BEHAVIOR-AWARE REWARD SHAPING FOR PERSONALIZED PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Language Models (LLMs) increasingly power web applications such as chatbots and recommender systems, making personalization critical for enhancing user experience. Existing methods for personalized preference alignment often extend Reinforcement Learning from Human Feedback (RLHF) by incorporating user characteristics into reward shaping. While effective, these approaches mainly rely on semantic information and overlook behavioral similarity, which encodes collaborative relationships among users and responses—important cues for personalization. Yet, capturing such signals is challenging: at the response level, users interact with dynamically generated responses, causing cross-user interactions to appear orthogonal with no overlaps, preventing the capture of collaborative signals. To address this challenge, we propose **BehaviorRM**, a reward shaping framework that explicitly models behavioral similarity. First, Response Semantic Decomposition represents each query-response pair as a composition of shared semantic identifiers (SIDs) obtained via residual quantization, which induces interaction overlaps at the SID level and exposes latent collaborative structures. Next, Behavioral Interaction Modeling fits the SID-level interaction matrix by jointly learning embeddings for users and SIDs, capturing behavioral similarity beyond pure semantics. Finally, Behavior-aware Reward Shaping integrates the behavioral rewards with semantic rewards to form a unified signal, enhancing personalized preference modeling. Extensive experiments on three real-world datasets show that BehaviorRM significantly outperforms baseline methods. All codes are included in the Supplemental Material.

1 INTRODUCTION

The widespread success of Large Language Models (LLMs) Guo et al. (2025a); Achiam et al. (2023) has sparked strong interest in leveraging them to transform a variety of web applications, such as web chatbots Dam et al. (2024) and recommender systems Zhang et al. (2024a); Lin et al. (2025); Zhang et al. (2026), or to develop new types of web applications, such as generative search[1]. Unlike general-purpose LLMs, which are designed to handle broad tasks in a "*one-size-fits-all*" manner Qiu et al. (2025b), web applications must serve a diverse set of users, each with distinct preferences and needs Ko et al. (2025). Effectively applying LLMs in these scenarios requires aligning the models with user preferences to provide more personalized services—a crucial factor for enhancing user satisfaction and enabling broad adoption. Consequently, *personalized preference alignment* Chen et al. (2025a); Zhao et al. (2025a); Poddar et al. (2024) has emerged as an important research direction, attracting substantial attention from both academia Poddar et al. (2024); Zhao et al. (2025a); Shenfeld et al. (2025); Zollo et al. (2025) and industry Bose et al. (2025).

Preference alignment for LLMs is commonly approached via the paradigm in Reinforcement Learning from Human Feedback (RLHF) Ouyang et al. (2022), which encodes human preferences into optimizable reward signals for LLM alignment. Building upon this framework, personalized preference alignment can be achieved by extending reward shaping to incorporate individual user characteristics. Prior work can be grouped into two categories: (1) representation-learning methods, which learn user representations and integrate them into reward shaping Poddar et al. (2024); Chen et al. (2025a); and (2) factorization-based methods, which decompose rewards into different bases and learn user-specific

[1] https://blog.google/products/search/generative-ai-google-search-may-2024/

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

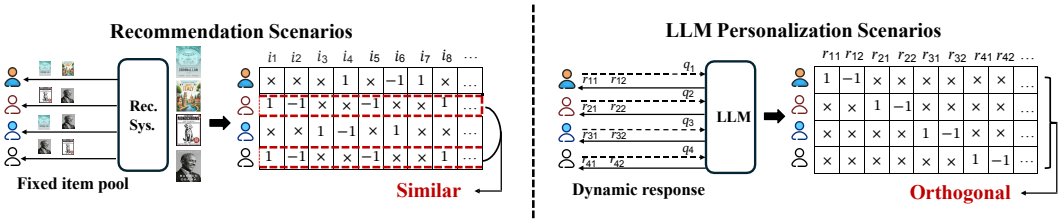

Figure 1: User interactions in (1) recommendation scenarios with a fixed item pool, and (2) LLM personalization scenarios with dynamically generated responses.

combination weights to capture personalized preferences Bose et al. (2025); Shenfeld et al. (2025). Both approaches demonstrate effective personalized reward shaping, representing notable progress toward aligning LLMs with individual users.

Despite considerable progress, we argue that existing methods share a common limitation: they lack explicit modeling of behavioral similarity Sun & Liu (2025). In these approaches, regardless of how users are represented, reward shaping is typically based on language models and primarily focuses on extracting semantic information to match user characteristics for reward prediction. However, behavioral signals themselves carry highly informative cues for user modeling, as widely demonstrated in the recommendation domain He et al. (2017); Liao et al. (2024). For instance, a user may prefer responses favored by behaviorally similar users, even if their semantics differ from the user's historical responses; likewise, different responses liked by the same user may be related, even if their semantics are distinct. Modeling solely from a semantic perspective makes it difficult to capture these behavior similarity signals (also termed *collaborative signals* Zhang et al. (2025b)) and inevitably limits the effectiveness of personalized modeling.

This work investigates incorporating explicit behavioral similarity modeling into reward shaping for personalized preference alignment. A common approach for capturing the similarity is to use latent factors to fit an interaction matrix representing user behaviors. For example, in recommender systems, collaborative filtering He et al. (2017) uses user and item embeddings to fit the user–item interaction matrix, capturing collaborative signals of "similar users–similar items" relationships. In personalized preference alignment, however, users do not interact with fixed items; instead, they engage with responses dynamically generated by LLMs, where overlaps on the same response are extremely rare. As shown in Figure 1, this sparsity often produces a nearly orthogonal user–response interaction matrix, making it difficult to capture collaborative relationships between users or between responses. Consequently, directly fitting such a matrix is infeasible for modeling behavioral similarity, limiting the ability of reward shaping to leverage these informative behavioral signals.

To address these challenges, we first represent each response and its corresponding query as a composition of semantic identifiers (SIDs), obtained via quantization techniques such as Residual-Quantized Variational AutoEncoder (RQ-VAE) Lee et al. (2022). Each SID is drawn from a dedicated codebook and can be shared across responses, allowing coarse-grained interaction overlaps to emerge and revealing collaborative relationships hidden at the raw response level. We then fit the SID-level interaction matrix by jointly learning embeddings for users and SIDs, capturing behavioral similarity beyond semantics. Finally, the resulting behavioral rewards are integrated with semantic rewards to form a unified signal for more effective personalized reward modeling. These designs collectively form our framework, termed **BehaviorRM**.

The main contributions of this work are as follows:

- We introduce a new perspective by explicitly modeling behavioral similarity in reward shaping for enhancing personalized preference alignment in LLMs.
- We present BehaviorRM, a behavior-aware reward shaping method that leverages quantization to construct semantic IDs, enabling the capture of behavior similarity signals.
- We integrate BehaviorRM with the state-of-the-art reward shaping method and evaluate it on three real-world datasets, widely demonstrating its effectiveness in enhancing personalized preference modeling for LLMs.

