix.com

Abstract

Direct preference optimization (DPO) is a simple and effective alignment strategy for large language models (LLMs) based on pairwise preferences. In recommender systems, however, user feedback is rarely pairwise. For a given context, e.g., a user, a session, or a conversation, we typically observe **set-wise preferences** with *multiple* positive items, where every positive item should outrank every unobserved or explicitly negative item, with no prescribed order among the positives or the negatives themselves. A natural generalization is to use the Plackett–Luce (PL) reward model, which extends the Bradley–Terry reward model underlying vanilla DPO from pairwise preferences to full rankings of candidates. However, we show that adapting the PL model to set-wise preferences requires marginalizing over all positive orderings, where the resulting expression is combinatorial in complexity. To address this fundamental challenge, we propose **Mult-DPO**, a novel DPO objective with a tractable multinomial surrogate likelihood over set-wise preference events for the user-preference alignment of LLM-based recommender systems. The multinomial construction is not itself a ranking distribution, but it is defined on the same reward-induced weight space and admits a closed-form DPO-style objective, enabling direct alignment of LLMs with multiple candidates through a classification-style objective. In addition, we prove that the multinomial DPO loss is a tractable *upper bound* on the marginalized PL DPO loss when optimizing against the set-wise preference data. We further characterize the tightness of this bound in terms of the relative total weight of positives versus negatives, which provides insights into tightening the bound with richer or harder negatives. Finally, we extend Mult-DPO to the alignment of LLMs with multiple preference levels. Code is available at https://github.com/yaochenzhu/Mult_DPO.

1 Introduction

Large language models (LLMs) have achieved remarkable success through large-scale pretraining. To further align these models with human preferences, reinforcement learning from human feedback (RLHF) [Christiano et al., 2017, Ouyang et al., 2022] emerged as a popular framework: It first trains a reward model from pairwise human preferences, and then optimizes the LLM policy via reinforcement learning (RL). Recently, direct preference optimization (DPO) [Rafailov et al., 2023] offers a simpler alternative by leveraging a closed-form connection between the optimal policy and the reward function in RLHF, eliminating the RL stage and reducing the alignment to a classification-style objective on preference data. Owing to its conceptual simplicity and strong empirical performance, DPO has been widely adopted in LLM alignment tasks such as summarization [Liu et al., 2023], code generation [Weyssow et al., 2026], and mathematical reasoning [Pang et al., 2024].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

As LLMs increasingly serve as the backbone of recommender systems (RSs) [Bao et al., 2023, Liao et al., 2024, He et al., 2023], it is natural to ask whether DPO can play the same role in aligning LLM-based RSs with user preferences. The answer, however, is complicated by a fundamental mismatch between the preference data structure assumed by vanilla DPO and that observed in recommendation tasks. Specifically, DPO assumes *pairwise* preferences between two candidates per context, which suits the setting of long responses for QA-style tasks where candidates are difficult to generate [Bai et al., 2022, Stiennon et al., 2020]. In contrast, user feedback in RSs is rarely pairwise. For a given context in RSs, e.g., an interaction history or a conversation between user and RS, we typically observe *multiple* positive items (e.g., items liked, clicked by the user) and negative items (e.g., items unobserved or explicitly rejected) [Rendle et al., 2009, Liang et al., 2018]. From the perspective of preference alignment, every positive should be preferred by the LLM over every negative, yet no order should be enforced among the positives or among the negatives themselves. Applying vanilla DPO to set-wise preferences therefore requires enumerating all positive-negative combinations, which is computationally expensive and discards the joint rank structure of the preference data.

Recently, several works have adapted DPO to preference alignment of LLMs with multiple candidates, which can be categorized into two technical routes. One route is to generalize the Bradley-Terry (BT) [Bradley and Terry, 1952] reward model underlying vanilla DPO from pairwise preferences to its listwise counterpart, the Plackett-Luce (PL) model [Plackett, 1975, Luce, 1959], including methods such as PRO [Song et al., 2024], KPO [Cai et al., 2025], etc. However, when applied to set-wise preference, the PL likelihood must be marginalized over all positive orderings consistent with the observation (where the orderings of the negatives can be directly summed-out), which is combinatorially expensive. Another route sidesteps this intractability by restricting the supervision to one single positive per context paired with multiple negatives. For example, DMPO [Bai et al., 2024] contrasts the positive against the arithmetic mean of the negative log-ratios within the BT model, whereas S-DPO [Chen et al., 2024] shows that, under the single-positive restriction, the marginalized PL likelihood collapses to a closed-form softmax DPO loss that contrasts the positive jointly against all sampled negatives. Nevertheless, neither route faithfully retains the joint set-wise structure of the feedback, leaving alignment with multi-positive preferences a fundamentally unresolved challenge.

To address the challenge, we propose **Mult-DPO**, a novel DPO-style objective for aligning LLM-based RSs with set-wise preferences. We first introduce a multinomial surrogate event model as a tractable lower-bound surrogate for the marginalized PL likelihood over set-wise preferences between the positive and negative candidate sets. Although this surrogate is not a ranking model, it is defined on the same reward-induced weights as BT and PL and yields a closed-form DPO-style objective through the standard RLHF policy-ratio reparameterization. Theoretically, we prove that minimizing the multinomial DPO loss optimizes a tractable upper bound on the otherwise intractable marginalized PL DPO loss. Furthermore, we characterize the tightness of this bound in closed form in terms of the relative total weights (i.e., exponentiated rewards) of positives versus negatives, and show that the bound becomes tighter when richer or harder negatives are selected. Finally, we extend Mult-DPO to alignment with multiple preference levels when fine-grained user preference data is available (e.g., explicit ratings) via a *sequential multinomial (SMN)* surrogate, and show that all theoretical properties of Mult-DPO from the two-set case carry over to this generalization. On both general recommendation and conversational recommendation benchmarks, we demonstrate that Mult-DPO and its multi-level extension consistently outperform various DPO-based baselines.

2 Preliminaries

2.1 Problem Formulation

In this paper, we aim to tackle the fundamental challenge of aligning LLM-based RSs with user preferences over multiple candidates. Specifically, let x denote the recommendation context, which may include a user profile, an interaction history, or a conversation, and let $\mathcal{C}$ denote the item catalog from which recommendations should be generated, where each item e is rendered as a token sequence $y(e) = (y_1, \ldots, y_{m_e})$. For each context x in the preference dataset $\mathcal{D}$, we observe user preference over candidate items in the form of $(x, \mathcal{E}^p, \mathcal{E}^d)$, where $\mathcal{E}^p = \{e_1, \ldots, e_k\}$ and $\mathcal{E}^d = \{e_{k+1}, \ldots, e_K\}$ denote disjoint positive and negative item sets associated with x, respectively. Here, $\mathcal{E}^p$ contains items receiving positive feedback (e.g., liked or clicked), whereas $\mathcal{E}^d$ contains uninteracted or rejected items. We use $\mathcal{E} = \mathcal{E}^p \cup \mathcal{E}^d \subset \mathcal{C}$ to denote the full candidate set, with $k = |\mathcal{E}^p|$ and $K - k = |\mathcal{E}^d|$ for the set-wise contrastive setting studied below. We note that the user preference tuple implies a

set-wise ranking constraint that every positive item in $\mathcal{E}^p$ should outrank every negative item in $\mathcal{E}^d$:

$$\Omega_x := \{e \succ e' \mid e \in \mathcal{E}^p,\ e' \in \mathcal{E}^d\}, \tag{1}$$

but no order is implied among the positives or the negatives themselves. Let $\pi_\theta(e \mid x)$ denote an LLM-based RS policy that assigns a generation probability to each candidate item e given the context x. This paper aims to align π_θ with the set-wise preference structure in Ω_x that fully leverages the joint multi-positive and negative structure while remaining computationally efficient, such that the items in catalog $\mathcal{C}$ can be effectively ranked for recommendation at inference time.

2.2 Theoretical Analysis: RLHF, DPO and Marginalized Plackett-Luce Extension

In this section, we first review reinforcement learning from human feedback and direct preference optimization, and discuss the naive extension of DPO to set-wise preferences with a marginalized Plackett-Luce likelihood, which forms the foundation of Mult-DPO introduced in Section 3.

2.2.1 Reinforcement Learning from Human Feedback (RLHF). RLHF aligns an LLM-based policy $\pi_\theta(e \mid x)$ with user preferences by optimizing it against a reward model derived from human preferences. Specifically, for each context-item pair (x, e), we assume a latent reward $r(x, e) \in \mathbb{R}$ indicating the utility of recommending item e under context x. We define the associated weight as $w(e \mid x) := \exp(r(x, e)) > 0$ for the convenience of linking $r(x, e)$ to probabilistic distributions. The **Bradley-Terry Preference Model** (BT) [Bradley and Terry, 1952] assumes that, given two candidate items e_p and e_d for context x with e_p preferred over e_d, the probability of observing such pair-wise preference can be formulated with the difference between their latent rewards as follows:

$$P(e_p \succ e_d \mid x) = \sigma\big(r(x, e_p) - r(x, e_d)\big) = \frac{w(e_p \mid x)}{w(e_p \mid x) + w(e_d \mid x)}, \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function. The reward model is then estimated by minimizing the negative log-likelihood of the observed comparisons induced by Eq. (2). Given a fixed reference policy $\pi_{\text{ref}}(e \mid x)$ and learned reward model $r(x, e)$, the RLHF objective can be formulated as follows:

$$\max_{\pi(\cdot|x)} \ \mathbb{E}_{e\sim\pi(\cdot|x)}[r(x, e)] - \beta\, \mathrm{KL}(\pi(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x))\,, \tag{3}$$

where $\beta > 0$ is an important hyperparameter that controls the regularization strength.

