Personalizing LLMs with Binary Feedback: A Preference-Corrected Optimization Framework

Xilai Ma[1], **Liye Zhao**[2], **Weijun Yao**[2], **Haibing Di**[2], **Wenya Wang**[3], **Jing Li**[1]✉
[1]Harbin Institute of Technology, Shenzhen, China
[2]Huawei Technologies Co., Ltd. [3]Nanyang Technological University
maxilai.hour1@gmail.com jingli.phd@hotmail.com

Abstract

Large Language Model (LLM) personalization aims to align model behaviors with individual user preferences. Existing methods often focus on isolated user histories, neglecting the essential role of inter-user differences. We propose C-BPO, a framework that personalizes LLMs via preference-calibrated binary signals. By treating target user data as positive feedback and other users' data as an auxiliary set of implicit negative signals, C-BPO captures distinct inter-user differences. To mitigate the preference overlap issue, where shared task knowledge is erroneously penalized, we derive an objective grounded in Positive-Unlabeled (PU) learning theory. This approach purifies negative signals by subtracting "positive bias", ensuring alignment with unique idiosyncrasies without compromising general helpfulness. Empirical experiments across various personalization tasks and backbone LLMs show C-BPO consistently outperforms baselines, demonstrating the efficacy of preference-calibrated binary signals in modeling inter-user differences.

1 Introduction

The advent of Large Language Models (LLMs) has driven the realization of advanced applications such as question answering, planning, and interactive agents (Ouyang et al., 2022; Touvron et al., 2023; Yang et al., 2024). These models not only excel at executing specific tasks guided by textual instructions, but also demonstrate exceptional potential in complex reasoning (Shao et al., 2024; Guo et al., 2026b) and tool-assisted decision-making (Zhang et al., 2025a; Guo et al., 2026a). However, most efforts to enhance LLM helpfulness adhere to a "one-size-fits-all" paradigm, optimizing for the preferences of an average user. Consequently, LLM personalization has emerged as a pivotal research direction (Salemi et al., 2024; Chen et al., 2024b; Tseng et al., 2024; Guan et al., 2025; Li et al., 2025), seeking to align model responses with individual user preferences.

Early prompt-based techniques (Mysore et al., 2024; Richardson et al., 2023) facilitate personalization by appending user-specific context, such as retrieved historical snippets or summarized user profiles, to the input query. Furthermore, parameter-efficient fine-tuning (PEFT) methods have been introduced to learn lightweight, user-specific modules directly from historical data (Tan et al., 2024b; Liu et al., 2025), with further research exploring collaborative generation through multi-LoRA architectures (Tan et al., 2024a; Zhang et al., 2025b). Despite their success, these methods primarily focus on the target user's isolated history. Drawing from behavioral science (Snyder and Fromkin, 2012; Irmak et al., 2010), where individuality is defined by inter-user variability, recent studies emphasize that effective personalization must capture how a user deviates from the general population. For instance, frameworks such as DPL (Qiu et al., 2025b) and DEP (Qiu et al., 2025a) have attempted to model these distinctions through either LLM-based textual comparisons or latent space embeddings of historical data. However, these methods face scalability and granularity constraints: textual analyzers are limited by the high computational cost of LLM reasoning, while latent embeddings often fail to capture fine-grained, phrase-level stylistic nuances.

While preference optimization frameworks like DPO (Rafailov et al., 2023) offer a solution, they require contrastive completions (y_{win}, y_{lose}) for identical inputs x, which are unavailable in individualized user historical data. Furthermore, such frameworks lack a mechanism to exploit inherent inter-user information for refined preference learning. This leads to a fundamental question: *Can we model individual preferences directly from raw, unpaired data by leveraging inter-user distinctions?*

✉ Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26539–26555

Inspired by recent advances (Ethayarajh et al., 2024; Jung et al., 2025) in preference optimization with binary feedback, it has been demonstrated that LLMs can be aligned using binary signals, such as "thumbs-up" or "thumbs-down" labels assigned to individual data points (x, y). This approach circumvents the requirement for shared inputs x and contrastive completions (y_{win}, y_{lose}) in traditional preference datasets. In this work, we argue that the target user's historical data can be treated as positive feedback, while data from other users serves as implicit negative feedback. This formulation allows the model to capture inter-user distinctions directly within a binary-feedback preference optimization framework. However, existing binary-feedback preference optimization (BPO) methods are typically designed for "clean-labeled preference" where negative samples are objectively inferior according to average human preferences. In personalized contexts, naively treating other users' data as purely negative signals incurs a **preference overlap** issue (as detailed in § 3.1). Specifically, this leads to the excessive penalization of **generic task-specific knowledge** and **community-wide preferences** that are inherently shared between the target user and the broader population.

To adapt BPO to personalized scenarios, we derive a preference-calibrated objective grounded in Positive-Unlabeled (PU) learning theory (Bekker and Davis, 2020). By leveraging the core concept of unbiased risk decomposition, we reformulate the optimization objective to account for the fact that data from other users acts as an unlabeled mixture of both common and user-specific features (as detailed in § 3.2). Specifically, we derive an estimator that recovers the true negative risk by subtracting the "positive bias" from the total auxiliary risk, where this bias represents the expected risk of target preferences found within the broader user data. This objective explicitly purifies the negative signals, ensuring that the model focuses on unique individual idiosyncrasies rather than suppressing general helpfulness. Furthermore, to handle the data imbalance between these two categories of data, we introduce an independent exponential moving average (EMA) reference point estimation to maintain a stable preference boundary during training (as detailed in § 3.3).

In summary, we propose C-BPO, a framework designed for personalizing LLMs through preference-calibrated binary-feedback signals. To evaluate its effectiveness, we conduct extensive experiments across five personalized generation tasks using various LLMs. Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to frame LLM personalization as a binary-feedback preference optimization (BPO) problem, utilizing inter-user variability as a natural contrastive signal for individual-level alignment.
- We uncover the **preference overlap** issue in standard BPO for personalization (§ 3.1) and formulate a preference-calibrated objective grounded in Positive-Unlabeled (PU) learning (§ 3.2).
- Extensive experiments demonstrate that C-BPO consistently outperforms strong baselines. We also provide in-depth analyses of how sample volume (§ 5.1) and preference overlap (§ 5.2) influence the optimization dynamics.

2 Preliminaries

To align large language models with general human preferences (e.g., helpfulness and harmlessness), representative approaches such as Reinforcement Learning from Human Feedback (RLHF, Ouyang et al., 2022) and Direct Preference Optimization (DPO, Rafailov et al., 2023) typically rely on paired preference data to explicitly or implicitly construct reward signals.

DPO. Rafailov et al. (2023) shows that the policy π_θ can be directly optimized from the paired preference dataset $\mathcal{D}$, and the implicit reward function can be defined as a function of the policy:

$$r_\theta(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} \tag{1}$$

Combining the BT model with the implicit reward, the loss function of DPO is

$$-\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(r_\theta(x, y_w) - r_\theta(x, y_l)\right)\right].$$

Here, y_w is a preferred completion and y_l is a non-preferred completion.

Preference Optimization with Binary Feedback. To overcome the limitation of RLHF that human feedback is provided as pairwise preferences over multiple outputs for the same input x (e.g., $y_w > y_l$ for the same input x), recent works (Ethayarajh et al., 2024; Jung et al., 2025) have shown that LLMs can be aligned using binary feedback, where "thumbs-up" or "thumbs-down" signals are assigned to individual (x, y) data point–completion

pairs without requiring shared inputs. KTO (Ethayarajh et al., 2024) and BCO (Jung et al., 2025) are two representative methods in this line of work. Both of them optimize a closely related objective under the implicit reward formulation induced by the policy–reference relationship in Eq. (1). We take the BCO objective as an illustrative example[1]:

$$\mathbb{E}_{(x,y_w)\sim\mathcal{D}^+}\left[-\log\sigma(r_\theta(x,y_w)-\delta)\right] + \mathbb{E}_{(x,y_l)\sim\mathcal{D}^-}\left[-\log\sigma(-(r_\theta(x,y_l)-\delta))\right], \tag{2}$$

where $\mathcal{D}^+$ and $\mathcal{D}^-$ denote "thumbs-up" and "thumbs-down" datasets, respectively, and δ is a reference point that anchors the implicit reward.