2 Related Work

• **LLM Personalization.** We defer detailed discussion to Appendix C. LLM personalization aims to move beyond one-size-fits-all alignment toward adapting models to individual user preferences and values Chen et al. (2024); Zhang et al. (2024b); Liu et al. (2025b). Existing work can be broadly divided into personalized generation and personalized preference alignment.

Personalized generation. Existing methods either adapt models in a training-free manner through prompting and retrieval Salemi et al. (2024b); Qiu et al. (2025b); Kim et al. (2025); Shi et al. (2025); Salemi et al. (2024a), or optimize model parameters to capture user-specific characteristics Tan et al. (2024); Qiu et al. (2025a); Zhao et al. (2026); Wang et al. (2026). While effective for personalized content generation, these methods mainly focus on generation quality rather than preference alignment.

Personalized preference alignment. Personalized preference alignment focuses on aligning LLMs with individual human preferences by modeling personalized reward functions Jang et al. (2023); Chen et al. (2025b); Shen et al. (2025). Representative methods include prototype- or latent-profile-based modeling Chen et al. (2025a); Poddar et al. (2024), low-dimensional reward representations Bose et al. (2025); Shenfeld et al. (2025), and synthetic persona construction Ryan et al. (2025). However, existing approaches mainly rely on semantic-level modeling and largely overlook behavioral similarity signals across users. In contrast, our work explicitly models collaborative behavioral patterns to improve personalized reward modeling.

• **Behavior Similarity Modeling.** Behavior similarity modeling has been extensively studied in recommender systems, where collaborative filtering methods capture user preference patterns from co-occurrence signals He et al. (2023). Traditional approaches range from heuristic CF methods Sarwar et al. (2001) to latent factor models Koren et al. (2009); Kumar et al. (2014) and neural recommendation models He et al. (2017); Kang & McAuley (2018); He et al. (2020). However, these methods assume that items are shared across users, which is often not the case in personalized LLM settings, where query-response pairs are highly sparse and unique.

Recent semantic-ID-based generative recommendation methods Wang et al. (2024); Zhou et al. (2025); Li et al. (2026); Rajput et al. (2023) replace raw item IDs with semantic representations to enable autoregressive modeling. Inspired by this idea, we map query-response pairs into semantic IDs to recover co-occurrence structures across users, enabling effective behavioral similarity modeling for personalized LLM alignment.

3 Preliminaries

In this section, we briefly introduce the background of preference alignment for LLMs.

• **Task Formulation.** Preference alignment for LLMs aims to ensure that model-generated responses are consistent with human preferences Ouyang et al. (2022), such that preferred responses receive higher scores than less preferred ones under the same query. In practice, human feedback is typically collected through relative comparisons between candidate responses, which are generally more reliable than absolute quality judgments.

Let x denote a query and y denote a response to x. Preference annotations are provided in the form of triplets (x, y^+, y^-), where y^+ denotes the *preferred* response and y^- denotes the *non-preferred* response for the same query x. Under this formulation, preference alignment can be naturally cast as a reward modeling problem, which aims to learn a scoring function $r(x, y)$ that assigns higher scores to preferred responses, *i.e.*, $r(x, y^+) > r(x, y^-)$.

• **Bradley-Terry (BT) Model.** BT model Bradley & Terry (1952) provides a standard probabilistic framework for modeling pairwise preference annotations. Under this model, the probability that a preferred response y^+ is chosen over a non-preferred response y^- for the same query x is given by:

$$p(y^+ \succ y^- \mid x) = \sigma\big(r(x, y^+) - r(x, y^-)\big), \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The reward model parameters are learned by maximizing the log-likelihood of the observed preferences. This approach encourages the model to assign higher

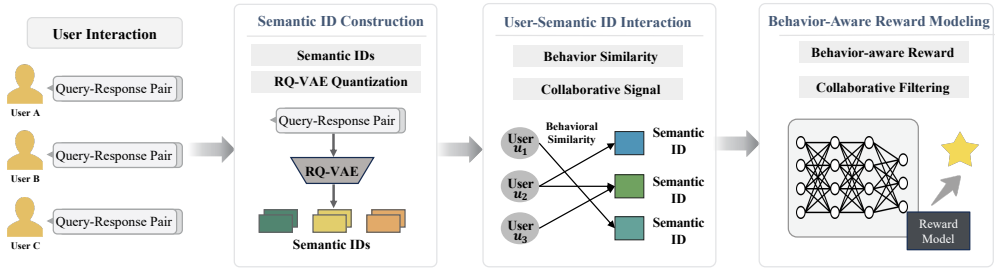

Figure 2: Overview of the BehaviorRM framework. BehaviorRM treats each user–(query, response) pair as a behavioral interaction and performs reward modeling in three stages: (1) Semantic ID construction, which maps query–response pairs into discrete semantic identifiers; (2) User–Semantic ID interaction modeling, which builds an interaction matrix capturing user–semantic relationships; and (3) Behavior-aware reward modeling, which integrates collaborative filtering with text-semantics–aware modeling to produce final reward estimates.

scores to preferred responses while providing a smooth probabilistic treatment of noisy human feedback, serving as a principled foundation for preference alignment.

However, the BT model assumes a global reward function shared across all users, which tends to average over diverse signals and overlooks user-specific distinctions. This assumption is particularly limiting in personalized settings, where preferences vary due to individual characteristics, prior experiences, or cultural backgrounds Zhang et al. (2024b); Xu et al. (2025b); Liu et al. (2025b); Xie et al. (2025a). To overcome this limitation, recent work has focused on *personalized preference alignment*, which replaces the global reward function with a user-conditioned formulation:

$$r(x, y) \longrightarrow r(u, x, y), \tag{2}$$

where u denotes the user. This formulation enables the model to more accurately capture heterogeneity in user preferences.

4 Methodology

In this section, we first provide an overview of the proposed methodology, followed by a detailed introduction of its components.

4.1 Overview

We aim to model personalized rewards by capturing both semantic consistency for each individual and behavioral similarity across users, thereby enabling more effective personalized preference modeling. Our key insight is that collaborative patterns are inherently latent within user interaction data and can be revealed through appropriate representations. Based on this, we propose **BehaviorRM**, a unified framework that comprises the following components (as shown in Figure 2):

- *Response Semantic Decomposition*. To mitigate the sparsity of overlaps among dynamically generated responses, each response with its corresponding quey is represented as a composition of shared semantic identifiers (Semantic IDs, SIDs) via quantization. This operation induces SID-level interaction overlaps and exposes latent collaborative structures, enabling the model to capture shared semantic components across users.
- *Behavioral Interaction Modeling*. The SID-level interaction matrix is learned by jointly embedding users and SIDs, allowing the model to capture behavioral similarity between users based on their interactions with semantic components.
- *Behavior-aware Reward Shaping*. Behavioral rewards derived from the interaction modeling are combined with semantic rewards to form a unified reward signal, reflecting both individual semantic preferences and cross-user behavioral similarities, thereby enhancing personalized preference alignment.