2.2.2 Direct Preference Optimization (DPO). DPO [Rafailov et al., 2023] avoids explicit reward modeling by observing that the RLHF objective in Eq. (3) admits a closed-form solution as follows:

$$r(x, e) = \beta \log \frac{\pi^*(e \mid x)}{\pi_{\text{ref}}(e \mid x)} + \beta \log Z(x), \tag{4}$$

where $Z(x)$ is an intractable partition function. Observing that the BT likelihood in Eq. (2) depends only on reward differences, $Z(x)$ cancels out after substituting Eq. (4). Replacing π^* with a trainable policy π_θ yields the DPO objective as follows:

$$\mathcal{L}_{\text{DPO}}(x, e_p, e_d) = -\log \sigma\left(\beta \log \frac{\pi_\theta(e_p \mid x)}{\pi_{\text{ref}}(e_p \mid x)} - \beta \log \frac{\pi_\theta(e_d \mid x)}{\pi_{\text{ref}}(e_d \mid x)}\right), \tag{5}$$

which reduces RLHF alignment to a classification-like objective on the preference data.

2.2.3 Marginalized Plackett-Luce DPO Objective for Set-wise Preference. Vanilla DPO can only align LLMs with two candidates due to the pairwise constraint of the BT preference model. A natural generalization of BT to multiple candidates is the Plackett-Luce (PL) model [Plackett, 1975, Luce, 1959]. As in BT, PL associates each item $e \in \mathcal{E}$ with a latent reward $r(x, e)$ (and weight $w(e \mid x)$). For a permutation τ of the candidate items in $\mathcal{E}$, the PL likelihood can be formulated as follows:

$$p_{\text{PL}}(\tau \mid x, \mathcal{E}; w) = \prod_{t=1}^{|\mathcal{E}|} \frac{w(e_{\tau(t)} \mid x)}{\sum_{j=t}^{|\mathcal{E}|} w(e_{\tau(j)} \mid x)}, \tag{6}$$

which denotes a sequential selection process where the item at rank t is sampled with probability proportional to its weight among the remaining candidates. To use PL to model the set-wise preference event Ω_x, however, since the positive and negative ordering are unknown, we must marginalize over

all consistent ranks of $\mathcal{E}$. Let S_k denote the permutations of the positive indices $\{1, \ldots, k\}$, and define the cumulative weights of the positive set, negative set, and full candidate set under context x:

$$A := \sum_{e \in \mathcal{E}^p} w(e \mid x), \;\; B := \sum_{e \in \mathcal{E}^d} w(e \mid x), \;\; W := A + B = \sum_{e \in \mathcal{E}} w(e \mid x), \tag{7}$$

With these definitions, we show that by marginalizing the PL likelihood in Eq. (6) (see Appendix A), the likelihood of Ω_x can be formulated as follows (which we dub the **marginalized PL event model**):

$$p_{\mathrm{PL}}(\Omega_x \mid x, \mathcal{E}; w) = \sum_{\rho \in S_k} \prod_{t=1}^{k} \frac{w(e_{\rho(t)} \mid x)}{B + \sum_{j=t}^{k} w(e_{\rho(j)} \mid x)}, \tag{8}$$

where the marginalization over the negative permutations gracefully vanishes into the cumulative weights B. An equivalent inclusion-exclusion form derived from the exponential-race interpretation of PL reduces the number of terms from $k!$ to 2^k (see A.3), but it still remains exponential w.r.t. k. Therefore, directly optimizing Eq. (8) as a DPO-style objective is infeasible even if k is moderate.

3 Mult-DPO: Multinomial Direct Preference Optimization

3.1 Multinomial Surrogate Event Model

To obtain an effective and tractable surrogate for the exact marginalized PL event model for aligning LLMs with set-wise preferences, we propose an alternative event model defined on the same weight space, i.e., $w(e \mid x), \; \forall e \in \mathcal{E}$, which we name the **multinomial (MN) surrogate**. The surrogate first normalizes the weights into a categorical distribution over candidates:

$$p(e \mid x) := \frac{w(e \mid x)}{W}, \;\; e \in \mathcal{E}. \tag{9}$$

The surrogate likelihood then defines the observed set-wise event Ω_x as the probability, under k independent item draws from $p(\cdot \mid x)$, that *each positive item in $\mathcal{E}^p$ appears exactly once while no negative item is sampled*. Since the k positives can appear in any draw order, there are $k!$ equivalent sequences that correspond to this event. Therefore, the MN surrogate likelihood of the set-wise preference event can be formulated as follows:

$$p_{\mathrm{MN}}(\Omega_x \mid x, \mathcal{E}; w) = k! \prod_{e \in \mathcal{E}^p} \frac{w(e \mid x)}{W}, \tag{10}$$

which can be computed in $\mathcal{O}(k)$ complexity. Unlike PL, the MN construction is not a distribution over permutations: it is an IID event likelihood and assigns probability mass to duplicate draw sequences outside the valid ranking space. We use it as a tractable surrogate for the observed positive set, and the lower-bound result justifies it as a conservative surrogate for the exact marginalized PL likelihood:

Theorem 1. *For any disjoint sets $\mathcal{E}^p$ and $\mathcal{E}^d$ with $|\mathcal{E}^p| \geq 1$, and any positive weights $\{w(e \mid x)\}_{e \in \mathcal{E}}$,*

$$p_{\mathrm{PL}}(\Omega_x \mid x, \mathcal{E}; w) \geq p_{\mathrm{MN}}(\Omega_x \mid x, \mathcal{E}; w). \tag{11}$$

In addition, we prove that the ratio between the PL likelihood and the MN surrogate likelihood is bounded, which will later imply a loss-gap bound between marginalized PL-DPO and Mult-DPO.

Theorem 2. *For any disjoint sets $\mathcal{E}^p$ and $\mathcal{E}^d$ with $|\mathcal{E}^p| \geq 1$ and $|\mathcal{E}^d| \geq 1$, and any positive weights $\{w(e \mid x)\}_{e \in \mathcal{E}}$, the ratio between the marginalized PL likelihood and the multinomial surrogate likelihood is bounded by the following relations*

$$1 \;\leq\; \frac{p_{\mathrm{PL}}(\Omega_x \mid x, \mathcal{E}; w)}{p_{\mathrm{MN}}(\Omega_x \mid x, \mathcal{E}; w)} \;\leq\; \left(1 + \frac{A}{B}\right)^{k-1}. \tag{12}$$

The proof of Theorems 1 and 2 is in A.5 and A.6. We note that the bound in Eq. (12) is exact when $k = 1$, in which case the MN surrogate likelihood reduces to the categorical likelihood in S-DPO [Chen et al., 2024], which generalizes DPO to alignment with **one** positive over multiple negatives.

3.2 Mult-DPO Objective for LLM Alignment with Set-wise Preference

Building on the MN surrogate, we further show that it is compatible with the same RLHF reward-policy reparameterization that underlies DPO, yielding a classification-like alignment objective on (now set-wise) preference data. Recall that, under RLHF, the reward can be expressed in terms of the optimal policy and the reference policy as $r(x, e) = \beta \log \frac{\pi^*(e|x)}{\pi_{\text{ref}}(e|x)} + \beta \log Z(x)$ (see Eq. (4)), and the item weight is defined as $w(e \mid x) := \exp(r(x, e))$. Combining these two relations and replacing the optimal policy π^* with a trainable policy π_θ yields the **policy-induced weights** as follows:

$$w_{\boldsymbol{\pi_\theta}}(e \mid x) \propto \left(\frac{\pi_\theta(e \mid x)}{\pi_{\text{ref}}(e \mid x)} \right)^\beta, \tag{13}$$

where the omitted proportionality constant $Z_{\pi_\theta}(x)^\beta$ is shared across candidates and cancels in both the PL and MN surrogate likelihoods. Substituting the policy-induced weights $w_{\pi_\theta}(e \mid x)$ into Eq. (10) and taking the negative log-likelihood leads to the proposed **Mult-DPO** objective as follows:

$$\begin{aligned} \mathcal{L}_{\text{Mult-DPO}}(x, \mathcal{E}^p, \mathcal{E}^d) &:= -\log p_{\text{MN}}(\Omega_x \mid x, \mathcal{E}; w_{\pi_\theta}) \\ &= -\beta \sum_{e \in \mathcal{E}^p} \log \frac{\pi_\theta(e \mid x)}{\pi_{\text{ref}}(e \mid x)} + k \log \sum_{e \in \mathcal{E}} \left(\frac{\pi_\theta(e \mid x)}{\pi_{\text{ref}}(e \mid x)} \right)^\beta + C. \end{aligned} \tag{14}$$

where C is a constant depending only on k and can be omitted during optimization.

3.3 Relation between Marginalized PL and Multinomial DPO Objective

To further characterize the relation between the PL-DPO and the tractable Mult-DPO objectives, we define the policy-induced cumulative weights of the positives and negatives analogous to Eq. (7) as:

$$A_{\pi_\theta} := \sum_{e \in \mathcal{E}^p} w_{\pi_\theta}(e \mid x), \;\; B_{\pi_\theta} := \sum_{e \in \mathcal{E}^d} w_{\pi_\theta}(e \mid x), \tag{15}$$

where both A and B now depend on policy π_θ. Substituting the policy-induced weights from Eq. (13) into Eq. (8) and taking the negative log-likelihood gives the ideal but intractable PL-DPO loss:

$$\mathcal{L}_{\text{PL-DPO}}(x, \mathcal{E}^p, \mathcal{E}^d) := -\log p_{\text{PL}}(\Omega_x \mid x, \mathcal{E}; w_{\pi_\theta}), \tag{16}$$

Theorems 1 and 2 immediately imply the following corollary between $\mathcal{L}_{\text{PL-DPO}}$ and $\mathcal{L}_{\text{Mult-DPO}}$.

Corollary 1. *For every context x and every training instance $(x, \mathcal{E}^p, \mathcal{E}^d)$,*

$$\mathcal{L}_{\text{PL-DPO}}(x, \mathcal{E}^p, \mathcal{E}^d) \leq \mathcal{L}_{\text{Mult-DPO}}(x, \mathcal{E}^p, \mathcal{E}^d), \tag{17}$$

and

$$0 \leq \mathcal{L}_{\text{Mult-DPO}}(x, \mathcal{E}^p, \mathcal{E}^d) - \mathcal{L}_{\text{PL-DPO}}(x, \mathcal{E}^p, \mathcal{E}^d) \leq (k-1) \log\left(1 + \frac{A_{\pi_\theta}}{B_{\pi_\theta}}\right). \tag{18}$$

The proof follows from Theorem 2 by substituting the policy-induced weights and taking negative logarithms. The corollary shows that the Mult-DPO loss is a tractable upper bound on the exact but otherwise intractable marginalized PL-DPO loss. In addition, it reveals that, for a fixed k-positive set and current policy, increasing the non-negligible policy-induced weight π_θ for the negatives B_{π_θ} sharpens the worst-case gap, suggesting that the MN surrogate becomes a better approximation of the intractable marginalized PL-DPO loss when richer or harder negatives are selected as candidates.