The main difference between KTO and BCO lies in how they define the reference point δ in Eq. (2). KTO sets δ as the average implicit reward of *other completions* in the same batch, with zero-clipping to ensure non-negativity (as shown in Eq. (12)).

In contrast, BCO improves upon KTO by taking the average implicit reward of contrastive pairs from the training batch as the reference point:

$$\delta_{\text{BCO}} = \frac{1}{2}\Big(\mathbb{E}_{(x,y)\sim\mathcal{D}^+}[r_\theta(x,y)] + \mathbb{E}_{(x,y)\sim\mathcal{D}^-}[r_\theta(x,y)]\Big). \tag{3}$$

This modification reduces bias introduced by batch-based reference points and preserves informative gradients for all samples, enabling more stable training purely from binary feedback.

3 Personalization with Binary-Feedback Preference Optimization

In this section, we introduce C-BPO, a framework designed for personalizing LLMs through preference-calibrated binary-feedback signals (Figure 1). We first formulate the personalization task within a binary-feedback preference optimization framework and analyze the potential biases arising from "negative data" in §3.1. Subsequently, in §3.2, we derive a preference-corrected optimization objective based on Positive-Unlabeled (PU) learning theory to achieve unbiased preference alignment. Finally, in §3.3, we discuss how to extend our framework to scenarios characterized by data imbalance between user-specific and "negative" sets.

[1]More detailed comparison are given in § B

3.1 Problem Formulation and Motivation

The goal of LLM personalization is to adapt a pre-trained model π_{base} to align with the specific preferences and behaviors of a target user, which are characterized by their historical interaction data $\mathcal{H}_{\text{user}} = \{(x_i, y_i)\}_{i=1}^N$. A standard practice for efficient adaptation is to employ Parameter-Efficient Fine-Tuning (PEFT) techniques, such as LoRA (Hu et al., 2021), to optimize a user-specific module Δ_{user}. The personalized model is thus defined as $\pi_{\text{user}} = \pi_{\text{base}} + \Delta_{\text{user}}$. Conventional approaches typically optimize Δ_{user} via supervised fine-tuning (SFT) on $\mathcal{H}_{\text{user}}$, minimizing the cross-entropy loss between the model's output and the historical ground-truth labels. While standard practices rely solely on $\mathcal{H}_{\text{target}}$ for supervised training, we investigate the following question: *can we directly capture inter-user distinctions by modeling preferences within the raw historical data?*

Our key motivation stems from **Binary-Feedback Preference Optimization** (§2), which facilitates preference alignment without requiring explicit pairwise signals. Under this framework, we can treat the **target user's data** $\mathcal{H}_{\textbf{tar}}$ as positive feedback ("thumbs-up") and the data from **auxiliary users** $\mathcal{H}_{\textbf{aux}}$[2] as implicit negative feedback ("thumbs-down"). This allows us to directly capture inter-user difference signals from raw data by employing the Eq. (2).

While existing binary-feedback preference optimization methods typically assume "clean scenarios", where $(x, y_w) \in \mathcal{D}^+$ and $(x, y_l) \in \mathcal{D}^-$ are explicitly labeled as preferred and non-preferred. Naively treating $\mathcal{H}_{\text{aux}}$ as the negative set $\mathcal{D}^-$ introduces a critical challenge: preference overlap. Specifically, different user data inevitably share common preferences (Zhang et al., 2025b), encompassing **generic task-specific knowledge** defined by the core requirements of a particular task, and **community-wide preferences** characterized by collective stylistic or aesthetic trends shared across the population. Directly penalizing $\mathcal{H}_{\text{aux}}$ via Eq. (2) forces the model to erroneously suppress shared features, impeding effective individual-level preference alignment. To mitigate this issue, we aim to "peel off" the common preference signals from the noisy auxiliary data, ensuring a more accurate and robust personalization.

[2]Throughout this paper, we refer to the user undergoing optimization as the target user and a selected pool of other users as auxiliary users.

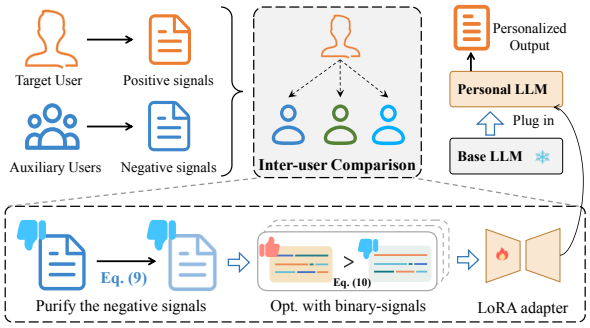


Figure 1: Overview of the **C-BPO** framework. We leverage the target user's data as positive signals and auxiliary users' data as implicit negative signals to align the LLM with the target user's distinct preferences.

3.2 Unbiased Personalization via PU Risk Reformulation

To address the negative bias discussed in §3.1, we propose to reformulate the binary-feedback preference optimization under Positive-Unlabeled (PU) learning theory (Kiryo et al., 2017; Bekker and Davis, 2020; Wang et al., 2023).

Unbiased Risk Estimation. In standard binary classification, let $L(g(x), y)$ be a loss function for a classifier g. The expected risk $R(g)$ is defined as:

$$R(g) = \pi_p R_p^+(g) + \pi_n R_n^-(g), \tag{4}$$

where π_p/π_n is the positive/negative class prior, and $R_p^+(g) = \mathbb{E}_{x\sim p_p}[L(g(x), +1)]$ and $R_n^-(g) = \mathbb{E}_{x\sim p_n}[L(g(x), -1)]$ denote the expected positive and negative risks, respectively.

In standard PU learning, the negative distribution p_n is likewise unavailable; instead, we utilize a set of unlabeled samples drawn from the marginal density $p_u(x) = \pi_p p_p(x) + \pi_n p_n(x)$. This relationship allows for the following risk decomposition:

Lemma 1 (Risk Decomposition)**.** *Let $p_u(x) = \pi_p p_p(x) + \pi_n p_n(x)$ be the marginal distribution of unlabeled data. The negative risk can be unbiasedly estimated using only positive and unlabeled data distributions:*

$$\pi_n R_n^-(g) = R_u^-(g) - \pi_p R_p^-(g). \tag{5}$$

The proof and further details on PU Learning are provided in Appendix C and D, respectively. Eq. (5) suggests that the negative bias (the "misclassified" positive component) inherent in p_u can be explicitly subtracted to recover the true underlying negative signal and ensure a stable estimation.

C-BPO Objective. Under the binary-feedback preference optimization framework, we define g as a discriminator capable of distinguishing between preferred and non-preferred samples, which in practice is parameterized by the user-specific module Δ_{user} to capture preference boundaries. Based on the binary signals introduced in § 2, we formulate the following preference loss functions:

$$l(g(x, y), +1) = -\log \sigma(r_\theta(x, y) - \delta), \tag{6}$$
$$l(g(x, y), -1) = -\log \sigma(-(r_\theta(x, y) - \delta)), \tag{7}$$

where $r_\theta(x, y)$ is derived from Eq. (1). Here, $l(g, +1)$ quantifies the degree to which a sample is preferred, whereas $l(g, -1)$ measures the extent to which it is non-preferred.

Following Eq. (2), the objective of standard binary-feedback preference optimization (with $\mathcal{D}^+$ and $\mathcal{D}^-$) can be viewed as the minimization of the empirical risk summed over both positive and negative samples:

$$R^+(g) + R^-(g) = \mathbb{E}_{(x,y_w)\sim\mathcal{D}^+}[l(g(x, y_w), +1)] + \mathbb{E}_{(x,y_l)\sim\mathcal{D}^-}[l(g(x, y_l), -1)]. \tag{8}$$

In the context of personalization, we treat the data from auxiliary users $\mathcal{H}_{\text{aux}}$ as an unlabeled set. By substituting the negative risk estimator derived in Eq. (5) into the objective, we obtain the raw C-BPO formulation:

$$\mathcal{L}_{\text{raw}} = \mathbb{E}_{\mathcal{H}_{\text{tar}}}[l(g, +1)] + \frac{1}{\pi_n}\Big(\mathbb{E}_{\mathcal{H}_{\text{aux}}}[l(g, -1)] - \pi_p \mathbb{E}_{\mathcal{H}_{\text{tar}}}[l(g, -1)]\Big). \tag{9}$$

This expression enables the model to recover an unbiased preference signal by correcting the negative gradient with information from the positive set during the optimization process.