4.2 Response Semantic Decomposition

In BehaviorRM, we model a user interaction as a triple (u, x, y), where the query–response pair (x, y), rather than the query alone, is treated as the analogue of an item in recommender systems. This design captures the open-ended and query-dependent nature of generated responses. Moreover, by embedding interactions into a shared latent semantic space, it mitigates the severe sparsity induced by exact (x, y) matching and enables effective collaboration among semantically similar query–response pairs.

We obtain the representation by encoding the concatenated (x, y) with a backbone language model and applying last-token pooling:

$$\boldsymbol{z} = \text{Pooling}\big(\text{LLM}((x, y))\big),$$

where $\boldsymbol{z} \in \mathbb{R}^d$ denotes the latent representation of the query–response pair, $\text{LLM}(\cdot)$ denotes the backbone model, and $\text{Pooling}(\cdot)$ corresponds to the last-token pooling operation.

We then learn a mapping that projects each query–response pair embedding into a sequence of discrete semantic identifiers (SIDs):

$$f(\boldsymbol{z}) = [c_1, c_2, \ldots, c_L],$$

where $f(\cdot)$ denotes the mapping function, c_l represents the semantic identifier at level l, and L is the length of the SID sequence. To implement this mapping, we adopt Residual-Quantized Variational AutoEncoders (RQ-VAE) Lee et al. (2022), which are widely used for representation quantization. Given a semantic embedding $\boldsymbol{z}$ of a query–response pair (x, y), an MLP-based encoder first projects it into a latent representation:

$$\boldsymbol{h} = \text{Encoder}(\boldsymbol{z}). \tag{3}$$

The latent vector $\boldsymbol{h}$ is then discretized into L hierarchical SIDs through residual vector quantization with L independent codebooks. Specifically, at level l, the codebook $\mathcal{C}_l = \{\boldsymbol{e}_l^k\}_{k=1}^K$ contains K codeword embeddings. The SID assignment and residual update are defined as:

$$c_l = \arg\min_k \|\boldsymbol{v}_l - \boldsymbol{e}_l^k\|^2, \tag{4}$$

$$\boldsymbol{v}_l = \boldsymbol{v}_{l-1} - \boldsymbol{e}_{l-1}^{c_{l-1}}, \tag{5}$$

where $\boldsymbol{v}_1 = \boldsymbol{h}$ and $\boldsymbol{v}_l$ denotes the residual representation at level l. After quantization across all levels, the discrete SID sequence $[c_1, \ldots, c_L]$ induces a reconstructed latent representation by aggregating the selected codewords:

$$\tilde{\boldsymbol{h}} = \sum_{l=1}^{L} \boldsymbol{e}_l^{c_l}. \tag{6}$$

An MLP-based decoder then maps $\tilde{\boldsymbol{h}}$ back to the semantic space:

$$\tilde{\boldsymbol{z}} = \text{Decoder}(\tilde{\boldsymbol{h}}). \tag{7}$$

The model is trained by minimizing a combination of reconstruction and quantization losses:

$$\mathcal{L}_{\text{SID}} = \underbrace{\|\tilde{\boldsymbol{z}} - \boldsymbol{z}\|^2}_{\text{reconstruction}} + \underbrace{\sum_{l=1}^{L} \Big(\|\text{sg}(\boldsymbol{v}_l) - \boldsymbol{e}_l^{c_l}\|^2 + \beta \|\boldsymbol{v}_l - \text{sg}(\boldsymbol{e}_l^{c_l})\|^2 \Big)}_{\text{quantization}}, \tag{8}$$

where $\text{sg}(\cdot)$ is the stop-gradient operator and β balances the encoder updates and codebook adaptation. The reconstruction term preserves semantic fidelity, while the quantization terms encourage stable and consistent discretization across hierarchical levels. After training, the SID construction is performed offline, and the resulting SID sequences are stored for subsequent model training.

4.3 Behavioral Interaction Modeling

Based on the obtained data in the $(u, [c_1, \ldots, c_L])$ format, we construct an SID-level interaction matrix to alleviate extreme sparsity. This representation is analogous to interaction data in recommendation systems, allowing us to adopt similar modeling techniques. For simplicity, we employ a matrix

factorization (MF) model Koren et al. (2009). Specifically, for each user and each SID, we learn embeddings and measure their compatibility using cosine similarity, which can be formally expressed as:

$$r_{\text{behavior}} = \frac{\boldsymbol{b}_u \cdot \boldsymbol{b}_s}{\|\boldsymbol{b}_u\|\|\boldsymbol{b}_s\|}, \quad \boldsymbol{b}_s = \sum_{l=1}^{L} \boldsymbol{b}_{c_l} \tag{9}$$

where $\boldsymbol{b}_u \in \mathbb{R}^{d_b}$ denotes the embedding of user u, $\boldsymbol{b}_{c_l} \in \mathbb{R}^{d_b}$ is the embedding of the l-th SID token, and $\boldsymbol{b}_s \in \mathbb{R}^{d_b}$ represents the aggregated embedding of the entire SID sequence, and d_b is the behavioral embedding dimension. The cosine similarity r_{behavior} quantifies the behavioral alignment between the user and the corresponding query–response pair in the latent SID space.

4.4 Behavior-aware Reward Shaping

Beyond capturing collaborative patterns at the behavior level, it is also crucial to model semantic relevance between users and textual content. To this end, we additionally learn a text-aware embedding for each user, which enables effective matching with the previously obtained query–response pair embeddings. However, the representations $\boldsymbol{z}$ produced by LLMs are typically high-dimensional, and directly learning user embeddings in this space can lead to representation collapse and optimization instability Guo et al. (2023).

To address this issue, we introduce a linear projection matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d_t}$ that maps the original LLM representations into a lower-dimensional semantic space. This projection facilitates stable optimization while preserving essential semantic information. We then measure the compatibility between a user and a query–response pair using the dot product:

$$r_{\text{text}} = \boldsymbol{t}_u \cdot (\boldsymbol{z}\boldsymbol{A}), \tag{10}$$

where $\boldsymbol{t}_u \in \mathbb{R}^{d_t}$ denotes the text-aware embedding of user u, and d_t is the textual embedding dimension. Following prior work Bose et al. (2025), we constrain $\boldsymbol{t}_u$ to lie on the probability simplex by applying a softmax operation, ensuring that its elements are non-negative and sum to one. This constraint further stabilizes training and improves interpretability.

Overall Reward. We combine the semantic relevance reward and the behavior-aware collaborative reward using a simple linear weighting scheme. Specifically, the overall reward is defined as:

$$r(u, x, y) = \alpha\, r_{\text{text}}(u, x, y) + (1 - \alpha)\, r_{\text{behavior}}(u, x, y), \tag{11}$$

where r_{text} captures the semantic consistency between the user and the query–response pair, r_{behavior} denotes the behavior-aware collaborative reward, and $\alpha \in [0, 1]$ is a trade-off hyperparameter that balances the contribution of the two components. Following the BT model Bradley & Terry (1952), we optimize the reward using a pairwise loss, which can be expressed as:

$$\mathcal{L}_{\text{Reward}} = -\sum_{(u,x,y^+,y^-)} \log \sigma\Big(r(u, x, y^+) - r(u, x, y^-)\Big), \tag{12}$$

where (y^+, y^-) represents a positive and negative response pair for user u given query x, and $\sigma(\cdot)$ is the sigmoid function.