3.4 Extension of Mult-DPO to Multi-Level Preference

The preceding sections assume binary preferences from users, which cover the most common RS setting, including recommendation with implicit feedback [Rendle et al., 2009]. In practice, however, user feedback can be more fine-grained. For example, users may provide different ratings for the items. Such feedback naturally induces a **multi-level preference structure**, where items are partitioned into multiple preference groups. To capture this richer preference structure, we extend Mult-DPO from binary set-wise preferences to multi-level set-wise preferences, denoted **Mult^2-DPO**.

3.4.1 Problem Formulation of Multi-Level Preference. Suppose that, for a recommendation-relevant context x, the candidate set $\mathcal{E}$ is partitioned into $G \geq 2$ disjoint preference groups as:

$$\mathcal{E} = \bigcup_{g=1}^{G} \mathcal{E}^{(g)}, \ |\mathcal{E}^{(g)}| = k_g, \ \sum_{g=1}^{G} k_g = K, \tag{19}$$

where the group index denotes the preference level, i.e., items in $\mathcal{E}^{(g)}$ should outrank items in $\mathcal{E}^{(h)}$ if $g < h$, with no order among items within each group. We denote the multi-level preference event as:

$$\Omega_x^{\text{grp}} := \Big\{ e \succ e' \;\Big|\; e \in \mathcal{E}^{(g)},\ e' \in \mathcal{E}^{(h)},\ 1 \leq g < h \leq G \Big\}. \tag{20}$$

For each boundary $g = 1, \ldots, G-1$, we further define the **group-g-specific preference event** $\Omega_x^{(g)} := \{e \succ e' \mid e \in \mathcal{E}^{(g)},\ e' \in \bigcup_{h=g+1}^{G} \mathcal{E}^{(h)}\}$. This event has the same form as the two-set preference event Ω_x in Eq. (1) if we view $\mathcal{E}^{(g)}$ as $\mathcal{E}^p$ and $\bigcup_{h=g+1}^{G} \mathcal{E}^{(h)}$ as $\mathcal{E}^d$. By definition, we have $\Omega_x^{\text{grp}} = \bigcap_{g=1}^{G-1} \Omega_x^{(g)}$. For each boundary $g = 1, \ldots, G-1$ and any positive weights, we define

$$A_g := \sum_{e \in \mathcal{E}^{(g)}} w(e \mid x),\ B_g := \sum_{h=g+1}^{G} \sum_{e \in \mathcal{E}^{(h)}} w(e \mid x),\ W_g := A_g + B_g = \sum_{h=g}^{G} \sum_{e \in \mathcal{E}^{(h)}} w(e \mid x), \tag{21}$$

where A_g is the total weight of group g and B_g is the total weight of all lower-preference groups.

3.4.2 Marginalized PL Reward Model for Multi-Level Preference. By the sequential selection property of PL, once items from the higher-preference groups $\{\mathcal{E}^{(1)}, \ldots, \mathcal{E}^{(g-1)}\}$ have been placed, the events $\Omega_x^{(g)}$ become conditionally independent across g, which yields the following recursive factorization of the marginalized PL likelihood for the multi-level preference data:

$$p_{\text{PL}}(\Omega_x^{\text{grp}} \mid x, \mathcal{E}; w) = \prod_{g=1}^{G-1} p_{\text{PL}} \left(\Omega_x^{(g)} \;\middle|\; x,\ \bigcup_{h=g}^{G} \mathcal{E}^{(h)};\ w \right), \tag{22}$$

where each factor is a set-wise marginalized PL likelihood of the same form studied in Section 3.1.

3.4.3 Sequential Multinomial (SMN) Surrogate Model and SMN-DPO. This factorization in Eq. (22) naturally extends the multinomial lower bound group by group. Applying the two-set MN surrogate to each group-specific event yields the sequential multinomial (SMN) surrogate

$$p_{\text{SMN}}(\Omega_x^{\text{grp}} \mid x, \mathcal{E}; w) := \prod_{g=1}^{G-1} \left(k_g! \prod_{e \in \mathcal{E}^{(g)}} \frac{w(e \mid x)}{W_g} \right), \tag{23}$$

By applying Theorem 1 at each factor, Eq. (23) provides a tractable lower bound on the exact PL likelihood in Eq. (22). Substituting the policy-induced weight $w_{\pi_\theta}(e \mid x)$ from Eq. (13) into Eq. (23) and taking the negative log-likelihood leads to the multi-group Mult-DPO (**Mult^2-DPO**) objective

$$\begin{aligned} \mathcal{L}_{\text{Mult}^2\text{-DPO}} &:= -\log p_{\text{SMN}}(\Omega_x^{\text{grp}} \mid x, \mathcal{E}; w_{\pi_\theta}) \\ &= \sum_{g=1}^{G-1} \left[-\beta \sum_{e \in \mathcal{E}^{(g)}} \log \frac{\pi_\theta(e \mid x)}{\pi_{\text{ref}}(e \mid x)} + k_g \log \sum_{h=g}^{G} \sum_{e \in \mathcal{E}^{(h)}} \left(\frac{\pi_\theta(e \mid x)}{\pi_{\text{ref}}(e \mid x)} \right)^{\beta} \right] + C', \end{aligned} \tag{24}$$

where $C' = -\sum_{g=1}^{G-1} \log k_g!$ is a constant independent of θ and can be omitted during optimization. When $G = 2$, Eq. (24) reduces to the binary Mult-DPO objective in Eq. (14). The same boundary-wise reduction also gives a multi-level analogue of Corollary 1. If we define the PL-DPO loss

$$\mathcal{L}_{\text{PL-DPO}}^{\text{grp}} := -\log p_{\text{PL}}(\Omega_x^{\text{grp}} \mid x, \mathcal{E}; w_{\pi_\theta}), \tag{25}$$

and let $A_{\pi_\theta, g}$ and $B_{\pi_\theta, g}$ denote the group masses in Eq. (21) after substituting w_{π_θ} for w, we have:

Corollary 2 (Multi-level loss-gap bound)**.** *For every context x and every multi-level preference instance with nonempty ordered groups, we have*

$$0 \leq \mathcal{L}_{\text{Mult}^2\text{-DPO}} - \mathcal{L}_{\text{PL-DPO}}^{\text{grp}} \leq \sum_{g=1}^{G-1} (k_g - 1) \log \left(1 + \frac{A_{\pi_\theta, g}}{B_{\pi_\theta, g}} \right). \tag{26}$$

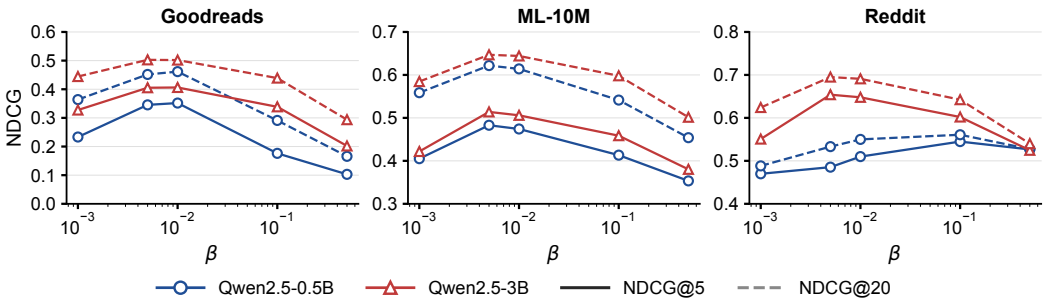


Figure 1: Validation NDCG@5 (solid) and NDCG@20 (dashed) versus the regularization strength β.

3.5 Complexity Analysis

Although Mult-DPO aligns LLM-based RSs with user preferences over more candidates per training step than its single-positive counterparts (e.g., vanilla DPO and S-DPO), the per-step time complexity remains comparable thanks to the shared prompt prefix that allows **KV-cache reuse** of the prompt across all candidates. Concretely, let N_x denote the number of prompt tokens and N_i the average number of tokens per item. With KV-cache reuse, the per-step self-attention cost of every DPO-style method takes the form $\mathcal{O}(N_x^2 + cN_xN_i + cN_i^2)$. Here, c is the number of candidates scored per step: $c = 2$ for vanilla DPO, $c = 1 + (K - k)$ for S-DPO, and $c = K$ for Mult-DPO, where k and K are the numbers of positives and full candidates, respectively. Therefore, when $K \ll N_x$, which is typical in the RS settings as the prompt usually contains user context, user/item features, interaction history, or conversation, the prompt-level self-attention term N_x^2 dominates; in our experiments, we indeed find that Mult-DPO exhibits similar wall-clock times to vanilla DPO and S-DPO.

4 Empirical Study

In this section, we evaluate Mult-DPO and its extension to multi-level preference alignment, i.e., Mult^2-DPO, on both general and conversational recommendation tasks, aiming to answer the following three research questions: **(RQ1)** how does Mult-DPO perform compared to other supervised finetuning (SFT)-based baselines or DPO-style baselines for aligning LLM-based RSs under set-wise feedback; **(RQ2)** how well does the MN DPO loss of Mult-DPO match the theoretical upper bound on the marginalized PL DPO loss from Corollary 1; and **(RQ3)** does Mult^2-DPO yield further performance gains over Mult-DPO when explicit multi-level preferences are available?

4.1 Datasets and Experimental Setup

For general recommendation tasks, we use the `MovieLens-10M` [Harper and Konstan, 2015] and `Goodreads` [Wan and McAuley, 2018] datasets. For `MovieLens-10M`, where explicit user-item ratings are available, we first treat items rated 5 as positives and randomly sampled unrated items as negatives, and then treat different ratings as different preference levels in Section 4.5 to evaluate Mult^2-DPO. For conversational recommendation, we adapt the `Reddit-V2` [He et al., 2023, Zhu et al., 2026] dataset, where we combine recommendations from different Reddit users on the same conversation as multiple ground truths. For validation, we use sampled negatives with 200 candidates in total for efficient model selection, and the test results are reported against the full catalog. We report NDCG@K for $K \in \{5, 15, 20\}$. Full implementation details are provided in Appendix C.1.