Assumption Adaptation. Applying Eq. (9) requires re-examining the foundational assumptions of PU learning (Hyttinen et al., 2013). Conventionally, the unlabeled set is required to be a mixture of positive and negative distributions, $p_u(x) = \pi_p p_p(x) + (1 - \pi_p) p_n(x)$. While $\mathcal{H}_{\text{tar}}$ and $\mathcal{H}_{\text{aux}}$ are physically disjoint, we posit that this mixture assumption remains applicable within the shared preference manifold. Specifically, we treat $\mathcal{H}_{\text{aux}}$ as an implicit mixture where shared features (e.g., generic task knowledge) overlap with the target user's preference space. This allows us to apply the risk decomposition in Eq. (5) as an estimator

for personalized calibration. A formal discussion on the adaptation of the assumption is provided in Appendix E.

Under this relaxation, the class prior π_p is reinterpreted as an empirical correction coefficient $\alpha \in (0, 1)$. This coefficient α[3] quantifies the degree of preference feature overlap between the target user and auxiliary users.

Furthermore, as observed by Kiryo et al. (2017), if the model is highly flexible, the empirical estimator for $R^{-}(g)$ may become negative during training, leading to overfitting. To ensure stability and robustness, we adopt the non-negative constraint (Kiryo et al., 2017). The final C-BPO optimization objective is formulated as:

$$\mathcal{L}_{\text{C-BPO}} = \mathcal{L}_{\text{pos}} + \frac{1}{1-\alpha} \max\left\{0, \mathcal{L}_{\text{pure\_neg}}\right\} \quad (10)$$

where

$$\mathcal{L}_{\text{pos}} = \underbrace{\mathbb{E}_{\mathcal{H}_{\text{tar}}}[l(g, +1)]}_{\text{Positive Alignment}},$$

$$\mathcal{L}_{\text{pure\_neg}} = \underbrace{\mathbb{E}_{\mathcal{H}_{\text{aux}}}[l(g, -1)] - \alpha \mathbb{E}_{\mathcal{H}_{\text{tar}}}[l(g, -1)]}_{\text{Purified Negative Loss}}.$$

where the first term ensures the model aligns with the target user's historical preferences, and the second term (Purified Negative Loss) provides a debiased negative signal by explicitly correcting the contamination from common preference features.

3.3 Adaptation to Imbalanced Preference Data

In personalized scenarios, the auxiliary user data $\mathcal{H}_{\text{aux}}$ typically far exceeds the target user's data $\mathcal{H}_{\text{tar}}$ in volume. While leveraging a larger $\mathcal{H}_{\text{aux}}$ may facilitate the modeling of inter-user information, it poses a challenge for the stability of the reference point δ (§ 2).

In standard BCO (Jung et al., 2025), δ is estimated by averaging rewards across positive and negative samples jointly within each batch. While effective in balanced settings, this joint estimator is highly sensitive to sampling ratios; an increased proportion of negative samples can cause δ to disproportionately drift toward the dominant distribution, thereby undermining its role as a neutral baseline for accurate preference discrimination.

To mitigate this, we decouple the tracking of reward statistics. Specifically, we maintain **independent** Exponential Moving Averages (EMA) for positive and auxiliary rewards. The calibrated reference point, δ_{EMA}, is then dynamically computed as the mean of these two decoupled statistics during each optimization step. This ensures a stable decision boundary that remains invariant to batch-level data imbalance.

[3]We provide an empirical rationale (§ E.2) and the corresponding estimation procedure (§ E.3) for α.

4 Experiments

4.1 Experimental Setup

Datasets and Backbone Models. We evaluate the effectiveness of C-BPO on personalized generation tasks from the LaMP (Salemi et al., 2024) and LongLaMP (Kumar et al., 2024) benchmarks. Specifically, we select the following representative tasks: News Headline Generation (LaMP-4), Scholarly Title Generation (LaMP-5), Abstract Generation (LongLaMP-2), Review Writing (LongLaMP-3), and Topic Writing (LongLaMP-4). These datasets provide per-user behavioral history, query inputs, and ground-truth outputs. Following prior work (Tan et al., 2024b; Bu et al., 2025), we report ROUGE-1 and ROUGE-L scores for all tasks. To ensure the robustness of our evaluation, we conduct experiments across several backbone series, including LLaMA (Touvron et al., 2023; Dubey et al., 2024), Qwen (Yang et al., 2024), and Mistral (Jiang et al., 2023).

Baselines and Implementation Details. We compare C-BPO against several categories of baselines: (1) **Retrieval-based methods**: We include **RAG** (Lewis et al., 2020), which retrieves user-related histories and appends them to the prompt as in-context evidence, and **PAG** (Richardson et al., 2023), which incorporates a summarized user profile directly into the prompt to guide the generation process. (2) **SFT-based adapters**: We consider **TAM** (Tan et al., 2024b), a task-specific LoRA module trained on general task data excluding the target user's data, representing a non-personalized upper bound for general task performance. We also compare with **OPPU** (Tan et al., 2024b), which fine-tunes individual adapters exclusively on each user's historical interaction data via standard supervised learning. (3) **Preference-based methods**: We compare against **CoPE** (Bu et al., 2025), which constructs negative samples for each user-specific instance through rejection sampling and optimizes user-specific adapters using DPO (Rafailov et al., 2023). Additionally, we evaluate **KTO** (Ethayarajh et al., 2024) and **BCO** (Jung et al., 2025), which

LaMP Bench. → Method ↓	Abstract Gen.		Review Writing		Topic Writing		News Headline		Scholarly Title	
	R-1	R-L	R-1	R-L	R-1	R-L	R-1	R-L	R-1	R-L
Non-Tuned										
Base Model	0.341	0.186	0.287	0.126	0.246	0.105	0.119	0.105	0.409	0.324
RAG	0.347	0.205	0.272	0.128	0.243	0.115	0.141	0.124	0.425	0.347
PAG	0.344	0.186	0.256	0.125	0.262	0.107	0.118	0.102	0.372	0.289
SFT-based										
TAM	0.357	0.204	0.289	0.122	0.253	0.107	0.200	0.179	0.514	0.456
OPPU	0.378	0.218	0.319	0.134	0.278	0.112	0.203	0.182	0.510	0.454
Preference Opt.										
CoPE ♣	0.392	0.239	0.335	0.146	0.281	**0.120**	0.205	0.184	0.519	0.461
KTO †	0.370	0.229	0.298	0.126	0.269	0.109	0.191	0.173	0.491	0.431
BCO †	0.373	0.231	0.315	0.132	0.272	0.112	0.197	0.179	0.507	0.443
C-BPO (Ours)†	**0.398**	**0.269**	**0.353**	**0.154**	**0.291**	0.118	**0.215**	**0.198**	**0.539**	**0.481**

Table 1: Results on the LaMP benchmark. R-1 and R-L denote ROUGE-1 and ROUGE-L, respectively. **Bold** and underline mark the best and second-best results. "†" denotes the series of binary-feedback preference optimization methods. ♣ indicates the methods implemented based on DPO.

enable preference optimization using binary feedback, allowing for flexible alignment even when the prompts for positive and negative instances differ.

For binary-feedback preference optimization methods (KTO, BCO, and C-BPO), unless otherwise specified, we randomly sample data from other users as the auxiliary set during user-specific training, maintaining a balanced 1:1 ratio with the positive (user-specific) data. For C-BPO, the calibration coefficient α is pre-estimated prior to training, following the procedure detailed in Appendix E.3. To ensure a fair comparison with previous studies, user-specific training is initiated from the TAM checkpoint, and we employ greedy decoding for all evaluations. Hyperparameters for LoRA modules are kept consistent across all training-based methods. More details on the experimental setup can be found in Appendix F.