5 Experiments

5.1 Experimental Setup

Datasets. To validate the effectiveness of our proposed BehaviorRM, we conduct experiments on several representative personalized reward datasets, including *ChatbotArena* Zheng et al. (2023), *PRISM* Kirk et al. (2024), *RedditTLDR* Stiennon et al. (2020), and *PersonalLLM* Zollo et al. (2025). Following previous work on personalized rewards Bose et al. (2025); Ryan et al. (2025), we first filter out users with fewer than 5 historical records. Then we partition the interaction history for each user into training, validation, and test sets according to an approximate ratio of 3:1:1; for *PersonalLLM*, we adopt a 9:1:1 ratio instead, following the settings of LoRe. Details of these datasets are provided in Appendix D.

Table 1: Performance comparison of baseline methods and BehaviorRM across datasets and backbone models. "Avg." denotes the average performance across all backbone models for each dataset. Accuracy scores are reported as percentages (%). Higher values indicate better results, where the best-performing methods in each comparison are highlighted in bold.

Datasets	Models	Baselines					Ours
		Ref	BT	PAL	VPL	LoRe	BehaviorRM
ChatbotArena	Qwen3-1.7B	57.64	64.42	49.35	73.54	73.66	**74.17**
	Qwen3-4B	40.21	66.48	48.67	73.75	73.54	**74.88**
	Llama3.2-1B	41.64	64.00	55.11	72.99	73.08	**73.92**
	Llama3.2-3B	72.29	65.73	66.55	72.11	72.28	**73.45**
	Avg.	52.95	65.16	54.92	73.10	73.14	**74.11**
PRISM	Qwen3-1.7B	47.77	49.15	48.09	55.09	**55.30**	54.56
	Qwen3-4B	45.97	50.00	51.59	53.81	**54.87**	54.24
	Llama3.2-1B	47.55	50.21	47.99	52.86	53.07	**54.24**
	Llama3.2-3B	54.46	51.49	50.64	54.77	53.81	**54.87**
	Avg.	48.94	50.21	49.58	54.13	54.26	**54.48**
RedditTLDR	Qwen3-1.7B	47.05	63.31	48.69	69.31	69.26	**69.80**
	Qwen3-4B	46.35	67.63	48.56	72.32	71.79	**72.79**
	Llama3.2-1B	45.20	63.82	48.95	68.08	68.66	**69.59**
	Llama3.2-3B	69.14	67.86	42.47	72.16	71.72	**72.95**
	Avg.	51.94	65.66	47.17	70.47	70.36	**71.28**
PersonalLLM	Qwen3-1.7B	74.67	80.46	80.62	88.34	89.76	**90.00**
	Qwen3-4B	48.20	78.94	80.10	87.46	87.96	**89.44**
	Llama3.2-1B	34.29	80.24	81.14	87.62	89.20	**89.86**
	Llama3.2-3B	74.63	79.24	85.36	86.86	88.50	**89.28**
	Avg.	57.95	79.72	81.81	87.57	88.86	**89.65**

Evaluation Metrics. Following the common evaluation standards Bose et al. (2025), we employ *Pairwise Accuracy* as the primary evaluation metric. This metric measures the proportion of response pairs for which the reward model correctly assigns a higher scalar reward to the user-preferred response than to the rejected one.

Baselines. We compare our proposed BehaviorRM with the following baseline methods: Ref Liu et al. (2025a), BT Bradley & Terry (1952), PAL Chen et al. (2025a), VPL Poddar et al. (2024), LoRe Bose et al. (2025). Details of these baselines are provided in Appendix E.

Implementation Details. To comprehensively evaluate the generalizability of BehaviorRM, we adopt four pre-trained reward models from the Skywork-Reward-V2 series[2] Liu et al. (2025a), including two architectures and varying scales: Qwen3 (1.7B and 4B) and Llama3.2 (1B and 3B). More implementation details are provided in Appendix F.

5.2 Main Results

We begin by evaluating the overall performance of all methods. The main experimental results across datasets and models are presented in Table 1, from which we draw the following observations:

- **Standard (non-personalized) reward modeling approaches are insufficient for personalized preference alignment.** The Ref baseline exhibits poor performance across the majority of datasets, revealing the inability of zero-shot, general-purpose reward models to align with the specific preference distributions of individual users. While BT baseline improves accuracy through task-

[2] https://huggingface.co/collections/Skywork/skywork-reward-v2

Table 2: Ablation study on semantic modeling and behavioral similarity modeling. *"Behavior"* represents behavior-only modeling methods, *"Semantic"* denotes semantic-only methods, and *"Fusion"* combines behavioral and semantic signals.

Category	Methods	Chatbot Arena	PRISM	Reddit TLDR	Personal LLM
Behavior	**Dot-Product**	55.66	45.02	59.01	63.64
	Cosine Similarity	61.72	46.93	64.01	87.32
Semantic	**Semantic Only**	73.24	52.44	**69.97**	88.40
Fusion	**Feature Concatenation**	74.08	**54.56**	68.63	87.70
	Multiplicative Interaction	73.96	53.07	69.79	88.44
	Joint Projection	73.92	54.45	69.53	88.48
	BehaviorRM	**74.17**	**54.56**	69.80	**90.00**

specific adaptation, its monolithic architecture assumes a single shared reward function, failing to capture individual preference conflicts.

- **Existing personalized baselines achieve relatively superior performance by modeling individual preferences than standard non-personalized methods.** By explicitly incorporating user-level features, these methods generally surpass the non-personalized BT model. While the gains from PAL are limited in diverse real-world scenarios, it achieves a notable performance gain on the PersonalLLM dataset. In contrast, VPL and LoRe deliver more consistent and stable improvements across all evaluated tasks. Among these, LoRe emerges as the most competitive baseline by representing rewards in a low-rank subspace. Overall, these results highlight the critical role of personalization modeling in reward models for the task.

- **BehaviorRM establishes a new state-of-the-art across diverse datasets and backbone models.** Our method achieves state-of-the-art results on ChatbotArena, RedditTLDR, and PersonalLLM across all backbones, with an average accuracy of ***74.11%***, ***71.28%***, and ***89.65%***, with statistical significance verified via t-tests reported in the Appendix G.1. These results validate the effectiveness of leveraging behavioral similarity to enhance preference modeling. On PRISM, BehaviorRM surpasses the strongest baseline LoRe with the LLaMA backbone and, despite a slight drop with Qwen, still yields a higher average across the two backbones. We attribute this narrower margin to the substantially higher sparsity of PRISM relative to our other datasets (*c.f.*, Table 3 in Appendix), which makes behavioral similarity signals themselves sparse and inherently harder to learn through ID embedding updates—a well-known bottleneck in collaborative filtering.