4.2 Training Dynamics and Hyperparameter Analysis

We first study the effects of the regularization strength β in Eq. (3) on the validation sets, summarized in Figure 1. Due to space limitations and constraint of computational budgets, we present the results with the `Qwen-2.5-0.5B-Instruct` [Qwen et al., 2024] and `Qwen-2.5-3B-Instruct` backbones, while we use the insight to select β to run Mult-DPO on the Goodreads dataset with $\{0.5B, 1.5B, 3B, 7B\}$ backbones in C.3 to show its generalization ability w.r.t. model scale. We observe that too small a β lets π_θ drift too far from a strong π_{ref}, while too large a β over-regularizes

Table 1: Comparison between Mult-DPO and DPO-style baselines on the test sets of Goodreads, MovieLens-10M, and Reddit-V2. Evaluation is conducted with the entire catalog as the candidates. Most results are significant as standard error is between 0.0020-0.0035 for N@5, N@15 and N@20.

Method	Goodreads			MovieLens-10M			Reddit-V2		
	N@5	N@15	N@20	N@5	N@15	N@20	N@5	N@15	N@20
Qwen2.5-0.5B-Instruct (zero-shot)	0.0085	0.0090	0.0090	0.0136	0.0156	0.0156	0.0175	0.0207	0.0216
SFT-based methods									
+ BigRec [Bao et al., 2025]	0.0776	0.1069	0.1177	**0.0657**	0.0884	0.0962	0.1043	0.1056	0.1083
+ D^3 [Bao et al., 2024]	0.0818	0.1122	0.1210	0.0612	0.0881	0.0948	0.1015	0.1037	0.1064
DPO-based methods									
+ Vanilla DPO [Rafailov et al., 2023]	0.0389	0.0558	0.0622	0.0426	0.0710	0.0834	0.0816	0.0819	0.0838
+ DMPO [Bai et al., 2024]	0.0586	0.0805	0.0845	0.0461	0.0701	0.0758	0.0875	0.0893	0.0912
+ S-DPO [Chen et al., 2024]	0.0762	0.1105	0.1192	0.0592	0.0920	0.1049	0.0931	0.0938	0.0985
+ LiPO [Liu et al., 2025]	0.0862	0.1198	0.1294	0.0622	0.0980	0.1100	0.0963	0.1020	0.1060
+ Mult-DPO (ours)	**0.0947**	**0.1292**	**0.1406**	0.0650	**0.1001**	**0.1103**	**0.1097**	**0.1101**	**0.1154**
Qwen2.5-3B-Instruct (zero-shot)	0.0149	0.0184	0.0184	0.0232	0.0340	0.0359	0.0633	0.0617	0.0641
SFT-based methods									
+ BigRec [Bao et al., 2025]	0.1109	0.1527	0.1682	0.0747	0.1069	0.1180	0.1228	0.1247	0.1364
+ D^3 [Bao et al., 2024]	0.1254	0.1531	0.1678	0.0710	0.1038	0.1159	0.1195	0.1214	0.1332
DPO-based methods									
+ Vanilla DPO [Rafailov et al., 2023]	0.0870	0.1120	0.1202	0.0559	0.0806	0.0903	0.0915	0.0960	0.0998
+ DMPO [Bai et al., 2024]	0.0932	0.1200	0.1359	0.0562	0.0875	0.0938	0.0981	0.1009	0.1032
+ S-DPO [Chen et al., 2024]	0.1181	0.1586	0.1693	0.0631	0.0989	0.1122	0.1043	0.1132	0.1219
+ LiPO [Liu et al., 2025]	0.1252	0.1611	0.1731	0.0672	0.1046	0.1185	0.1147	0.1234	0.1329
+ Mult-DPO (ours)	**0.1288**	**0.1678**	**0.1785**	**0.0751**	**0.1155**	**0.1300**	**0.1369**	**0.1431**	**0.1503**

and erodes the alignment signal. The optimal β is consistently smaller on MovieLens-10M and Goodreads than on Reddit-V2. This may be because the reference backbone provides a stronger initialization for conversational recommendation, where the dialogue context can be exploited via language understanding [He et al., 2023], than for general recommendation, where collaborative-filtering signals that are unavailable to pretrained LLMs play a crucial role [Zhu et al., 2025].

4.3 Comparison with Baselines

We then evaluate Mult-DPO against representative baselines on the test set, where the best β for each DPO-style method is selected on the validation set as specified in Section 4.2. All DPO-style baselines share the same reference backbone and training protocol as Mult-DPO for fair comparison.

Baselines. We compare Mult-DPO against two categories of LLM-based recommenders: SFT-only baselines (BigRec [Bao et al., 2025] and D^3 [Bao et al., 2024]) that fine-tune the LLM on context–item demonstrations, and DPO-style alignment baselines (vanilla DPO [Rafailov et al., 2023], DMPO [Bai et al., 2024], S-DPO [Chen et al., 2024], and LiPO [Liu et al., 2025]). Full details are provided in Appendix C.2. For D^3 on conversational recommendation, we follow Zhu et al. [2024] to extract collaborative-filtering signals and train an EASE [Steck, 2019] model as the text-free assistant.

Results. From Table 1 we can find that DMPO extends vanilla DPO with multiple negatives for one positive item, but the arithmetic mean it takes for the negatives inside the BT sigmoid is an ad-hoc construction without a coherent ranking-likelihood interpretation. S-DPO replaces this mean with the softmax derived as the closed-form $k=1$ case of the marginalized PL likelihood (a special case of Theorem 2), providing a principled treatment of multiple negatives that respects the underlying preference model. LiPO further admits multi-positive supervision by summing NDCG-projected pair losses, yet it still factorizes the set-wise event into independent BT pairs and discards the constraint that the positives jointly dominate the negative set. In contrast, Mult-DPO fundamentally removes this spurious independence: we note that under the multinomial surrogate every positive is contrasted against the same cumulative negative weight B, which preserves the joint set-wise structure and is provably an upper bound on the marginalized PL-DPO loss (Corollary 1). The margin over LiPO is most pronounced on Reddit-V2, where multi-positive ground truths are densest, and widens with backbone scale, suggesting that the joint signal pays off most when the policy has the capacity to exploit it. Mult-DPO also surpasses BigRec and D^3 from the same SFT initialization, indicating

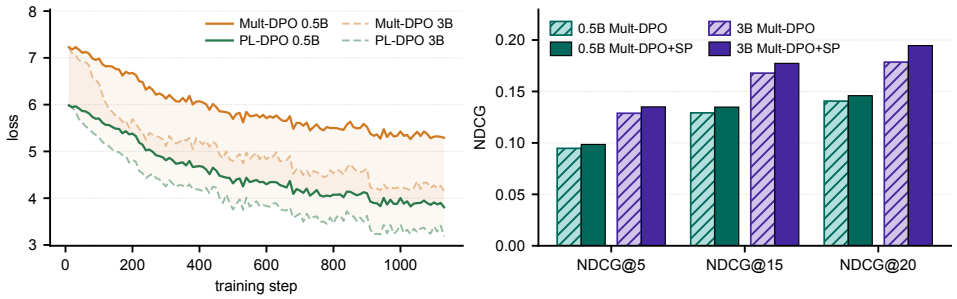


Figure 2: **Left**: Mult-DPO loss and exact marginalized PL-DPO loss; **Right**: NDCG on the `Goodreads` dataset of Mult-DPO with SPRec-style epoch-level dynamic hard negatives.

that set-wise alignment supplies supervision that demonstration matching cannot recover. Beyond accuracy, we note that compared with the strongest baseline LiPO, in Mult-DPO, the sharing B across positives reduces the loss-aggregation cost from $\mathcal{O}((K-k)\cdot k)$ to $\mathcal{O}(K)$.

4.4 On Upper Bound of MN-DPO to Marginalized PL-DPO and Its Tightness

For **RQ2**, we first restrict the training set to samples with at most three positives, for which the exact marginalized PL-DPO loss can be computed. As shown in Fig. 2 (left), on this filtered subset, the training dynamics of Mult-DPO and the exact marginalized PL-DPO verify the upper-bound relation in Corollary 1. However, this restriction removes most training instances, making the resulting test-set evaluation uninformative. To further examine whether harder negatives tighten the bound and thereby improve alignment with multi-positive supervision, we consider dynamic negative sampling. Since the hardness of a negative item is policy-dependent and changes throughout training, we adapt SPRec [Gao et al., 2025] and resample negatives at the epoch level. Concretely, at the beginning of each epoch, negatives are drawn from a temperature-scaled categorical distribution induced by the current policy weights with temperature set to 0.1. Fig. 2 (right) reports the test performance, where we find that introducing harder negatives can indeed improve the alignment ability of Mult-DPO.

4.5 Extension to Multi-Level Preference on the MovieLens Dataset

To answer **RQ3**, on `MovieLens-10M`, we partition items by their ratings into $G = 4$ ordered preference groups, with rating $= 5$ (i.e., the positive set in the two-set case) as the group with the highest preference, and with randomly sampled unrated items appended after the high-rated groups (rating > 3) to keep the candidate set comparable to the binary setting of Section 4.3. Mult^2-DPO aggregates the per-boundary multinomial losses across the three group boundaries via Eq. (24), and is compared against the binary Mult-DPO baseline that merges the four groups into one positive-negative split. Figure 3 shows that Mult^2-DPO outperforms the binary counterpart at every cutoff, with a $\sim 12\%$ gain in NDCG at the 0.5B backbone (0.0732 vs. 0.0650 at NDCG@5). In addition, the improvement can be generalized to the 3B backbone. This further confirms that Mult^2-DPO provides a stronger alignment signal by preserving the richer preference structure of explicit ratings.

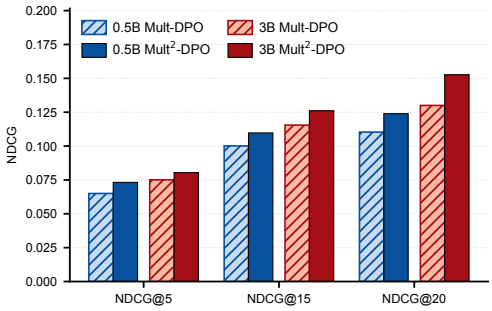


Figure 3: Comparison of Mult-DPO and Mult^2-DPO NDCG on `MovieLens-10M` dataset.