4.2 Main Results

We present the primary experimental results in Table 1 (using *Mistral-7B-Instruct-v0.3*), with extensive evaluations across different base models provided in Appendix G.1 (Figure 4). From these results, we derive several key observations:

First, directly applying binary-feedback preference optimization (BPO) methods, such as KTO and BCO, to personalization tasks does not guarantee performance gains. In several cases, these methods even underperform SFT-based baselines. This phenomenon suggests that **negative signals from other users can exert a detrimental effect if treated naively**. As analyzed in §3.1, this is likely due to the preference overlap inherent in personalization scenarios, where generic task-specific knowledge and community-wide preferences are mistakenly penalized.

In contrast, **C-BPO effectively exploits the negative signals from auxiliary users**, consistently outperforming both SFT-based and standard BPO-based approaches. This demonstrates that our method successfully leverages inter-user distinctions while avoiding the excessive penalization of shared preferences. Furthermore, C-BPO achieves competitive or even superior performance compared to the DPO-based method, CoPE. These results underscore the critical importance of incorporating inter-user information to refine individual preference modeling in LLM personalization.

Furthermore, results presented in Appendix G.1 (Figure 4) demonstrate that the superior performance of C-BPO generalizes across various backbone LLMs, confirming its robustness and model-agnostic effectiveness.

5 Analysis

In this section, we analyze the impact of key components in C-BPO, including the quantity and quality of auxiliary data, the sensitivity of the correction coefficient α, and the effectiveness of the EMA-based reference point estimation.

Figure 2: Performance comparison across varying proportions of user historical data ($\mathcal{H}_{\text{tar}}$). x denotes the ratio of auxiliary data to target user data.

5.1 The Role of Auxiliary Data

To further investigate how auxiliary data influences the optimization of personalized LLMs[4], we evaluate the performance of C-BPO under varying proportions of user historical data and different reference point estimation strategies, as illustrated in Figure 2. Specifically, we examine the behavior of C-BPO by introducing two variables: First, we adjust the data ratio x (where $x = \#\mathcal{H}_{\text{aux}}/\#\mathcal{H}_{\text{tar}}$) to 0.5 and 1.5, to observe the model's sensitivity to the volume of negative signals. Second, we compare the standard reference point estimation with our proposed independent EMA-based strategy (§ 3.3). A more detailed experimental setup is elaborated in § F.3.

Trade-off in Auxiliary Data Volume. We find that increasing the volume of auxiliary data does not always yield monotonic performance gains. As shown in Figure 2, when the target user history $\mathcal{H}_{\text{tar}}$ is scarce (e.g., $< 50\%$), a higher data ratio ($x = 1.5$, *purple dashed line*) yields inferior results compared to the balanced setting ($x = 1.0$, *orange solid line*). Superiority for $x = 1.5$ is only achieved after the volume of $\mathcal{H}_{\text{tar}}$ surpasses a certain threshold. This suggests that effectively distilling inter-user information from $\mathcal{H}_{\text{aux}}$ requires a sufficient amount of $\mathcal{H}_{\text{tar}}$ to provide a strong enough "positive" anchor to guide the debiasing process.

Robustness of EMA-based Reference Point Estimation. The independent EMA-based estimation proves crucial for handling imbalanced data scenarios. In the absence of EMA (*green dashed line*), C-BPO exhibits a noticeable performance drop as x deviates from 1.0. Notably, when $x = 1.5$ and $\mathcal{H}_{\text{tar}}$ is limited, the model without EMA even underperforms the SFT-based baseline (OPPU), indicating that a biased reference point can hinder the model's ability to utilize negative signals. In contrast, the EMA-based strategy (*purple dashed line*) effectively mitigates the sensitivity to x by maintaining a stable decision boundary. We also observe that with sufficient $\mathcal{H}_{\text{tar}}$, the model becomes more resilient to estimation methods, eventually capturing inter-user signals even without EMA. This further reinforces that the proposed EMA mechanism is a vital stabilizer for personalized alignment in data-constrained or imbalanced regimes.

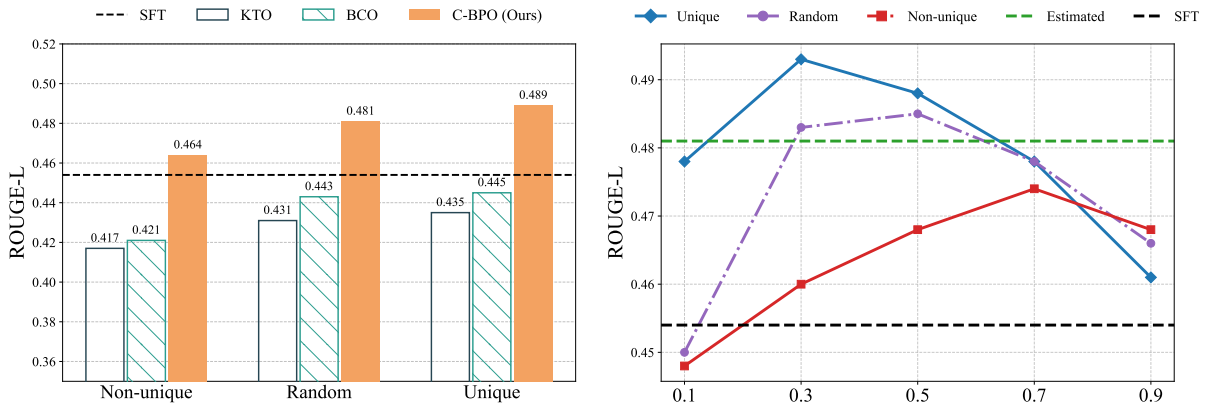
5.2 The Impact of User Uniqueness

This section further examines how the degree of user uniqueness (**preference overlap**) influences personalization performance. Following prior research (Qiu et al., 2025b,a; Liu et al., 2025), which suggests that the embeddings of users' history effectively reflect user characteristics, we construct different experimental groups by retrieving specific sets of auxiliary data. Using Euclidean distance in the embedding space relative to $\mathcal{H}_{\text{tar}}$, we curate three distinct configurations: a *Unique* group (retrieving users with the largest distances), a *Non-unique* group (retrieving those with the smallest distances), and a *Random* group. A more detailed experimental setup is elaborated in § F.4. We then evaluate C-BPO and its counterparts across these controlled settings, as illustrated in Figure 3[5].

Analysis across User Groups. As illustrated in Figure 3 (a), standard BPO-based methods (KTO and BCO) exhibit a significant performance degradation when transitioning from the *Random* to the *Non-unique* setting. Furthermore, they fail to achieve substantial gains even in the *Unique* setting, consistently underperforming the SFT baseline. This suggests that traditional BPO approaches are unable to navigate the dense preference overlap in non-unique scenarios and fail to effectively extract useful auxiliary signals when preferences are distinct. In contrast, C-BPO consistently facilitates preference optimization across all settings by leveraging inter-user information, with the most pronounced improvements observed in the *Unique* group, where individual idiosyncrasies are more prominent and distinct.

[4] In § G.3, we examine how our calibration prevents the erosion of overlap preference during personalization.

[5] More results on different datasets are given in § G.2



(a) User Uniqueness Analysis **(b) Sensitivity Analysis on α**

Figure 3: Analysis of user uniqueness and the sensitivity of α. (a) Performance across groups retrieved via embedding distances (*Unique*, *Random*, *Non-unique*). (b) C-BPO performance under varying α, with "Estimated" denoting the performance under the estimation method in § E.3.

Sensitivity and Estimation of the Correction Coefficient α. Intuitively, the correction coefficient α should scale with the degree of preference overlap. Figure 3 (b) confirms this intuition: the *Non-unique* group requires a higher α to filter out common preferences, whereas the optimal performance for the *Unique* group is achieved at a lower α value. These findings validate the necessity and effectiveness of our preference-calibrated objective. To address the challenge of hyperparameter tuning, we provid an estimation method based on the embedding of user history in § E.3. As shown by the "Estimated" markers in Figure 3 (b) (for the *Random* group), this heuristic method provides a reliable starting point that closely aligns with the empirically optimal coefficient.