5.3 Impact of Semantic and Behavioral Modeling

To investigate how semantic modeling and behavioral similarity modeling contribute to the performance of BehaviorRM, we conduct an ablation study by comparing different reward formulations under the same `Qwen3-1.7B` backbone model. Specifically, we evaluate: (1) *Semantic-only methods*, (2) *Behavior-only methods*, including Dot-Product and Cosine Similarity to model user–SID interactions, and (3) *Fusion modeling methods*, which integrates both behavioral and semantic signals via four strategies: *Feature Concatenation* (merging features into a single vector before scoring), *Multiplicative Interaction* (capturing non-linear dependencies via element-wise products), *Joint Projection* (mapping combined features through a shared latent layer), and our advanced strategy.

As shown in Table 2, behavior-only modeling captures fundamental structural regularities through discrete SIDs and provides meaningful predictions in most cases (>0.5), yet it consistently underperforms semantic approaches across all datasets. While semantic-only modeling serves as a strong baseline, it lacks the explicit modeling of the behavioral similarity grounding inherent in user interaction histories, leading to suboptimal performance in most cases. Together, these results demonstrate that relying on either signal in isolation is suboptimal, underscoring the need for an integrated fusion framework.

Across datasets, including ChatbotArena, PRISM, and PersonalLLM, all fusion strategies generally outperform both behavior-only and semantic-only methods. This success demonstrates that fusion

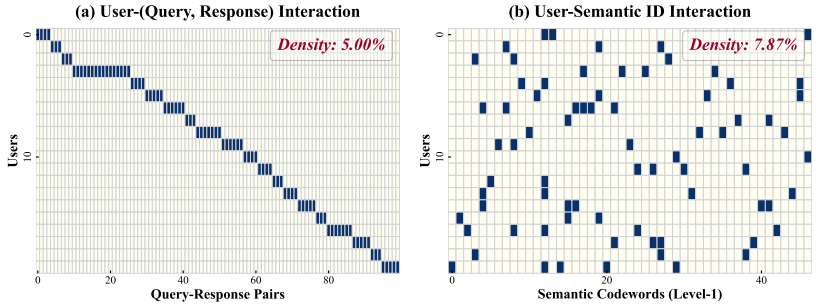

Figure 3: Visualization of interaction sparsity before and after semantic abstraction. (a) User–(Query, Response) interaction matrix built from raw prompt–response pairs, showing extreme sparsity (always being orthogonal between row vectors). (b) User–Semantic ID interaction matrix (level-1 codewords), where semantic IDs induce denser interactions (becoming non-orthogonal among row vectors).

successfully maintains deep semantic understanding while leveraging structural behavioral similarity (collaborative signals), thereby significantly enhancing personalized alignment performance. Among these variants, our advanced approach achieves the best results across all strategies by more effectively balancing and utilizing both signals, ensuring their complementary strengths are fully captured for superior, robust personalized alignment.

Notably, beyond the ablation study, we also analyzed the impact of different semantic ID generation strategies and hyperparameter settings in Appendix G.2 and G.3. The results show that our method is generally robust to these variations.

5.4 Visualization of Interaction Matrix

As discussed in Section 1, capturing behavioral similarity in LLM personalization is particularly challenging because user–(query, response) interactions lack co-occurrence signals. Consequently, collaborative signals cannot be captured in a straightforward manner as in traditional recommender systems. To address this issue, we propose a semantic ID–based modeling approach. In the following, we visualize the interaction matrices to illustrate how semantic IDs effectively alleviate this limitation.

Specifically, we randomly sample the first 100 interactions corresponding to 20 users from the ChatbotArena dataset. We visualize the interaction matrices from two perspectives: (1) the direct interaction matrix between users and raw (query, response) pairs; and (2) the interaction matrix between users and the quantized semantic IDs derived from (query, response) pairs. The results are shown in Figure 3. Due to space constraints, we present only the results from the first semantic layer, and we visualize only positive interactions. As shown in Figure 3(a), in the user–raw (query, response) interaction matrix, each (query, response) pair is uniquely associated with a single user. Consequently, the row vectors corresponding to different users are mutually orthogonal. Under 20 users, the overall matrix density is only 5%. In contrast, in the user–semantic ID interaction matrix (Figure 3(b)), the user row vectors are no longer strictly orthogonal. Instead, partial similarities emerge, as different users may interact with the same semantic IDs. Overall, the matrix density increases to 7.87%. This demonstrates that semantic ID transformation successfully introduces co-occurrence relationships, facilitating the capture of collaborative signals, i.e., behavioral similarity. Combined with the performance improvements observed in downstream tasks, these results validate the effectiveness of our method for behavior similarity modeling.

6 Conclusion

In this work, we investigated personalized reward modeling for personalized preference alignment. Moving beyond existing approaches that rely primarily on semantic information, we introduced BehaviorRM, a behavior-aware reward modeling framework that captures behavioral similarity through collaborative signals encoded by semantic IDs. Extensive experiments across four standard personalization backbones and four reward-model backbones demonstrate the consistent effectiveness of BehaviorRM, while detailed analyses further confirm its robustness and stability.

AI Assistance

We used large language models (LLMs) or other AI tools to aid or polish the writing of this paper.

Ethics Statement

We have read and will adhere to the ICLR Code of Ethics.

Reproducibility Statement

The method is described in Section 4 and the experimental setup in Section 5.1. Appendix D describes the datasets, Appendix E the baselines, Appendix F the implementation and hyperparameter settings, and Appendix G reports additional experimental details, including results from three independent runs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Avinandan Bose, Zhihan Xiong, Yuejie Chi, Simon Shaolei Du, Lin Xiao, and Maryam Fazel. Lore: Personalizing LLMs via low-rank reward modeling. In *Second Conference on Language Modeling*, 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.

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

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

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. PAD: Personalized alignment at decoding-time. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025a.

Xingzhuo Guo, Junwei Pan, Ximei Wang, Baixu Chen, Jie Jiang, and Mingsheng Long. On the embedding collapse when scaling up recommendation models. *arXiv preprint arXiv:2310.04400*, 2023.

Zhiliang Guo, Teng Tu, Yunshan Ma, and Xun Yang. Vibemus: Proactive agentic system for music personalization. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pp. 1–3, 2025b.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pp. 173–182, 2017.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 639–648, 2020.

Zhicheng He, Weiwen Liu, Wei Guo, Jiarui Qin, Yingxue Zhang, Yaochen Hu, and Ruiming Tang. A survey on user behavior modeling in recommender systems. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023*, pp. 6656–6664. ijcai.org, 2023.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

Song Jin, Juntian Zhang, Yong Liu, Xun Zhang, Yufei Zhang, Fei Jiang, Guojun Yin, Wei Lin, and Rui Yan. Tagging the thought: Unlocking personalization reasoning via reinforcement learning. *arXiv preprint arXiv:2509.23140*, 2025.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pp. 197–206. IEEE, 2018.

Jieyong Kim, Tongyoung Kim, Soojin Yoon, Jaehyung Kim, and Dongha Lee. Llms think, but not in your flow: Reasoning-level personalization for black-box large language models. *arXiv preprint arXiv:2505.21082*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

Yunyong Ko, Seongeun Ryu, and Sang-Wook Kim. Crown: A novel approach to comprehending users' preferences for accurate personalized news recommendation. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, pp. 1911–1921, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Rajeev Kumar, BK Verma, and Shyam Sunder Rastogi. Social popularity based svd++ recommender system. *International Journal of Computer Applications*, 87(14), 2014.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11523–11532, 2022.