5 Conclusion

In this paper, we presented **Mult-DPO**, a new DPO-style framework for aligning LLM-based RSs with preferences over multiple candidates. The key contribution is a multinomial surrogate event likelihood that yields a tractable closed-form DPO-style loss that is a provable upper bound on the otherwise intractable marginalized Plackett-Luce loss. We also characterize the tightness of this bound in

terms of the number of positives and the cumulative policy-induced weights of the selected negatives. Empirically, Mult-DPO and its multi-level extension, i.e., Mult^2-DPO, consistently outperform various DPO-based baselines on both general and conversational recommendation benchmarks, suggesting that explicitly modeling the joint set-wise structure of user feedback is both effective and efficient for preference-based alignment in LLM-based recommendation tasks.

Acknowledgment

Yaochen Zhu, Yinhan He, and Jundong Li are supported in part by the Office of Naval Research (ONR) under grant N000142412636 and the gift funding from Netflix.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *AISTATS*, pages 4447–4455. PMLR, 2024.

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

Zhuoxi Bai, Ning Wu, Fengyu Cai, Xinyi Zhu, and Yun Xiong. Aligning large language model with direct multi-preference optimization for recommendation. In *CIKM*, pages 76–86, 2024.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. TALLRec: An effective and efficient tuning framework to align large language model with recommendation. In *RecSys*, 2023.

Keqin Bao, Jizhi Zhang, Yang Zhang, Xinyue Huo, Chong Chen, and Fuli Feng. Decoding matters: Addressing amplification bias and homogeneity issue for LLM-based recommendation. In *EMNLP*, 2024.

Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yanchen Luo, Chong Chen, Fuli Feng, and Qi Tian. A bi-step grounding paradigm for large language models in recommendation systems. *ACM Transactions on Recommender Systems*, 3(4):1–27, 2025.

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

Shihao Cai, Chongming Gao, Yang Zhang, Wentao Shi, Jizhi Zhang, Keqin Bao, Qifan Wang, and Fuli Feng. K-order ranking preference optimization for large language models. In *Findings of ACL*, pages 4844–4859, 2025.

Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. On softmax direct preference optimization for recommendation. In *NeurIPS*, volume 37, pages 27463–27489, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, volume 30, 2017.

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

Chongming Gao, Ruijun Chen, Shuai Yuan, Kexin Huang, Yuanqing Yu, and Xiangnan He. SPRec: Self-play to debias LLM-based recommendation. In *WWW*, pages 5075–5084, 2025.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (RLP): A unified pretrain, personalized prompt & predict paradigm (P5). In *RecSys*, pages 299–315, 2022.

F Maxwell Harper and Joseph A Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. Large language models as zero-shot conversational recommenders. In *CIKM*, pages 720–730, 2023.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *ECIR*, 2024.

Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *WWW*, pages 689–698, 2018.

Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. LLaRA: Large language-recommendation assistant. In *SIGIR*, 2024.

Jiacheng Lin, Tian Wang, and Kun Qian. Rec-R1: Bridging generative large language models and user-centric recommendation systems via reinforcement learning. *Transactions on Machine Learning Research*, 2025.

Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, et al. How can recommender systems benefit from large language models: A survey. *ACM Transactions on Information Systems*, 2023.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. In *ICLR*, 2023.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. LiPO: Listwise preference optimization through learning-to-rank. In *NAACL*, pages 2404–2420, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

R Duncan Luce. *Individual Choice Behavior*. Wiley, 1959.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *NeurIPS*, volume 37, pages 124198–124235, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, volume 35, pages 27730–27744, 2022.

Richard Y Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. In *NeurIPS*, volume 37, pages 116617–116637, 2024.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C*, 24 (2):193–202, 1975.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Lin, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, volume 36, pages 53728–53741, 2023.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC*, 2020.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *UAI*, pages 452–461, 2009.

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

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *AAAI*, volume 38, pages 18990–18998, 2024.

Harald Steck. Embarrassingly shallow autoencoders for sparse data. In *WWW*, pages 3251–3257, 2019.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *NeurIPS*, 33: 3008–3021, 2020.

Mengting Wan and Julian McAuley. Item recommendation on monotonic behavior chains. In *RecSys*, 2018.

Martin Weyssow, Aton Kamanda, Xin Zhou, and Houari Sahraoui. CodeUltraFeedback: An LLM-as-a-judge dataset for aligning large language models to coding preferences. *ACM Transactions on Software Engineering and Methodology*, 35(3):1–36, 2026.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. A survey on large language models for recommendation. *World Wide Web*, 27(5):60, 2024.

Junjie Zhang, Ruobing Xie, Yupeng Hou, Xin Zhao, Leyu Lin, and Ji-Rong Wen. Recommendation as instruction following: A large language model empowered recommendation approach. *ACM Transactions on Information Systems*, 43(5):1–37, 2025.

Yaochen Zhu, Liang Wu, Qi Guo, Liangjie Hong, and Jundong Li. Collaborative large language model for recommender systems. In *WWW*, pages 3162–3172, 2024.

Yaochen Zhu, Chao Wan, Harald Steck, Dawen Liang, Yesu Feng, Nathan Kallus, and Jundong Li. Collaborative retrieval for large language model-based conversational recommender systems. In *WWW*, pages 3323–3334, 2025.

Yaochen Zhu, Harald Steck, Dawen Liang, Yinhan He, Vito Ostuni, Jundong Li, and Nathan Kallus. Rank-GRPO: Training LLM-based conversational recommender systems with reinforcement learning. In *ICLR*, 2026.

Appendix

A Proofs

In this section, we provide the proofs and derivations for Section 3 in the main paper, including the ***(i)*** marginalized Plackett-Luce likelihood, ***(ii)*** its inclusion-exclusion form, ***(iii)*** the multinomial surrogate likelihood, ***(iv)*** the lower-bound relation between MN surrogate and PL likelihoods, ***(v)*** the tightness analysis, and ***(vi)*** the extension of Mult-DPO to multi-level preference alignment of LLMs.

A.1 Full-Rank Plackett-Luce Model and Its Recursions.

Recall that in the main paper, we use $\mathcal{E}$ to denote the candidate set associated with a recommendation-related context x. Let $K := |\mathcal{E}|$ and let τ be a permutation of $\{1, \dots, K\}$ representing a full ranking of the items in $\mathcal{E}$. We write $e_{\tau(t)}$ for the item placed at rank t under τ. For any $t \in \{1, \dots, K\}$, define the prefix ranking of length t and the suffix ranking of the remaining items as

$$\tau_{1:t} := (\tau(1), \tau(2), \dots, \tau(t)), \;\; \tau_{t+1:K} := (\tau(t+1), \tau(t+2), \dots, \tau(K)).$$

The suffix candidate set after selecting the first t ranked items is defined as

$$\mathcal{E}_{t+1} := \mathcal{E} \setminus \{e_{\tau(1)}, \dots, e_{\tau(t)}\}.$$

The Plackett-Luce (PL) [Plackett, 1975, Luce, 1959] model assigns the likelihood to a full ranking τ:

$$p_{\mathrm{PL}}(\tau \mid x, \mathcal{E}; w) = \prod_{j=1}^{K} \frac{w(e_{\tau(j)})}{\sum_{l=j}^{K} w(e_{\tau(l)})} \tag{27a}$$

$$= \frac{w(e_{\tau(1)})}{\sum_{e \in \mathcal{E}} w(e)} \cdot \prod_{j=2}^{K} \frac{w(e_{\tau(j)})}{\sum_{l=j}^{K} w(e_{\tau(l)})}$$

$$= \frac{w(e_{\tau(1)})}{\sum_{e \in \mathcal{E}} w(e)} \cdot p_{\mathrm{PL}}(\tau_{2:K} \mid x, \mathcal{E}_2; w) \tag{27b}$$

$$= \left(\prod_{t=1}^{k} \frac{w(e_{\tau(t)})}{\sum_{e \in \mathcal{E}_t} w(e)} \right) \cdot \prod_{j=k+1}^{K} \frac{w(e_{\tau(j)})}{\sum_{l=j}^{K} w(e_{\tau(l)})}$$

$$= \left(\prod_{t=1}^{k} \frac{w(e_{\tau(t)})}{\sum_{e \in \mathcal{E}_t} w(e)} \right) \cdot p_{\mathrm{PL}}(\tau_{k+1:K} \mid x, \mathcal{E}_{k+1}; w)\,, \tag{27c}$$

where Eqs. (27b) and (27c) are the one-step and k-step recursions, respectively.

Lemma 1 (PL normalization on a finite candidate set)**.** *Let $\mathcal{E} = \{e_1, \dots, e_K\}$ be a finite candidate set and let S_K denote the set of all permutations of $\{1, \dots, K\}$. Then the PL likelihood in Eq.* (27a) *defines a valid distribution over full rankings:*

$$\sum_{\tau \in S_K} p_{\mathrm{PL}}(\tau \mid x, \mathcal{E}; w) = 1. \tag{28}$$

Proof. When $K = 1$, Eq. (28) trivially holds. For arbitrary K, we have

$$\sum_{\tau \in S_K} p_{\mathrm{PL}}(\tau \mid x, \mathcal{E}; w) = \sum_{i=1}^{K} \sum_{\rho \in \mathrm{Perm}(\mathcal{E} \setminus \{e_i\})} p_{\mathrm{PL}}\big([e_i, \rho] \mid x, \mathcal{E}; w\big) \tag{29a}$$

$$= \sum_{i=1}^{K} \sum_{\rho \in \mathrm{Perm}(\mathcal{E} \setminus \{e_i\})} \frac{w(e_i)}{\sum_{i=1}^{K} w(e_i)} \cdot p_{\mathrm{PL}}\big(\rho \mid x, \mathcal{E} \setminus \{e_i\}; w\big) \tag{29b}$$

$$= \sum_{i=1}^{K} \frac{w(e_i)}{\sum_{i=1}^{K} w(e_i)} \left(\sum_{\rho \in \mathrm{Perm}(\mathcal{E} \setminus \{e_i\})} p_{\mathrm{PL}}\big(\rho \mid x, \mathcal{E} \setminus \{e_i\}; w\big) \right)$$

$$= \sum_{i=1}^{K} \frac{w(e_i)}{\sum_{i=1}^{K} w(e_i)} = 1, \tag{29c}$$

where the last step follows by induction on K. □

A.2 Proof of the Marginalized PL Likelihood in Eq. (8).