6 Related Work

Existing LLM personalization methods primarily fall into three lines: retrieval-based prompting, fine-tuning with user history, and training with auxiliary data. Retrieval-based approaches integrate user-specific context into the input via few-shot demonstrations (Brown et al., 2020), relevant history snippets (RAG, Mysore et al., 2024; Salemi et al., 2024), or summarized profiles (PAG, Richardson et al., 2023; Guan et al., 2025). While scalable and interpretable, these methods often result in shallow personalization as they rely heavily on the inherent capacity of a powerful LLM to analyze user information within prompt-length constraints. Fine-tuning-based methods, in contrast, adapt model parameters through Parameter-Efficient Fine-Tuning (PEFT) for individuals (Tan et al., 2024b,a; Liu et al., 2025) or specific user groups (Zhang et al., 2025b). While standard approaches rely solely on historical user data to learn corresponding adapters or embedding prefixes, CoPE (Bu et al., 2025) further constructs negative samples via rejection sampling to train user-specific adapters using DPO (Rafailov et al., 2023), which significantly complicates the overall training and inference pipeline.

Recent advancements have further explored incorporating other users' information as auxiliary signals, such as using inter-user differences (DPL, Qiu et al., 2025b) or collaborative filtering-inspired retrieval (CFRAG, Shi et al., 2025) to highlight individual uniqueness. Along this line, DEP (Qiu et al., 2025a) models user difference information within the latent space based on the embedding of the user history data. Yet, these methods either demand complex reasoning from a large LLM or are confined to latent embedding spaces that may miss fine-grained linguistic nuances. Moreover, their heavy reliance on sufficient data for high-quality latent representation extraction hinders their ability to generalize or rapidly extend to new users.

In contrast to the aforementioned methods, our framework migrates binary-feedback preference optimization into the personalization context, relying solely on the target and auxiliary other users' data to construct contrastive signals. By shifting the focus from high-resource LLM reasoning or holistic embedding alignment to explicit sample-level comparisons, our method effectively captures term- and phrase-level relative information while streamlining the personalization pipeline, enabling

more precise alignment with genuine user intent without the necessity for massive datasets or ultra-powerful backbone models.

7 Conclusion

This paper introduces C-BPO, a framework designed for personalizing LLMs through preference-calibrated binary-feedback signals. Unlike existing methods that rely on isolated user histories, C-BPO captures individuality by treating target user data as positive signals and auxiliary user data as implicit negative feedback. To address the fundamental challenge of preference overlap, we derive a calibrated objective grounded in Positive-Unlabeled (PU) learning, which purifies negative signals from the auxiliary data. Comprehensive experiments across five personalized generation tasks demonstrate that C-BPO consistently outperforms competitive baselines, offering a robust, theoretically sound, and scalable solution for aligning LLMs with nuanced individual preferences.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (62476070), Shenzhen Science and Technology Program (JCYJ20241202123503005, GXWD20231128103232001, ZDSYS20230626091203008, KQTD20240729102154066), Department of Science and Technology of Guangdong (2024A1515011540), National Key R&D Program of China (SQ2024YFE0200592) and Suzhou Science and Technology Program (SYG2025072).

Limitations

While C-BPO demonstrates consistent improvements across personalized generation tasks, it possesses several limitations. First, the framework's effectiveness is sensitive to the calibration of the correction coefficient α. Although we have proposed an initial estimation strategy in Appendix E.3, more sophisticated and dynamic methods for adaptive parameter estimation remain to be explored. Second, our method requires access to a centralized auxiliary dataset comprising other users' raw historical data to construct the unlabeled set. It poses privacy risks and deployment hurdles in specific real-world scenarios, such as Federated Learning or on-device personalization, where accessing raw user data is prohibited. Lastly, our current evaluation is centered on generative benchmarks. Adapting this preference-calibrated objective to broader personalization domains, such as dialogue scenarios or recommendation systems, requires further study to fully validate its cross-domain generalizability.

Ethical Considerations

While our work aims to personalize large language models (LLMs), we acknowledge the potential ethical concerns. The datasets used in our study may contain biases, which could be reflected in the model's outputs. Mitigating such biases is crucial for ensuring fairness. Additionally, the use of LLMs may generate offensive or biased content. We suggest that practitioners should carefully examine the potential bias before deploying the model in real-world applications.

References

Jessa Bekker and Jesse Davis. 2020. Learning from positive and unlabeled data: A survey. *Machine learning*, 109(4):719–760.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems (NeurIPS)*, 33:1877–1901.

Hyungjune Bu, ChanJoo Jung, Minjae Kang, and Jaehyung Kim. 2025. Personalized LLM decoding via contrasting personal preference. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 33946–33966.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, and 1 others. 2024b. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web*, 27(4):42.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 213–220.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning (ICML)*.

Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, and Wei Wu. 2025. A survey on personalized Alignment—The missing piece for large language models in real-world applications. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5313–5333. Association for Computational Linguistics.

Weiyang Guo, Zesheng Shi, Liye Zhao, Jiayuan Ma, Zeen Zhu, Junxian He, Min Zhang, and Jing Li. 2026a. E3-tir: Enhanced experience exploitation for tool-integrated reasoning. *arXiv preprint arXiv:2604.09455*.

Weiyang Guo, Zesheng Shi, Zeen Zhu, Yuan Zhou, Min Zhang, and Jing Li. 2026b. Backdoors in rlvr: Jailbreak backdoors in llms from verifiable reward. *arXiv preprint arXiv:2604.09748*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. 2013. Experiment selection for causal discovery. *The Journal of Machine Learning Research*, 14(1):3041–3071.

Caglar Irmak, Beth Vallen, and Sankar Sen. 2010. You like what i like, but i don't like what you like: Uniqueness motivations in product preferences. *Journal of Consumer Research*, 37(3):443–455.

Shantanu Jain, Justin Delano, Himanshu Sharma, and Predrag Radivojac. 2020. Class prior estimation with biased positives and unlabeled examples. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, volume 34, pages 4255–4263.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. 2025. Binary classifier optimization for large language model alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1858–1872.

Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis, and Masashi Sugiyama. 2017. Positive-unlabeled learning with non-negative risk estimator. In *Advances in neural information processing systems (NeurIPS)*, volume 30.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, and 1 others. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in neural information processing systems (NeurIPS)*, volume 33, pages 9459–9474.

Zhuo Li, Guodong Du, Weiyang Guo, Yigeng Zhou, Xiucheng Li, Wenya Wang, Fangming Liu, Yequan Wang, Deheng Ye, Min Zhang, and 1 others. 2025. Multi-objective large language model alignment with hierarchical experts. *arXiv preprint arXiv:2505.20925*.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025. Llms+ persona-plug= personalized llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9373–9385.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 198–219.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems (NeurIPS)*, volume 35, pages 27730–27744.

Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. 2025a. Latent inter-user difference modeling for llm personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10610–10628.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. 2025b. Measuring what makes you unique: Difference-aware user modeling for enhancing LLM

personalization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21258–21277.

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

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7370–7392.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Han Li. 2025. Retrieval augmented generation with collaborative filtering for personalized text generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1294–1304.

Charles R Snyder and Howard L Fromkin. 2012. *Uniqueness: The human pursuit of difference*.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024a. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6459–6475.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024b. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6476–6491.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631.

Xutao Wang, Hanting Chen, Tianyu Guo, and Yunhe Wang. 2023. Pue: Biased positive-unlabeled learning enhancement by causal inference. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:19783–19798.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, and 1 others. 2025a. The landscape of agentic reinforcement learning for llms: A survey. *arXiv preprint arXiv:2509.02547*.

Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. 2025b. Proper: A progressive learning framework for personalized large language models with group-level adaptation. In *The 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Appendix

A The Use of AI Assistants

Throughout the preparation of this manuscript, large language models were employed exclusively for light stylistic refinement and grammatical adjustment. Furthermore, these tools assisted in generating structural templates for the illustrative diagrams.