Jialei Li, Yang Zhang, Yimeng Bai, Shuai Zhu, Ziqi Xue, Xiaoyan Zhao, Dingxian Wang, Frank Yang, Andrew Rabinovich, and Xiangnan He. Unigrec: Unified generative recommendation with soft identifiers for end-to-end optimization. *arXiv preprint arXiv:2601.17438*, 2026.

Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. Llara: Large language-recommendation assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 1785–1795, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314.

Zijie Lin, Yang Zhang, Xiaoyan Zhao, Fengbin ZHU, Fuli Feng, and Tat-Seng Chua. IGD: Token decisiveness modeling via information gain in LLMs for personalized recommendation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ygNaCTGUwJ.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, et al. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025a.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. A survey of personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*, 2025b.

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

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52516–52544, 2024.

Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. Latent inter-user difference modeling for LLM personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10599–10617. Association for Computational Linguistics, 2025a.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. Measuring what makes you unique: Difference-aware user modeling for enhancing LLM personalization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21258–21277. Association for Computational Linguistics, 2025b.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315, 2023.

Michael J. Ryan, Omar Shaikh, Aditri Bhagirath, Daniel Frees, William Held, and Diyi Yang. SynthesizeMe! inducing persona-guided prompts for personalized reward models in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8045–8078. Association for Computational Linguistics, 2025.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. Optimization methods for personalizing large language models through retrieval augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 752–762, 2024a.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392. Association for Computational Linguistics, 2024b.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pp. 285–295, 2001.

Jingyan Shen, Jiarui Yao, Rui Yang, Yifan Sun, Feng Luo, Rui Pan, Tong Zhang, and Han Zhao. MiCRo: Mixture modeling and context-aware routing for personalized preference learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17447–17463. Association for Computational Linguistics, 2025.

Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. Language model personalization via reward factorization. In *Second Conference on Language Modeling*, 2025.

Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Han Li. 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*, pp. 1294–1304, 2025.

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. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Yu Sun and Qicheng Liu. Collaborative filtering recommendation based on k-nearest neighbor and non-negative matrix factorization algorithm. *The Journal of Supercomputing*, 81(1):79, 2025.

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

Chengbing Wang, Yang Zhang, Wenjie Wang, Xiaoyan Zhao, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Think-while-generating: On-the-fly reasoning for personalized long-form generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lle0aGQyQb.

Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. Learnable item tokenization for generative recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2400–2409, 2024.

Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, et al. A survey on personalized and pluralistic preference alignment in large language models. *arXiv preprint arXiv:2504.07070*, 2025a.

Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, et al. A survey on personalized and pluralistic preference alignment in large language models. *arXiv preprint arXiv:2504.07070*, 2025b.

Yiyan Xu, Wenjie Wang, Yang Zhang, Biao Tang, Peng Yan, Fuli Feng, and Xiangnan He. Personalized image generation with large multimodal models. In *Proceedings of the ACM on Web Conference 2025*, pp. 264–274, 2025a.

Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. Personalized generation in large model era: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24607–24649, 2025b.

Chao Zhang, Shiwei Wu, Haoxin Zhang, Tong Xu, Yan Gao, Yao Hu, and Enhong Chen. Notellm: A retrievable large language model for note recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 170–179, 2024a.

Kepu Zhang, Teng Shi, Weijie Yu, and Jun Xu. Prlm: Learning explicit reasoning for personalized rag via contrastive reward optimization. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 5484–5488, 2025a.

Yang Zhang, Fuli Feng, Jizhi Zhang, Keqin Bao, Qifan Wang, and Xiangnan He. Collm: Integrating collaborative embeddings into large language models for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2329–2340, 2025b.

Yang Zhang, Wenxin Xu, Xiaoyan Zhao, Wenjie Wang, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Reinforced latent reasoning for LLM-based recommendation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=eUtIZT2ONS.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*, 2024b.

Weixiang Zhao, Xingyu Sui, Yulin Hu, Jiahe Guo, Haixiao Liu, Biye Li, Yanyan Zhao, Bing Qin, and Ting Liu. Teaching language models to evolve with users: Dynamic profile modeling for personalized alignment. In *Advances in Neural Information Processing Systems*, 2025a.

Xiaoyan Zhao, Ming Yan, Yilun Qiu, Haoting Ni, Yang Zhang, Fuli Feng, Hong Cheng, and Tat-Seng Chua. Steerx: Disentangled steering for llm personalization. *arXiv preprint arXiv:2510.22256*, 2025b.

Xiaoyan Zhao, Juntao You, Yang Zhang, Wenjie Wang, Hong Cheng, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. Nextquill: Causal preference modeling for enhancing LLM personalization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Guorui Zhou, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Qiang Luo, Qianqian Wang, Qigen Hu, Rui Huang, Shiyao Wang, et al. Onerec technical report. *arXiv preprint arXiv:2506.13695*, 2025.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. PersonalLLM: Tailoring LLMs to individual preferences. In *The Thirteenth International Conference on Learning Representations*, 2025.

A Perspective and Future Directions

BehaviorRM opens a new perspective for personalized preference alignment by introducing behavioral similarity as a critical pathway for reward modeling. Our current exploration focuses on modeling behavioral similarity via semantic IDs, representing only an initial step in this direction. Future work can investigate more general approaches, including end-to-end joint modeling of semantic IDs and rewards, as well as non–semantic–ID–based methods. Beyond reward modeling in personalized preference alignment, we believe that behavioral similarity modeling may also benefit other lines of LLM personalization — personalized text generation, by providing richer collaborative signals of user preference. We encourage further exploration along this promising research direction.

B Scope and Limitations

Consistent with prior work, we focus on reward modeling rather than the full RL with human/AI feedback pipeline. As established by extensive literature, reward modeling remains the most critical component of this pipeline. Thus, this design choice reflects the current scope of our study rather than an inherent limitation of BehaviorRM. Our method can be easily applied to the full RL tuning process.

In addition, our approach relies on learned SID embeddings to capture behavioral similarity signals, and—much like collaborative filtering in recommender systems—it imposes certain requirements on data density. When user interactions are extremely sparse, behavioral similarity signals are themselves scarce and inherently harder to learn through ID embedding updates, which limits the gains BehaviorRM can deliver. Mitigating this sparsity bottleneck is a promising direction for future work.

C Additional Related Work

C.1 LLM Personalization

LLM personalization has emerged as a pivotal research direction, moving beyond the "one-size-fits-all" alignment paradigm toward a pluralistic alignment that respects the diversity of individual human values Chen et al. (2024); Zhang et al. (2024b); Liu et al. (2025b). Current research in this field can be broadly categorized into two primary directions: personalized generation Xu et al. (2025b) and personalized preference alignment Xie et al. (2025b).