Recall that the positives are $\mathcal{E}^p = \{e_1, \dots, e_k\}$ and the negatives are $\mathcal{E}^d = \{e_{k+1}, \dots, e_K\}$. Let S_k denote the set of permutations of $\{1, \dots, k\}$ for the positive items. Then

$$
\begin{aligned}
p_{\mathrm{PL}}(\Omega_x \mid x, \mathcal{E}; w) &= \sum_{\rho \in S_k} \sum_{\tau_{k+1:K}} p_{\mathrm{PL}}([\rho, \tau_{k+1:K}] \mid x, \mathcal{E}; w) && \text{(30a)} \\
&= \sum_{\rho \in S_k} \sum_{\tau_{k+1:K}} \left(\prod_{t=1}^{k} \frac{w(e_{\rho(t)})}{\sum_{e \in \mathcal{E} \setminus \{e_{\rho(1)}, \dots, e_{\rho(t-1)}\}} w(e)} \right) p_{\mathrm{PL}}\big(\tau_{k+1:K} \mid x, \mathcal{E}^d; w\big) && \text{(30b)} \\
&= \sum_{\rho \in S_k} \left(\prod_{t=1}^{k} \frac{w(e_{\rho(t)})}{\sum_{e \in \mathcal{E} \setminus \{e_{\rho(1)}, \dots, e_{\rho(t-1)}\}} w(e)} \right) \sum_{\tau_{k+1:K}} p_{\mathrm{PL}}\big(\tau_{k+1:K} \mid x, \mathcal{E}^d; w\big) \\
&= \sum_{\rho \in S_k} \prod_{t=1}^{k} \frac{w(e_{\rho(t)})}{\sum_{e \in \mathcal{E} \setminus \{e_{\rho(1)}, \dots, e_{\rho(t-1)}\}} w(e)} && \text{(30c)} \\
&= \sum_{\rho \in S_k} \prod_{t=1}^{k} \frac{w(e_{\rho(t)})}{B + \sum_{j=t}^{k} w(e_{\rho(j)})}. && \text{(30d)}
\end{aligned}
$$

In the last step, the remaining candidate set at rank t consists of the unselected positives $\{e_{\rho(t)}, \dots, e_{\rho(k)}\}$ and all negatives $\mathcal{E}^d$, so the denominator decomposes into $B + \sum_{j=t}^{k} w(e_{\rho(j)})$. Note that the marginalization over the latent rankings of the negatives vanishes from step (30b) to step (30c) because the PL likelihood over the suffix ranking of $\mathcal{E}^d$ sums to one by Lemma 1.

A.3 Proof of the Inclusion-Exclusion Form for the Marginalized PL Likelihood.

Lemma 2 (Exponential-race representation)**.** *Let $\{T_e\}_{e \in \mathcal{E}}$ be independent random variables with $T_e \sim \mathrm{Exp}(w(e))$. Ranking items by increasing T_e induces the PL distribution in Eq.* (27)*.*

Proof. This is a standard characterization of the Plackett-Luce model. The minimum exponential clock is selected first with probability proportional to its rate, and the memoryless property gives the same rule recursively for the remaining items. □

Using Lemma 2, we have

$$p_{\mathrm{PL}}(\Omega_x \mid x, \mathcal{E}; w) = \Pr\left(\max_{e\in\mathcal{E}^p} T_e < \min_{e\in\mathcal{E}^d} T_e\right) \tag{31a}$$

$$= \Pr(T_e < U \ \ \forall e \in \mathcal{E}^p)\,, \ \ U := \min_{e\in\mathcal{E}^d} T_e$$

$$= \int_0^\infty \Pr(T_e < u \ \ \forall e \in \mathcal{E}^p)\ f_U(u)\, du \tag{31b}$$

$$= \int_0^\infty Be^{-Bu} \prod_{e\in\mathcal{E}^p} \Pr(T_e < u)\, du \tag{31c}$$

$$= \int_0^\infty Be^{-Bu} \prod_{e\in\mathcal{E}^p} \left(1 - e^{-w(e)u}\right) du \tag{31d}$$

$$= \int_0^\infty Be^{-Bu} \sum_{S\subseteq\mathcal{E}^p} (-1)^{|S|} \prod_{e\in S} e^{-w(e)u}\, du \tag{31e}$$

$$= \sum_{S\subseteq\mathcal{E}^p} (-1)^{|S|} \int_0^\infty B \exp\left(-u\Big(B + \sum_{e\in S} w(e)\Big)\right) du$$

$$= \sum_{S\subseteq\mathcal{E}^p} (-1)^{|S|} \frac{B}{B + \sum_{e\in S} w(e)}. \tag{31f}$$

This proves the inclusion-exclusion form for the marginalized PL likelihood, which reduces the number of terms in naive marginalization (i.e., Eq. (8)) from $k!$ to 2^k. However, it still remains intractable for moderate k, motivating the multinomial surrogate developed in Mult-DPO.

A.4 Derivation of the Multinomial Surrogate Likelihood in Eq. (10).

We formally derive the multinomial surrogate likelihood. Recall from Section 3.1 that the categorical distribution over $\mathcal{E}$ is defined as $p(e \mid x) = w(e \mid x)/W$. The MN surrogate assigns to Ω_x the probability of the count event induced by k i.i.d. draws from $p(\cdot \mid x)$ in which each positive item in $\mathcal{E}^p$ is sampled exactly once and no negative item in $\mathcal{E}^d$ is sampled. Letting $c(e)$ denote the count of item e across the k draws, this event can be formulated as

$$c(e) = 1 \ \forall e \in \mathcal{E}^p, \ \ c(e) = 0 \ \forall e \in \mathcal{E}^d.$$

Substituting these counts into the multinomial probability mass function then yields

$$p_{\mathrm{MN}}(\Omega_x \mid x, \mathcal{E}; w) = \frac{k!}{\prod_{e\in\mathcal{E}} c(e)!} \prod_{e\in\mathcal{E}} p(e \mid x)^{c(e)} \tag{32a}$$

$$= k! \prod_{e\in\mathcal{E}^p} p(e \mid x) \tag{32b}$$

$$= k! \prod_{e\in\mathcal{E}^p} \frac{w(e \mid x)}{W}. \tag{32c}$$

A.5 Proof of Theorem 1.

We now prove the main theorem of the paper, i.e., the multinomial surrogate likelihood is a pointwise lower bound on the intractable marginalized PL likelihood. Recall from Eq. (8) that

$$p_{\mathrm{PL}}(\Omega_x \mid x, \mathcal{E}; w) = \sum_{\rho\in S_k} \prod_{t=1}^{k} \frac{w(e_{\rho(t)} \mid x)}{B + \sum_{j=t}^{k} w(e_{\rho(j)} \mid x)}.$$

Since the per-rank denominator satisfies $B + \sum_{j=t}^{k} w(e_{\rho(j)} \mid x) \le B + \sum_{e \in \mathcal{E}^p} w(e \mid x) = W$, we obtain the following chain of pointwise inequalities for any $\rho \in S_k$ and any $t \in \{1, \dots, k\}$:

$$\begin{aligned} & \frac{1}{B + \sum_{j=t}^{k} w(e_{\rho(j)} \mid x)} \ge \frac{1}{W} \\ \Longrightarrow \quad & \frac{w(e_{\rho(t)} \mid x)}{B + \sum_{j=t}^{k} w(e_{\rho(j)} \mid x)} \ge \frac{w(e_{\rho(t)} \mid x)}{W} \\ \Longrightarrow \quad & \prod_{t=1}^{k} \frac{w(e_{\rho(t)} \mid x)}{B + \sum_{j=t}^{k} w(e_{\rho(j)} \mid x)} \ge \prod_{t=1}^{k} \frac{w(e_{\rho(t)} \mid x)}{W}. \end{aligned}$$

Summing over $\rho \in S_k$ gives

$$p_{\mathrm{PL}}(\Omega_x \mid x, \mathcal{E}; w) \ge \sum_{\rho \in S_k} \prod_{t=1}^{k} \frac{w(e_{\rho(t)} \mid x)}{W} = k! \prod_{e \in \mathcal{E}^p} \frac{w(e \mid x)}{W} = p_{\mathrm{MN}}(\Omega_x \mid x, \mathcal{E}; w).$$

A.6 Proof of Theorem 2.

We then characterize the tightness of the lower bound, i.e., the ratio between the marginalized PL likelihood and the multinomial surrogate likelihood is sandwiched between 1 and $(1 + A/B)^{k-1}$. Specifically, we first fix an arbitrary $\rho \in S_k$ and define the prefix sums of positive weights

$$H_{t-1}(\rho) := \sum_{j=1}^{t-1} w(e_{\rho(j)} \mid x), \;\; t = 2, \dots, k.$$

For the PL model, the denominator at rank t is

$$D_t(\rho) = B + \sum_{j=t}^{k} w(e_{\rho(j)} \mid x) = W - H_{t-1}(\rho).$$

Therefore, the ratio between the per-ρ PL term and the corresponding MN term satisfies

$$\begin{aligned} \frac{\prod_{t=1}^{k} \frac{w(e_{\rho(t)} \mid x)}{D_t(\rho)}}{\prod_{t=1}^{k} \frac{w(e_{\rho(t)} \mid x)}{W}} &= \prod_{t=2}^{k} \frac{W}{W - H_{t-1}(\rho)} && \text{(33a)} \\ &= \prod_{t=2}^{k} \frac{1}{1 - \frac{H_{t-1}(\rho)}{W}} && \text{(33b)} \\ &\le \prod_{t=2}^{k} \frac{1}{1 - \frac{A}{W}} = \left(\frac{W}{W - A} \right)^{k-1} = \left(1 + \frac{A}{B} \right)^{k-1}. && \text{(33c)} \end{aligned}$$

Since this bound holds uniformly over all $\rho \in S_k$, summing over ρ yields

$$1 \le \frac{p_{\mathrm{PL}}(\Omega_x \mid x, \mathcal{E}; w)}{p_{\mathrm{MN}}(\Omega_x \mid x, \mathcal{E}; w)} \le \left(1 + \frac{A}{B} \right)^{k-1}.$$

A.7 Derivation for the DPO-style Extension.