B The Detailed Difference between BCO and KTO

KTO Ethayarajh et al. (2024) proposed alignment framework that trains on binary signal of thumbs-up or thumbs-down collected on a per-sample basis for every unique prompt and completion combination. Given a dataset of { prompt, completion } pairs with respective binary signals, KTO defines a value function

$$v_{KTO}(x, y; \theta) = \begin{cases} \sigma(r_\theta(x, y) - z_{\text{ref}}) & \text{if } y \sim y_{\text{desirable}} \mid x \\ \sigma(z_{\text{ref}} - r_\theta(x, y)) & \text{if } y \sim y_{\text{undesirable}} \mid x, \end{cases} \tag{11}$$

where z_{ref} is a reference point. In practice, z_{ref} is implemented as

$$z_{\text{ref}} = \max\left(0, \frac{1}{|\mathcal{B}|} \sum_{y' \in \mathcal{B} \setminus y} \log \frac{\pi_\theta(y'|x)}{\pi_{\text{ref}}(y'|x)}\right) \tag{12}$$

for $(x, y) \in \mathcal{B}$ and $\mathcal{B} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{B}$ is a batch of samples.

Finally, the loss function of KTO is defined as

$$\mathcal{L}_{\text{KTO}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[w(y)(1 - v_{\text{KTO}}(x, y; \theta)\right] \tag{13}$$

where the weighting factor $w(y)$ is λ_D if y is a completion from thumbs-up dataset and λ_U if y is a completion from thumbs-down dataset.

BCO Jung et al. (2025) proposed Binary Classifier Optimization (BCO), a theoretically grounded alignment framework that learns directly from binary feedback. BCO views alignment from binary signals as training a binary classifier whose logit is the implicit reward induced by the policy–reference log-ratio,

$$r_\theta(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}. \tag{14}$$

Under this formulation, minimizing the binary cross-entropy (BCE) loss over thumbs-up and thumbs-down samples provably upper bounds the DPO objective, enabling alignment without explicit preference pairs.

To tighten this bound, BCO introduces a reward-shifting term δ defined as the average implicit reward of positive and negative samples,

$$\delta_{\text{BCO}} = \frac{1}{2}\Big(\mathbb{E}_{(x,y)\sim\mathcal{D}^+}[r_\theta(x, y)] + \mathbb{E}_{(x,y)\sim\mathcal{D}^-}[r_\theta(x, y)]\Big). \tag{15}$$

where $\mathcal{D}^+$ and $\mathcal{D}^-$ denote thumbs-up and thumbs-down datasets, respectively. The resulting BCO objective is

$$\mathbb{E}_{(x,y_w)\sim\mathcal{D}^+}\left[-\log\sigma(r_\theta(x, y_w) - \delta)\right] + \mathbb{E}_{(x,y_l)\sim\mathcal{D}^-}\left[-\log\sigma(-(r_\theta(x, y_l) - \delta))\right], \tag{16}$$

which preserves informative gradients across samples and enables stable, effective alignment purely from binary feedback.

C Proofs for Risk Decomposition

In this section, we provide the detailed derivation of the risk decomposition. For completeness, we first restate Lemma 1 from the main text before proceeding to its formal proof.

Lemma 2 (Risk Decomposition, Restated)**.** *Let $p_u(x) = \pi_p p_p(x) + \pi_n p_n(x)$ be the marginal distribution of unlabeled data. The negative risk $\pi_n R_n^-(g)$ can be unbiasedly estimated using only positive and unlabeled data distributions as:*

$$\pi_n R_n^-(g) = R_u^-(g) - \pi_p R_p^-(g). \tag{17}$$

By the definition of the expected risk on unlabeled data $\mathcal{D}_U$, which follows the marginal density

$p(x)$, we have:

$$
\begin{aligned}
R_u^-(g) &= \int L(g(x), -1)p(x)dx \\
&= \int L(g(x), -1)\left[\pi_p p_p(x) + \pi_n p_n(x)\right] dx \\
&= \pi_p \int L(g(x), -1)p_p(x)dx \\
&\quad + \pi_n \int L(g(x), -1)p_n(x)dx \\
&= \pi_p R_p^-(g) + \pi_n R_n^-(g). \qquad (18)
\end{aligned}
$$

In Eq. (18), the first term $\pi_p R_p^-(g)$ represents the risk contribution from positive samples that are mistakenly treated as negative in the unlabeled set. Rearranging the terms, we obtain the estimator for the true negative risk: $\pi_n R_n^-(g) = R_u^-(g) - \pi_p R_p^-(g)$.

D Review of PU Learning

In this section, we provide a more formal background on Positive-Unlabeled (PU) learning (Kiryo et al., 2017; Bekker and Davis, 2020; Wang et al., 2023) and the derivation of the unbiased risk estimator used in our framework.

From PN to PU Risk In an ideal scenario where both positive and negative preference data are available, a model g parametrized by θ is optimized by minimizing the expected risk:

$$R(g) = \pi R_p^+(g) + (1-\pi)R_n^-(g), \qquad (19)$$

where $\pi = P(y = 1)$ is the class prior, while $R_p^+(g) = \mathbb{E}_{x\sim p_p}[L(g(x), +1)]$ and $R_n^-(g) = \mathbb{E}_{x\sim p_n}[L(g(x), -1)]$ denote the expected positive and negative risks. In the PU learning setting, a representative negative set $\mathcal{D}_N$ is unavailable. Instead, we possess a positive set $\mathcal{D}_P$ and an unlabeled set $\mathcal{D}_U$ drawn from the marginal density $p(x) = \pi p_p(x) + (1-\pi)p_n(x)$.

Unbiased Risk Estimator The key challenge in PU learning is to estimate the negative risk $R_n^-(g)$ without negative samples. Utilizing the marginal density relationship, we can express the expected negative loss on the unlabeled distribution as a mixture:

$$R_u^-(g) = \pi R_p^-(g) + (1-\pi)R_n^-(g), \qquad (20)$$

where $R_u^-(g) = \mathbb{E}_{x\sim p(x)}[L(g(x), -1)]$ and $R_p^-(g) = \mathbb{E}_{x\sim p_p}[L(g(x), -1)]$. By rearranging this identity, we obtain an unbiased estimator for the weighted negative risk:

$$(1-\pi)R_n^-(g) = R_u^-(g) - \pi R_p^-(g). \qquad (21)$$

Substituting this back into Eq. (19), the total risk for PU learning becomes:

$$R_{PU}(g) = \pi R_p^+(g) + R_u^-(g) - \pi R_p^-(g). \qquad (22)$$

E Empirical Adaptation of SCAR in Preference Spaces

E.1 Assumption Adaptation

A potential concern in applying Positive-Unlabeled (PU) learning to personalization is the validity of the **Selected Completely At Random (SCAR)** assumption. In its classic form, SCAR posits that labeled positive instances are sampled randomly from the underlying positive distribution, implying that the unlabeled set $\mathcal{D}_U$ serves as a representative mixture of both positive and negative classes. While this assumption is traditionally applied to classification tasks with well-defined class separability, it requires a nuanced reinterpretation within the context of high-dimensional, personalized preference alignment.

In our personalization context, although $\mathcal{H}_{\text{aux}}$ (historical data from auxiliary users) does not physically encompass the target user's specific historical data, we argue that the preference distributions across different users are fundamentally entangled in the latent feature space.

Preference Feature-Space Decomposition. For the sake of simplicity, we assume that the representation of a preference sample x in the latent manifold can be decomposed into two orthogonal components: a *general* component x_{gen} (representing *generic task-specific knowledge* and *community-wide preferences*) and a *specific* component x_{spec} (representing *individual-specific preference*). Formally, we define the feature mappings as:

$$\text{Feature}(U) \approx w_{u,1} \cdot \Phi_{\text{gen}} + w_{u,2} \cdot \Phi_{\text{spec\_neg}}, \qquad (23)$$

$$\text{Feature}(P) \approx w_{p,1} \cdot \Phi_{\text{gen}} + w_{p,2} \cdot \Phi_{\text{spec\_pos}}. \qquad (24)$$

The Mechanism of Bias and Correction. The bias arises because $\mathcal{H}_{\text{aux}}$ and $\mathcal{H}_{\text{tar}}$ overlap significantly in the Φ_{gen} subspace. If we were to naively penalize all samples in $\mathcal{H}_{\text{aux}}$ (i.e., minimizing $\mathbb{E}_{\mathcal{H}_{\text{aux}}}[l^-]$), the model would inevitably receive

negative gradients for the general features Φ_{gen}. This leads to the undesirable effect of the model becoming less helpful or coherent simply because it overfits to the target user's specific taste at the expense of general capabilities.