• **Personalized generation**. Personalized generation tailors multi-modal outputs to individual user contexts Zhao et al. (2025b); Xu et al. (2025a); Guo et al. (2025b). In the textual domain, existing methods generally fall into training-free and training-based methods. Training-free adaptation primarily leverages personalized prompting and retrieval-augmented generation (RAG) to inject user-specific context into the model input Salemi et al. (2024b); Qiu et al. (2025b); Kim et al. (2025). For instance, CFRAG Shi et al. (2025) introduces collaborative filtering into the RAG framework to leverage histories of similar users, while ROPG Salemi et al. (2024a) optimizes the dense retriever to select the most supportive historical records. In addition, some training-free methods adopt a steering perspective, directly modifying the model's internal states (*e.g.*, activations) to guide the output toward individual user preferences. In contrast, training-based optimization directly optimizes the model's parameters to capture user-specific nuances Salemi et al. (2024b); Zhao et al. (2026); Zhang et al. (2025a); Wang et al. (2026); Jin et al. (2025). Examples include OPPU Tan et al. (2024), which trains dedicated per-user PEFT modules, and DEP Qiu et al. (2025a), which models inter-user differences in the latent space to construct soft prompts. There are also works exploring more advanced tuning-based methods, such as the causality-driven method NextQuill Zhao et al. (2026) or the reasoning-driven method FlyThinker Wang et al. (2026). Although both focus on personalization, our work is less related to this line of research; instead, it is more related to another personalized preference alignment direction.

• **Personalized preference alignment**. Personalized preference alignment is closely related to preference alignment in LLM post-training, but it targets alignment with **individual** human values and preferences. Existing work typically focuses on modeling personalized reward functions Jang

et al. (2023); Chen et al. (2025b); Shen et al. (2025). Individual human preferences are inherently heterogeneous. To effectively capture such preference heterogeneity, recent researchers have proposed various personalized reward modeling strategies. PAL Chen et al. (2025a) and VPL Poddar et al. (2024) utilize prototype-based clusters and latent virtual profiles, respectively, to represent user-specific rewards. More recently, LoRe Bose et al. (2025) and PReF Shenfeld et al. (2025) both represent personalized rewards in a low-dimensional subspace with basis functions. Other approaches, such as SynthesizeMe Ryan et al. (2025), induce synthetic user personas from historical interactions to guide reward modeling. Although these methods improve alignment accuracy to some extent, they primarily focus on semantic-level modeling and overlook the rich information embedded in similar behavioral patterns across user histories. For personalization, behavioral signals are crucial for modeling user interests, as has been widely demonstrated in recommender systems. Motivated by this observation, we propose BehaviorRM to explicitly address this limitation.

C.2 Behavior Similarity Modeling

Behavioral similarity modeling aims to support preference modeling by analyzing similarities in users' historical behaviors, and it has been widely studied in recommender systems He et al. (2023). In recommendation, behavioral similarity is often referred to as collaborative signals, typically captured through co-occurrence patterns (*e.g.*, items that are jointly consumed). A classic example is the beer and diapers phenomenon, where semantically unrelated items nonetheless frequently co-occur in user interactions. To model these signals, a wide range of collaborative filtering methods have been developed, evolving from early heuristic approaches (item-based and user-based CF) Sarwar et al. (2001), to latent factor models such as MF and SVD++ Koren et al. (2009); Kumar et al. (2014), and later to neural approaches, including MLP-based He et al. (2017), attention-based Kang & McAuley (2018), and graph-based models He et al. (2020), all of which have shown strong effectiveness in capturing collaborative signals. However, these methods inherently rely on the assumption that items are interacted with by multiple users, which breaks down in personalized LLM settings, where a (query, response) pair, as the most basic analogue of an item, is rarely shared across users.

Inspired by the success of LLMs, recent work has introduced semantic-ID-based generative recommendation methods Wang et al. (2024); Zhou et al. (2025); Li et al. (2026), exemplified by TIGER Rajput et al. (2023), which replace raw item IDs with quantized semantic IDs to enable autoregressive modeling. While these methods primarily aim to mimic language modeling for generation, our work takes a different perspective by extending semantic-ID modeling to LLM personalization. We transform (query, response) pairs into semantic IDs, thereby inducing effective co-occurrence structures. This design enables the extraction of behavioral similarity in personalization scenarios, providing a principled way to recover collaborative signals that are otherwise unavailable.

D Dataset Details

In this section, we briefly introduce the datasets used in our experiments:

- *ChatbotArena* Zheng et al. (2023) consists of real-world conversations where users provide pairwise judgments on outputs from different LLMs in a blind-test setting.
- *PRISM* Kirk et al. (2024) is a diverse feedback dataset featuring multiple live conversations between participants from 75 countries and 21 different LLMs.
- *RedditTLDR* Stiennon et al. (2020) is adapted from the Reddit summarization task, where each entry includes a post, two candidate summaries, and a unique worker ID.
- *PersonalLLM* Zollo et al. (2025) is a semi-synthetic dataset where each user is represented by a preference vector $\mathbf{w}$ sampled from a Dirichlet distribution $\mathrm{Dirichlet}(\alpha)$. We set $\alpha = 0.01$ to simulate a moderately diverse user population.

The statistics for these datasets after data pre-processing are presented in Table 3.

E Baseline Details

In this section, we provide a detailed description of each baseline method used in our experiments.

Table 3: Statistical summary of four personalized reward datasets, including the number of users, user sparsity, and the distribution for the training, validation, and test splits.

Dataset	#User	User Sparsity	#Data		
			Training	Validation	Test
ChatbotArena	904	13.12	7157	2328	2377
PRISM	731	6.46	2888	892	944
RedditTLDR	51	1820.71	55718	18571	18567
PersonalLLM	1000	55	45000	5000	5000

Table 4: Performance comparison between representative baselines and BehaviorRM. Mean and standard deviation are reported using the `Llama3.2-1B` backbone. Best results are bolded, and the symbol * indicates p-value < 0.05 in t-tests.

Methods	Chatbot Arena	PRISM	Reddit TLDR	Personal LLM
BT	$64.88_{\pm 0.59}$	$50.28_{\pm 0.96}$	$64.67_{\pm 0.57}$	$79.73_{\pm 0.17}$
LoRe	$73.01_{\pm 0.26}$	$53.71_{\pm 0.21}$	$68.00_{\pm 0.21}$	$87.54_{\pm 0.16}$
BehaviorRM	$\mathbf{73.83}^{*}{}_{\pm 0.18}$	$\mathbf{53.99}_{\pm 0.82}$	$\mathbf{69.43}^{*}{}_{\pm 0.16}$	$\mathbf{89.93}^{*}{}_{\pm 0.07}$

- **Ref** Liu et al. (2025a): A baseline that directly employs the pre-trained reward models for evaluation without any personalized adaptation, serving as a non-personalized performance reference.
- **BT** Bradley & Terry (1952): A standard reward model fine-tuned using the Bradley-Terry objective on the aggregated data of all users. It learns a unified, global preference model without considering individual user preferences.
- **PAL** Chen et al. (2025a): A modular reward modeling framework for pluralistic alignment that incorporates diverse human preferences by modeling individual preferences as combinations of K learnable preference prototypical vectors and user-specific weights.
- **VPL** Poddar et al. (2024): A variational framework that captures diverse individual preferences by inferring user-specific latent variables z from few-shot feedback to parameterize a conditional reward model $r_\phi(s, z)$, where s denotes the input state.
- **LoRe** Bose et al. (2025): A low-rank reward modeling framework that represents personalized reward functions as semantically weighted combinations of shared basis functions in a low-dimensional subspace. It achieves personalization by learning user-specific weights for these basis functions to reconstruct individual preference signals.