In this part, we connect weights in both the MN surrogate and the marginalized PL event model with the LLM-based RS policy π_θ through the analytical solution of the RLHF objective, i.e., $w_{\pi_\theta}(e \mid x)$, so that both bounds hold for the corresponding DPO-style losses. Let $\pi_{\mathrm{ref}}(e \mid x)$ be a reference policy and $\pi_\theta(e \mid x)$ a learned policy. The RLHF solution gives weights proportional to $(\pi_\theta(e \mid x)/\pi_{\mathrm{ref}}(e \mid x))^\beta$, with a context-dependent factor $Z(x)^\beta$ shared by all candidates. Since

both PL and MN likelihoods are invariant to multiplying all weights in the same candidate set by a common positive constant, we fix the representative

$$r_{\pi_\theta}(e \mid x) := \beta \log \frac{\pi_\theta(e \mid x)}{\pi_{\text{ref}}(e \mid x)}, \;\; w_{\pi_\theta}(e \mid x) := \exp(r_{\pi_\theta}(e \mid x)) = \left(\frac{\pi_\theta(e \mid x)}{\pi_{\text{ref}}(e \mid x)}\right)^\beta.$$

If we define $p_{\text{PL}}(\Omega_x \mid x, \mathcal{E}; w_{\pi_\theta})$ and $p_{\text{MN}}(\Omega_x \mid x, \mathcal{E}; w_{\pi_\theta})$ by replacing $w(e \mid x)$ with $w_{\pi_\theta}(e \mid x)$ in Eqs. (8) and (10), the lower bound and the tightness bound hold pointwise with A, B replaced by

$$A_{\pi_\theta} := \sum_{e \in \mathcal{E}^p} w_{\pi_\theta}(e \mid x), \;\; B_{\pi_\theta} := \sum_{e \in \mathcal{E}^d} w_{\pi_\theta}(e \mid x).$$

Equivalently, taking negative logarithms reverses the likelihood inequality: the Mult-DPO loss $\mathcal{L}^{\text{Mult-DPO}}(\pi_\theta) := -\log p_{\text{MN}}(\Omega_x \mid x, \mathcal{E}; w_{\pi_\theta})$ upper bounds the marginalized PL-DPO loss $\mathcal{L}^{\text{PL-DPO}}(\pi_\theta) := -\log p_{\text{PL}}(\Omega_x \mid x, \mathcal{E}; w_{\pi_\theta})$.

A.8 Proof of the Multi-Level PL Factorization in Eq. (22).

We further extend the theoretical analysis to the multi-level preference alignment setting discussed in Section 3.4, where user preference takes the form of the multi-level preference event Ω_x^{grp} defined in Eq. (20). We show that Ω_x^{grp} factorizes recursively across the boundaries between consecutive groups, reducing the multi-level case to a product of set-wise preference instances each handled by Theorem 1 and Theorem 2. By definition, the full multi-level preference event satisfies

$$\Omega_x^{\text{grp}} = \bigcap_{g=1}^{G-1} \Omega_x^{(g)}.$$

By the sequential selection property of PL, once items from the higher-preference groups $\mathcal{E}^{(1)}, \ldots, \mathcal{E}^{(g-1)}$ have been placed, the relative ordering among the remaining groups depends only on the remaining candidate set $\bigcup_{h=g}^{G} \mathcal{E}^{(h)}$, so the events $\Omega_x^{(g)}$ become conditionally independent across g. Therefore, the likelihood factorizes recursively as

$$\begin{aligned}
p_{\text{PL}}(\Omega_x^{\text{grp}} \mid x, \mathcal{E}; w) &= p_{\text{PL}}\left(\Omega_x^{(1)} \mid x, \bigcup_{h=1}^{G} \mathcal{E}^{(h)}; w\right) \cdot p_{\text{PL}}\left(\Omega_x^{\text{grp}} \text{ restricted to } \bigcup_{h=2}^{G} \mathcal{E}^{(h)}\right) \\
&= p_{\text{PL}}\left(\Omega_x^{(1)} \mid x, \bigcup_{h=1}^{G} \mathcal{E}^{(h)}; w\right) \cdot p_{\text{PL}}\left(\Omega_x^{(2)} \mid x, \bigcup_{h=2}^{G} \mathcal{E}^{(h)}; w\right) \cdot \ldots \\
&= \prod_{g=1}^{G-1} p_{\text{PL}}\left(\Omega_x^{(g)} \mid x, \bigcup_{h=g}^{G} \mathcal{E}^{(h)}; w\right),
\end{aligned} \tag{34}$$

which proves Eq. (22). For each fixed group g, the group-g-specific preference event $\Omega_x^{(g)}$ is exactly a set-wise preference event of the same form as Eq. (1), with $\mathcal{E}^{(g)}$ playing the role of $\mathcal{E}^p$ and $\bigcup_{h=g+1}^{G} \mathcal{E}^{(h)}$ playing the role of $\mathcal{E}^d$. Let

$$A_g := \sum_{e \in \mathcal{E}^{(g)}} w(e \mid x), \quad B_g := \sum_{h=g+1}^{G} \sum_{e \in \mathcal{E}^{(h)}} w(e \mid x), \quad W_g := A_g + B_g$$

denote the current-group, lower-group, and total remaining weights, respectively. Therefore, reusing the binary results yields

$$p_{\text{PL}}\left(\Omega_x^{(g)} \mid x, \bigcup_{h=g}^{G} \mathcal{E}^{(h)}; w\right) = \sum_{\rho \in S_{k_g}} \prod_{t=1}^{k_g} \frac{w(e_{\rho(t)}^{(g)} \mid x)}{B_g + \sum_{j=t}^{k_g} w(e_{\rho(j)}^{(g)} \mid x)}, \tag{35}$$

and equivalently

$$p_{\text{PL}}\left(\Omega_x^{(g)} \mid x, \bigcup_{h=g}^{G} \mathcal{E}^{(h)}; w\right) = \sum_{S \subseteq \mathcal{E}^{(g)}} (-1)^{|S|} \frac{B_g}{B_g + \sum_{e \in S} w(e \mid x)}. \tag{36}$$

A.9 Proof of the Sequential Multinomial Lower Bound and Tightness.

We finally prove that the DPO loss derived from the sequential multinomial (SMN) surrogate, i.e., the Mult^2-DPO loss in Eq. (24), is an upper bound on the DPO loss derived from the marginalized PL reward model in Eq. (22) in the multi-level setting. From Eq. (22) we have

$$p_{\mathrm{PL}}(\Omega_x^{\mathrm{grp}} \mid x, \mathcal{E}; w) = \prod_{g=1}^{G-1} p_{\mathrm{PL}}\left(\Omega_x^{(g)} \mid x, \bigcup_{h=g}^{G} \mathcal{E}^{(h)}; w\right). \tag{37}$$

For each group g, applying Theorem 1 to the group-g-specific preference event $\Omega_x^{(g)}$ on the remaining candidate set $\bigcup_{h=g}^{G} \mathcal{E}^{(h)}$ yields

$$p_{\mathrm{PL}}\left(\Omega_x^{(g)} \mid x, \bigcup_{h=g}^{G} \mathcal{E}^{(h)}; w\right) \geq k_g! \prod_{e \in \mathcal{E}^{(g)}} \frac{w(e \mid x)}{W_g}. \tag{38}$$

Here W_g is the total remaining weight from group g onward. Multiplying Eq. (38) over $g = 1, \ldots, G-1$ and using Eq. (37) gives

$$p_{\mathrm{PL}}(\Omega_x^{\mathrm{grp}} \mid x, \mathcal{E}; w) \geq \prod_{g=1}^{G-1} \left(k_g! \prod_{e \in \mathcal{E}^{(g)}} \frac{w(e \mid x)}{W_g}\right) = p_{\mathrm{SMN}}(\Omega_x^{\mathrm{grp}} \mid x, \mathcal{E}; w),$$

which establishes the SMN lower bound on the multi-level PL likelihood.

The tightness bound follows by applying Theorem 2 to each factor in Eq. (37). For every $g = 1, \ldots, G-1$,

$$1 \leq \frac{p_{\mathrm{PL}}\left(\Omega_x^{(g)} \mid x, \bigcup_{h=g}^{G} \mathcal{E}^{(h)}; w\right)}{k_g! \prod_{e \in \mathcal{E}^{(g)}} \frac{w(e|x)}{W_g}} \leq \left(1 + \frac{A_g}{B_g}\right)^{k_g - 1}.$$

Multiplying these inequalities across groups gives

$$1 \leq \frac{p_{\mathrm{PL}}(\Omega_x^{\mathrm{grp}} \mid x, \mathcal{E}; w)}{p_{\mathrm{SMN}}(\Omega_x^{\mathrm{grp}} \mid x, \mathcal{E}; w)} \leq \prod_{g=1}^{G-1} \left(1 + \frac{A_g}{B_g}\right)^{k_g - 1}.$$

Substituting the policy-induced weight $w_{\pi_\theta}(e \mid x)$ and taking negative logarithms flips the likelihood inequality, yielding the Mult^2-DPO loss in Eq. (24) as a tractable upper bound on the otherwise intractable multi-level marginalized PL-DPO loss and proving Corollary 2. Here, we note that when $G = 2$, the bound reduces exactly to the binary Mult-DPO upper bound established in Section 3.2.