By reinterpreting the class prior π as a correction coefficient α, our objective $\mathbb{E}_{\mathcal{H}_{\text{aux}}}[l^-] - \alpha\mathbb{E}_{\mathcal{H}_{\text{tar}}}[l^-]$ functions as a gradient counter-balancing mechanism:

- $\mathbb{E}_{\mathcal{H}_{\text{aux}}}[l^-]$ generates a negative gradient on $\Phi_{\text{gen}} + \Phi_{\text{spec\_neg}}$.
- $-\alpha\mathbb{E}_{\mathcal{H}_{\text{tar}}}[l^-]$ generates a *positive* gradient on $\Phi_{\text{gen}} + \Phi_{\text{spec\_pos}}$.

Since $\mathcal{H}_{\text{tar}}$ and $\mathcal{H}_{\text{aux}}$ are "entangled" within the Φ_{gen} manifold, the positive gradient from the term $-\alpha\mathbb{E}_{\mathcal{H}_{\text{tar}}}[l^-]$ precisely offsets the erroneous negative gradient produced by $\mathcal{H}_{\text{aux}}$ on the shared features.

Conclusion on SCAR Adaptation. Under this interpretation, the SCAR assumption holds not on the physical identity of the samples, but on the **shared feature manifold**. The target user's positive samples $P \in \mathcal{H}_{\text{tar}}$ can be viewed as being "randomly selected" from the broader distribution of "high-quality/preferred features" available in the unlabeled pool $\mathcal{H}_{\text{aux}}$. **The coefficient α thus reflects the density of these shared features Φ_{gen} relative to the total auxiliary set**, providing a theoretically grounded way to calibrate the debiasing strength without compromising the model's fundamental performance. We also provide a heuristic way to estimate the correction coefficient α in Appendix E.3.

E.2 Intuitive Understanding of the Correction Coefficient α

To provide a straightforward interpretation of α, we establish a conceptual mapping between the class prior π_p in PU learning and our correction coefficient. In binary classification, π_p represents the mixing proportion of positive instances within the unlabeled set. In our personalized generation context, this generalizes to the density of preference overlap: α quantifies the extent to which the auxiliary user data $\mathcal{H}_{\text{aux}}$ contains features that are intrinsically aligned with the target user's preference manifold. While a high π_p suggests a label-contaminated unlabeled set, a high α indicates that the target user shares significant commonalities (e.g., generic task-specific knowledge and community-wide preferences) with the broader population, necessitating stronger calibration to avoid suppressing these shared positive traits.

E.3 Estimation of the Correction Coefficient α

The correction coefficient α in our framework shares a fundamental conceptual link with the class prior π_p in Positive-Unlabeled (PU) learning (§ 3.2). Specifically, α quantifies the degree of preference overlap, representing the proportion of auxiliary data that functionally serves as positive signal relative to the target user. Inspired by the seminal Elkan-Noto method (Elkan and Noto, 2008; Jain et al., 2020), which estimates the class prior by evaluating the labeling propensity of unlabeled instances, we design a prior estimation scheme for α based on latent representations.

Given that historical embeddings effectively capture distinct user characteristics (Liu et al., 2025; Qiu et al., 2025a), we formalize the estimation as follows:

1. **Training the Proxy Classifier:** We train a probabilistic classifier $g(x)$ to distinguish between the target user history $\mathcal{H}_{\text{tar}}$ (proxy class 1) and the auxiliary user history set $\mathcal{H}_{\text{aux}}$ (proxy class 0) in the embedding space. This classifier learns to identify features indicative of the target user's specific preference manifold.

2. **Estimation of Propensity Score (c):** Following the SCAR assumption that labeled target instances are representative of the underlying positive distribution, we use a held-out validation set $h_t \subset \mathcal{H}_{\text{tar}}$ to estimate the constant labeling propensity c:
$$\hat{c} = \frac{1}{|h_t|} \sum_{x \in h_t} g(x). \tag{25}$$

3. **Calculation of α:** With the estimate $\hat{c}$ and the trained classifier $g(x)$, the correction coefficient α is derived by averaging the predictions over the auxiliary set $\mathcal{H}_{\text{aux}}$, representing the density of "target-like" preferences within the broader population:
$$\hat{\alpha} = \frac{1}{|\mathcal{H}_{\text{aux}}| \cdot \hat{c}} \sum_{x \in \mathcal{H}_{\text{aux}}} g(x). \tag{26}$$

This estimation provides a theoretically grounded heuristic for α before training, enabling the framework to adapt to varying degrees of user uniqueness as empirically validated in § 5.2.

Task	Other Users		Target Users	
	#Profile	Output Length	#Profile	Output Length
Abstract Generation	31,808	233.1 ± 117.5	1,296.7 ± 446.4	210.5 ± 92.8
Review Writing	19,649	407.2 ± 299.5	759.3 ± 324.2	511.8 ± 294.2
Topic Writing	21,119	358.3 ± 316.9	260.6 ± 314.0	358.3 ± 255.4
News Headline Generation	7,275	15.5 ± 6.0	270.1 ± 182.1	18.6 ± 5.2
Scholarly Title Generation	16,076	17.9 ± 6.1	444.0 ± 121.6	16.4 ± 5.8

Table 2: The statistics of the used dataset.

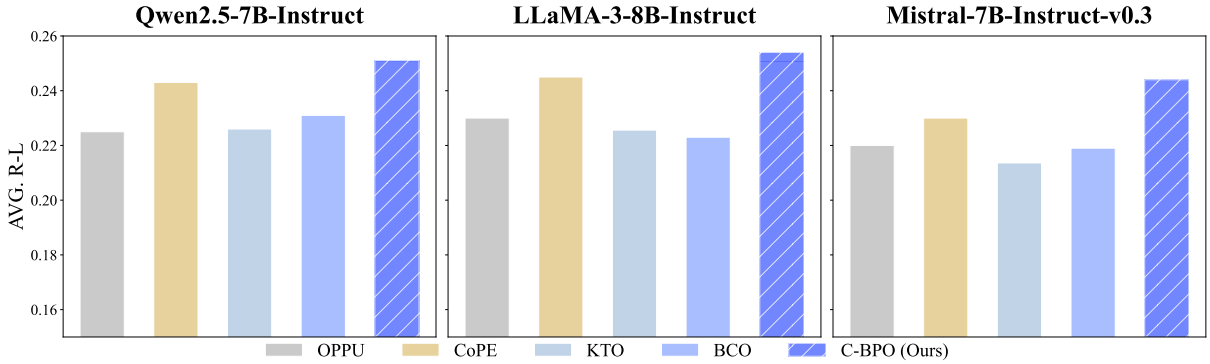


Figure 4: Average performance across 5 tasks for various LLMs.

F Details of Experimental Setup

F.1 Dataset

We evaluate our method on several text generation tasks from the **LaMP** (Salemi et al., 2024) and **LongLaMP** (Kumar et al., 2024) benchmarks, which are designed to test the personalization capabilities of LLMs across diverse contexts.

Following the established protocols from previous work (Tan et al., 2024b; Bu et al., 2025), we select the following representative tasks:

LaMP-4: News Headline Generation. This task requires the model to generate concise headlines for news articles. It emphasizes capturing an author's distinctive journalistic style based on their historical article-title pairs.

LaMP-5: Scholarly Title Generation. Models must generate titles for scholarly abstracts. Success depends on reflecting the author's specific academic writing style as evidenced in their publication history.

LongLaMP-2: Abstract Generation. This task evaluates the model's proficiency in generating scientific abstracts from titles and keywords. It requires emulating characteristic academic phrasing and domain-specific terminology from a user's previous publications.

LongLaMP-3: Review Writing. The objective is to generate product reviews based on specifications. The model must reflect the user's evaluative style and subjective perspective captured in their review history.

LongLaMP-4: Topic Writing. This task involves generating Reddit post content from provided summaries. The model must maintain the unique, often informal, writing style characteristic of individual users across their historical posts.

The detailed statistics of the datasets are provided in Table 2.