F Implementation Details

For baseline models, we rely on their author-provided implementations and follow the hyperparameter settings recommended in the original papers. In this work, we use the Adam Kingma & Ba (2015) optimizer for all methods. For our method, we tune the hyperparameter α over the range $\{0.05, 0.1, 0.3, 0.5, 0.7, 0.8, 0.9\}$, and further refine the search with a step size of 0.01 when necessary. As our method involves MF, we additionally introduce a regularization term to mitigate overfitting, implemented as the weight decay in the optimizer, and tune it within the range $\{1e\text{-}2, 1e\text{-}3, 1e\text{-}4, 0\}$. For the embedding dimension, we search it in {16, 32, 64, 96, 128}. For the RQ-VAE to obtain semantic IDs, we use three codebooks, each with a codebook size of 512. Experiments are run on one NVIDIA A100 GPU with 80GB of GPU memory.

G Additional Experiments

G.1 Statistical Significance

To rigorously evaluate the stability of BehaviorRM, we conduct multiple repeated trials to assess its robustness against stochastic variations. In this setting, we select two representative and competitive baselines: BT and LoRe (the best-performing baseline). We report the average performance accompanied by the standard deviation derived from three independent runs, each initialized with a distinct

random seed. To further validate the reliability of the observed improvements, we also perform paired t-tests compared with LoRe.

As shown in Table 4, BehaviorRM consistently achieves the highest mean accuracy across all four datasets while maintaining remarkably low standard deviations. Notably, on the ChatbotArena, RedditTLDR, and PersonalLLM datasets, our method is proven to be statistically significant via paired t-tests, its superior robustness and stability across different random initializations.

G.2 Impact of Quantization Methods

We evaluate the sensitivity of BehaviorRM to the underlying quantization strategy (*i.e.,* semantic ID construction strategy) by comparing our default Residual-Quantized Variational AutoEncoder (RQ-VAE) against a simplified Residual-Quantized KMeans (RQ-KMeans) variant Zhou et al. (2025). When implementing the RQ-Kmeans variant of our method, we directly replace RQ-VAE with RQ-Kmeans while keeping all other components fixed. All results are evaluated using the `Qwen3-1.7B` backbone.

Figure 4 summarizes the results, with the best-performing baseline LoRE as reference. As the figure shows, BehaviorRM exhibits strong robustness to the choice of quantization method. Specifically, on three of the four datasets, RQ-Kmeans achieves performance comparable to RQ-VAE and preserves a similar performance ranking relative to the baselines. Only on the PersonalLLM dataset does RQ-Kmeans exhibit a noticeable deviation from RQ-VAE. These results indicate that BehaviorRM is largely insensitive to the specific quantization strategy employed. Nevertheless, RQ-Kmeans generally underperforms RQ-VAE, which may be attributed to the learned nature of RQ-VAE enabling more effective quantization for semantic ID construction, whereas RQ-Kmeans relies on heuristic clustering guided by predefined similarity measures.

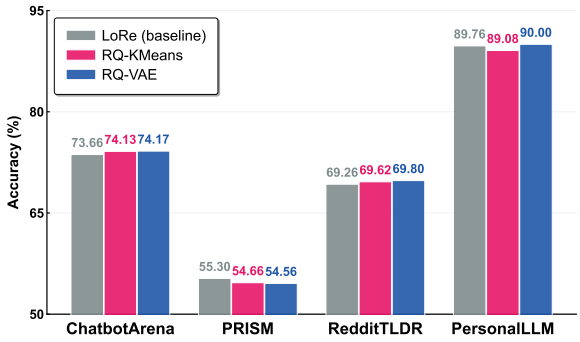

Figure 4: Performance comparison of LoRe and different BehaviorRM variants under two semantic ID quantization strategies (RQ-VAE and RQ-KMeans).

G.3 Hyperparameter Sensitivity Analysis

In this section, we analyze the sensitivity of BehaviorRM to two key hyperparameters: (1) the balancing coefficient α between semantic reward modeling and behavior similarity signals, and (2) the dimensionality d_b of SID embeddings, which controls the capacity of user–semantic ID interaction modeling. All results are reported on the PersonalLLM dataset using the `Qwen3-4B` backbone.

Effect of the balancing coefficient α. Figure 5 illustrates the sensitivity of accuracy to the balancing coefficient α. At low values of α, the behavior-aware component is insufficiently weighted, resulting in suboptimal alignment. Performance steadily improves as the behavioral signal is strengthened, reaching a peak at $\alpha = 0.5$. However, exceeding this

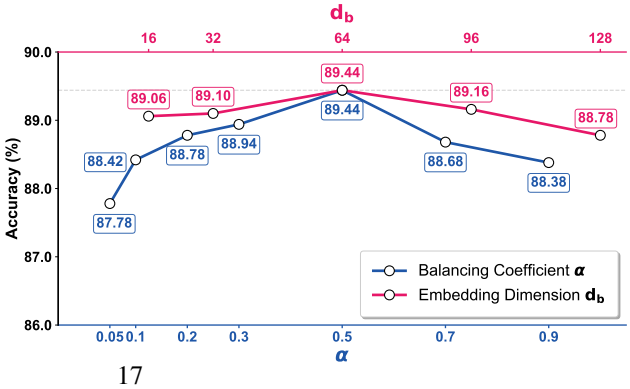

Figure 5: Hyperparameter sensitivity analysis of the proposed BehaviorRM with respect to two key parameters: (1) the balancing

threshold leads to a decline in performance, suggesting that overemphasizing behavior similarity may overshadow informative semantic reward signals. We adopt $\alpha = 0.5$ as our default configuration to maintain an optimal balance between these two signals. Notably, BehaviorRM is robust across a wide range of α (performing better than the best baseline, *c.f.,* Table 1), requiring minimal parameter tuning.

Effect of SID embedding dimension d_b. Figure 5 also shows the effect of varying the SID embedding dimension d_b. We observe that the accuracy improves as d_b increases from 16 to 64, suggesting that a moderate embedding capacity is essential for capturing latent user–semantic interactions. However, performance gradually declines beyond this threshold, indicating diminishing returns and potential over-parameterization. Overall, BehaviorRM achieves strong results with relatively low-dimensional embeddings. Moreover, similar to the trends observed for the α, our BehaviorRM is robust across a wide range of the embedding dimension, and could even be much more robust.

Together, these results demonstrate that the effectiveness of BehaviorRM is rooted in principled modeling of semantic–behavioral trade-offs, ensuring robustness across diverse configurations.