B Related Work

B.1 LLM-based Recommender Systems

Recently, LLM-based recommender systems (RSs) have attracted growing attention, where LLMs' knowledge and reasoning ability can be leveraged to generate high-quality recommendations [Lin et al., 2023, Wu et al., 2024]. Early works show that zero-shot LLMs can already serve as effective RSs by ranking candidate items provided in the prompt [Hou et al., 2024, He et al., 2023]. Post-training approaches further adapt pretrained LLMs to recommendation tasks, encompassing both supervised fine-tuning (SFT) and preference-based alignment. SFT-based methods fine-tune the LLM on prompt-completion pairs of context and item-recommendation demonstrations, with exemplar methods such as TallRec [Bao et al., 2023], BIGRec [Bao et al., 2025], P5 [Geng et al., 2022], InstructRec [Zhang et al., 2025], and LLaRA [Liao et al., 2024]. To further align LLM-based recommenders with user preferences beyond SFT, more recent works employ preference-based alignment. RL-free DPO-based methods include S-DPO [Chen et al., 2024], DMPO [Bai et al., 2024], and SPRec [Gao et al., 2025] (see Section B.2 for details). A parallel line built on reinforcement learning (RL), particularly group relative policy optimization (GRPO) [Shao et al., 2024], generates a list of recommendations via RL: Rec-R1 [Lin et al., 2025] and Rank-GRPO [Zhu et al., 2026] sample diverse recommendation trajectories from the LLM and update the policy with corresponding reward signals. However, the quadratic cost of autoregressive item generation limits these methods to generating only a small number of items per query. In contrast, our Mult-DPO natively supports multi-positive supervision and scales to full ranking over the entire catalog.

B.2 Direct Preference Optimization and Its Role in RS

Aligning LLMs with user preferences is a critical challenge. Reinforcement learning from human feedback (RLHF) [Christiano et al., 2017, Ouyang et al., 2022] first trains a reward model from pairwise preference data and then optimizes the LLM policy against this reward via RL. Direct preference optimization (DPO) [Rafailov et al., 2023] and its variants (including SimPO [Meng et al., 2024], IPO [Azar et al., 2024], and KTO [Ethayarajh et al., 2024]) exploit a closed-form connection between the optimal policy and the reward function in RLHF and reduce alignment to maximum likelihood estimation. Adapting DPO to RSs, however, introduces additional challenges due to the multi-item nature of user feedback. Prior work has explored two main strategies. The first generalizes the Bradley-Terry (BT) model underlying vanilla DPO to its listwise counterpart, the Plackett-Luce (PL) model, with methods such as PRO [Song et al., 2024] and KPO [Cai et al., 2025] fitting PL likelihoods to fully ordered candidate lists. The second restricts supervision to a single positive paired with multiple negatives: S-DPO [Chen et al., 2024] adopts a softmax-style loss corresponding to the single-positive special case of the marginalized PL likelihood, DMPO [Bai et al., 2024] contrasts the positive against the arithmetic mean of negative log-ratios via an ad-hoc BT sigmoid, and SPRec [Gao et al., 2025] introduces self-play with model-generated negatives. Nevertheless, neither strategy faithfully captures the joint multi-positive/multi-negative structure of user preference in RSs.

C Additional Experiments and Details

C.1 Implementation Details

We run all experiments on the 0.5B-1.5B backbones with 2 NVIDIA H100 GPUs, and on the 3B-7B backbones with 2 NVIDIA B200 GPUs. Since the maximum sequence length and per-instance candidate count vary across the three datasets, we adopt different parallelism strategies, ranging from pure data parallelism to ZeRO-2 [Rajbhandari et al., 2020], while keeping the parallelism strategy identical across baselines for fair comparison. All methods are optimized with AdamW [Loshchilov and Hutter, 2019] at a learning rate of $1\mathrm{e}{-6}$. Across both training and inference, we always reuse the KV cache of the prompt across the candidate items associated with the same context, which we find substantially reduces the total compute as detailed in the complexity analysis in Section 3.5.

C.2 Baselines

The baselines that we include in the main paper are summarized as follows:

- **BIGRec** [Bao et al., 2025] is an instruction-tuning framework for LLM-based recommendation that prompts the LLM to generate candidate items from a user's interaction history and then grounds the generated outputs to exact items in the catalog, providing a strong SFT-only baseline.
- **D^3** [Bao et al., 2024] improves alignment at decoding time through a debiasing-diversifying decoding strategy that mitigates amplification bias and homogeneity in LLM-based recommendation outputs, without modifying the underlying training procedure.
- **Vanilla DPO** [Rafailov et al., 2023] is the original DPO objective formulated under the Bradley-Terry (BT) model, where each training instance contains one positive and one negative item.
- **DMPO** [Bai et al., 2024] adapts DPO to recommendation by contrasting one positive item against the arithmetic mean of the negative log-ratios within the Bradley-Terry preference model, enabling preference-based alignment with multiple sampled negatives per context.
- **S-DPO** [Chen et al., 2024] extends vanilla DPO from a single negative to multiple negatives by optimizing a softmax-based preference objective over one preferred item and several dispreferred items, which corresponds to the single-positive special case of Mult-DPO under Theorem 2.
- **LiPO** [Liu et al., 2025] formulates listwise preference optimization through learning-to-rank, aggregating weighted positive-negative pairs at each step, with each pair weighted by the change in projected NDCG to align LLMs with full rankings with human-provided scores.

C.3 Evaluation of Mult-DPO with More Backbone Models

To complement the experimental results with 0.5B and 3B backbone models in the main text, we additionally train Mult-DPO on `Goodreads` with the Qwen2.5-1.5B and Qwen2.5-7B backbones, and report test NDCG@$\{5, 15, 20\}$ in Fig. 4. Rather than re-running the expensive β sweep at every backbone scale, we reuse the optimal β=0.005 identified from the 0.5B and 3B validation sweeps in Section 4.2 and apply it to the 1.5B and 7B backbone models. Empirically, we find that NDCG climbs steeply from 0.5B to 1.5B backbone model and gradually flattens between 1.5B, 3B, and 7B, with each successive doubling in scale yielding a smaller absolute gain. This pattern indicates that Mult-DPO already extracts most of the available set-wise signal at moderate scale, while the continued improvement at 7B confirms that the benefit of set-wise alignment can be potentially generalized to large backbone regimes.

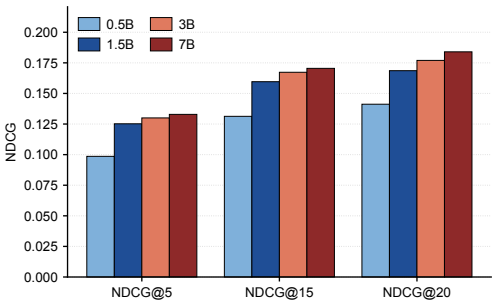


Figure 4: Mult-DPO test NDCG on `Goodreads` across four Qwen2.5 backbones at β=0.005 evaluated on the full item catalog.

D Limitations

We note that the multinomial (MN) construction developed in Mult-DPO is a surrogate event likelihood rather than a normalized ranking distribution. It provides one tractable surrogate for the intractable marginalized Plackett-Luce (PL) likelihood that simultaneously admits a closed-form DPO-style optimization, but it may not be the only or the tightest such surrogate. Because the resulting objective uses a uniform target over positives at the same preference level, it is best matched to settings where such positives can be treated as exchangeable; when positives have substantially different relevance, rating-aware groups or weighted variants may be preferable. Alternative methods could be derived from different angles, e.g., expectation-maximization that introduces latent assignment variables over the unobserved candidate orderings, or a variational lower bound with a flexible variational distribution over the positive permutations. Each alternative offers a different trade-off between tractability and tightness, and a systematic exploration of these directions for new DPO-style alignment with multi-candidate preferences is an interesting direction for future study.

E Broad Impact

Although this paper focuses on recommendation, Mult-DPO is applicable to broader tasks in which set-wise preference data arise, i.e., a set of preferred and dispreferred responses with no reliable internal order, including information retrieval and search ranking with multiple relevant documents, open-domain question answering with multiple acceptable answers, and code generation with several correct programs. As with any preference-based alignment method, the trained policy inherits the preference distribution of its training data, so we recommend pairing Mult-DPO with responsible-deployment practices, such as bias audits and human review, before consumer-facing applications.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions stated in the abstract and introduction (MN surrogate, closed-form Mult-DPO objective, upper-bound and tightness theorems, multi-level extension) map directly to Sections 3.1–3.4, and the empirical scope matches Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix D discusses the MN surrogate's exchangeability assumption, alternative surrogates (EM, variational), and the empirical scope (three datasets, Qwen-2.5 backbones up to 7B).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions (disjoint nonempty positive/negative sets, positive item weights, well-defined policy ratios) are stated in each theorem/corollary, and complete proofs are given in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 and Appendix C specify datasets, positive/negative construction, val/test protocol, Qwen-2.5 backbones, AdamW with lr $1e{-}6$, parallelism, and the β sweep, and an anonymous code repository is provided.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code (`https://github.com/yaochenzhu/Mult_DPO`) provides training and evaluation scripts for Mult-DPO, are publicly available as described in Section 4.1.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 and Appendix C report splits, optimizer (AdamW, lr $1e{-}6$), parallelism, KV-cache reuse, and the β-on-validation selection rule, with all DPO-style baselines sharing the same protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Tables report point estimates from single runs since repeating all backbone$\times$dataset$\times$method combinations on H100/B200 is prohibitive; consistency is instead shown across 3 datasets, 4 backbone scales, and the β sweep in Figure 1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C reports 2×H100 (0.5B–1.5B) and 2×B200 (3B–7B) with data-parallel/ZeRO-2, and Section 3.5 gives the per-step complexity.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We use only public datasets and openly-licensed Qwen-2.5 backbones, involve no human subjects, and submit anonymously per the double-blind policy.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The Broad Impact section in the appendix discusses transfer to IR/QA/code generation as positive impact and preference-data inheritance as the main negative pathway, recommending bias audits and human review before deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: No high-misuse-risk pretrained model, image generator, or scraped dataset is released; only a fine-tuning method and recommendation-specific checkpoints derived from public assets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets, backbones, and baselines are cited at first use, and each is used under its respective research-use license.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released asset is the anonymous code repository (`https://anonymous.4open.science/r/mult-dpo-41C7`) with training, data, and evaluation scripts and a README; no new models or datasets are released.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing or human-subjects research; only previously collected public datasets are used.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human-subjects research, so IRB review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: LLMs are the central object of study (the policy π_θ); the specific Qwen-2.5-Instruct backbones, reference policies, and SFT prerequisite are disclosed in Section 4 and Appendix C.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.