F.2 Implementation Details

We provide the additional implementation details for C-BPO and all baseline methods below:

- **General Training Configuration:** In alignment with prior work (Bu et al., 2025), all methods involving a Supervised Fine-Tuning (SFT) stage are optimized using AdamW (Loshchilov and Hutter, 2019) with a weight decay of 0.01. We employ a linear

learning rate scheduler with a warm-up ratio of 0.1. SFT is conducted for 2 epochs, with learning rates set to 1×10^{-4} for LongLaMP and 1×10^{-5} for LaMP. For LoRA-based adaptation, we configure the rank $r = 8$ and scaling factor $\alpha = 16$. All experiments are executed on NVIDIA L40S GPUs.

- **Configuration for Binary-Feedback Preference Optimization:** For methods based on binary-feedback preference optimization (i.e., KTO, BCO, and C-BPO), we utilize data from auxiliary users as the negative set during user-specific training. Unless otherwise specified, we maintain a 1:1 ratio between the positive (target user) and negative (auxiliary users) samples, where the auxiliary data are randomly drawn from the histories of other users. These methods are trained for 3 epochs with a unified learning rate of 1×10^{-6}. The hyperparameter α in C-BPO is pre-estimated prior to training, following the procedure detailed in Appendix E.3.

F.3 Experimental Settings for Auxiliary Data Analysis

To further investigate how auxiliary data and reference point estimation strategies influence the optimization of personalized LLMs, we conduct a series of controlled experiments. We specifically select the top 40 users with the most extensive historical data from the LaMP benchmark's test user set to ensure sufficient data for scaling analysis. During training, we systematically vary the proportion of utilized historical data (controlled by percentage) to observe model performance across diverse data-density regimes.

Two primary dimensions are explored in this analysis:

- **Sensitivity to Auxiliary Data Volume:** We adjust the data ratio x (defined as $x = |\mathcal{H}_{aux}|/|\mathcal{H}_{tar}|$) to values of 0.5 and 1.5. By dynamically controlling the proportion of training data, we observe the model's sensitivity to the volume of negative signals relative to target user data.

- **Impact of Reference Point Calibration:** Across varying data ratios x, we evaluate the effectiveness of our proposed independent EMA-based strategy (§ 3.3). This comparison highlights the framework's robustness under different degrees of data imbalance.

The corresponding experimental results are illustrated in Figure 2.

F.4 Experimental Settings for User Uniqueness Analysis

To investigate the impact of **preference overlap** (§ 3.1) on personalized training, we follow prior research (Qiu et al., 2025a; Liu et al., 2025) which suggests that embedding-based representations of user history effectively reflect individual characteristics. We construct distinct experimental groups by retrieving auxiliary data based on their semantic proximity to the target user. Specifically, we utilize **BGE-M3** (Chen et al., 2024a) as the embedding model to map each historical data point into a high-dimensional vector space. Using the **Euclidean distance** between the embeddings of the target user's history $\mathcal{H}_{tar}$ and the historical data of other users, we retrieve an equal volume of samples to form $\mathcal{H}_{aux}$. Based on this distance metric, we define three comparative groups:

- ***Non-unique group*:** Comprised of auxiliary data with the smallest Euclidean distance to $\mathcal{H}_{tar}$, representing a high degree of preference overlap.

- ***Unique group*:** Comprised of auxiliary data with the largest Euclidean distance, representing significant user divergence.

- ***Random group*:** Formed by randomly sampling auxiliary data to serve as a baseline reference.

The comparative results across these groups are illustrated in Figure 3 and Figure 5.

G Additional Experimental Results

G.1 Additional Results across Different Backbone LLMs

To evaluate the generalizability of our approach, we perform experiments on multiple backbone LLMs. The results in Figure 4 demonstrate that the superiority of C-BPO is consistent across different architectures, validating its robustness and model-agnostic nature.

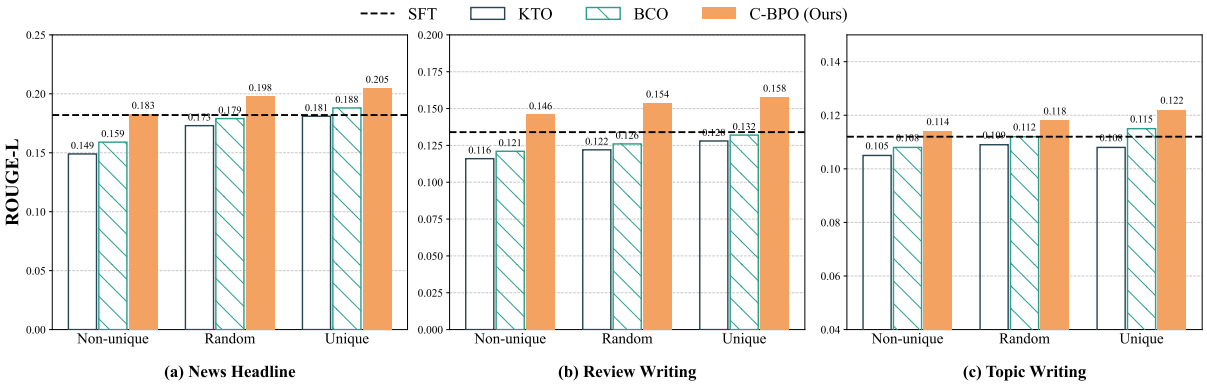


Figure 5: User uniqueness analysis across different tasks.

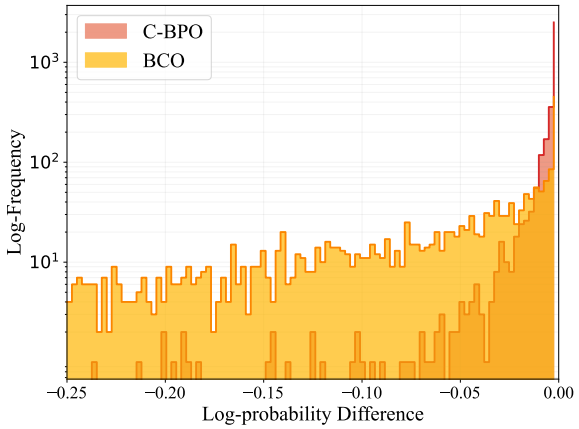


Figure 6: Token-level log-probability shift on auxiliary data. C-BPO effectively mitigates the over-penalization of shared preferences compared to the BCO.

G.2 Additional Results for User Uniqueness

We further extend our analysis by evaluating user grouping across different datasets, following the experimental setup detailed in § 5.2, with results presented in Figure 5. The empirical findings consistently align with the primary observations discussed in § 5.2.

G.3 Analysis on Preservation of Overlap Preference

To empirically verify that C-BPO effectively mitigates the excessive penalization of auxiliary data, we analyze the shift in token-level log-probabilities before and after personalization.

Experimental Setup. We focus on a "Non-Unique User" group identified via the embedding-based clustering strategy described in § F.4. Within this group, users exhibit high behavioral similarity. We first fine-tune a base LLM on the entire group to obtain a general preference model, π_{gen}. Subsequently, we randomly select a target user $\mathcal{H}_{\text{tar}}$ and treat the remaining users as the auxiliary set $\mathcal{H}_{\text{aux}}$. We then perform personalized training using both the standard BPO (the baseline) and our proposed C-BPO, resulting in two personalized models: π_{BPO} and $\pi_{\text{C-BPO}}$.

Evaluation Metric. We assess the distribution shift by measuring the token-level log-probability difference on the auxiliary data $\mathcal{H}_{\text{aux}}$:

$$\Delta \log P = \log \pi_{\text{pers}}(y|x) - \log \pi_{\text{gen}}(y|x), \quad (27)$$

where $\pi_{\text{pers}} \in \{\pi_{\text{BPO}}, \pi_{\text{C-BPO}}\}$. This metric quantifies the extent to which personalization alters the model's confidence in "overlap preference" knowledge shared between the target and auxiliary users.

Observations. As illustrated in Figure 6, the standard BCO objective leads to a significant decline in the log-probabilities of auxiliary tokens. Given the high similarity within the user group, this decline indicates that the standard objective erroneously treats shared linguistic patterns and common preferences as negative signals, suppressing them during optimization. In contrast, C-BPO maintains a higher and more stable probability distribution on $\mathcal{H}_{\text{aux}}$. This result confirms that our preference-calibrated objective effectively "protects" shared knowledge, ensuring that the model differentiates the user only on truly idiosyncratic traits without compromising the foundational commonalities of the user community.