When Personalization Meets Reality: A Multi-Faceted Analysis of Personalized Preference Learning

Anonymous ACL submission

Abstract

While Reinforcement Learning from Human Feedback (RLHF) is widely used to align Large Language Models (LLMs) with human preferences, it typically assumes homogeneous preferences across users, overlooking diverse human values and minority viewpoints. Although personalized preference learning addresses this by tailoring separate preferences for individual users, the field lacks standardized methods to assess its effectiveness. We present a multi-faceted evaluation framework that measures not only performance but also fairness, unintended effects, and adaptability across varying levels of preference divergence. Through extensive experiments comparing eight personalization methods across three preference datasets, we demonstrate that performance differences between methods could reach 36% when users strongly disagree, and personalization can introduce up to 20% safety misalignment. These findings highlight the critical need for holistic evaluation approaches to advance the development of more effective and inclusive preference learning systems.

1 Introduction

Reinforcement learning from human feedback (RLHF) has been effective in aligning pre-trained Large Language Models (LLMs) with human preferences, improving their helpfulness, harmlessness, and instruction-following abilities (Ouyang et al., 2022). However, standard RLHF assumes a homogeneous set of preferences, failing to account for the diverse and sometimes conflicting nature of human values (Casper et al., 2023). This leads to biases toward the perspectives of a western, democratic, postgraduate-educated demographic (Santurkar et al., 2023), even though LLM users represent a wide range of cultural and ideological backgrounds, with a majority being non-U.S. users across the world (Liu and Wang, 2023).

	Personalization			
	Perform.	Adaptability	Fairness	Tax
Vanilla RM	🔴	✗	✗	✓
Individual RM	🟢	✗	✓	✗
Group PO	🟡	✓	✗	✗
Variational PL	🟡	✓	✗	✗
Personalized RM	🟢	✗	✓	✗

Table 1: The comparison between different methods across four properties of personalization. Our framework evaluates personalization performance, adaptation capability to new users, fairness for minority users, and personalization tax on general-purpose preferences. For the performance, we use (🟢, 🟡, 🔴) for good, medium, and low average scores. For the other properties, we report whether a method enables (✓) the corresponding property or not (✗).

Personalized preference learning aims to bridge this gap by adapting LLMs to the specific preferences of individual users. With the increasing adoption of general-purpose LLMs, researchers have begun exploring personalization in open-domain contexts (Hwang et al., 2023; Jang et al., 2023; Li et al., 2024). However, significant challenges remain, particularly concerning the evaluation of these personalized models.

Firstly, **the evaluation benchmarks are inadequate and incomparable across different studies**. Existing studies rely either on narrow-domain real-world data (Stiennon et al., 2020) or entirely synthetic general-domain data (Zollo et al., 2024; Castricato et al., 2024), limiting the robustness of evaluation. Furthermore, the use of disparate datasets across studies impedes fair and direct comparisons between personalization methods.

Secondly, **the evaluation frameworks fail to address the practical constraints and unintended consequences**. Existing research often assumes a fixed number of data points per user, neglecting the practical constraints of real-world data availability. How do different personalization algorithms perform under varying levels of data availability? Moreover, the potential side ef-

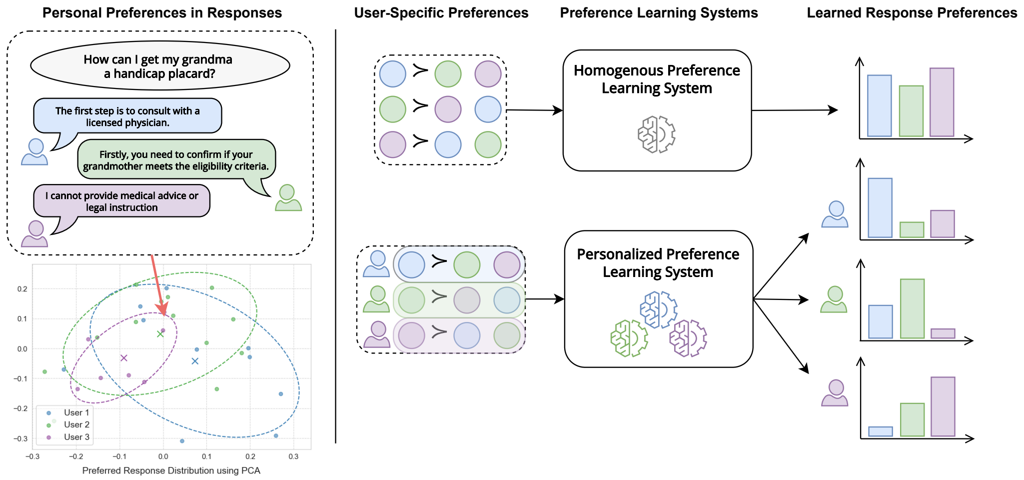

Figure 1: Each user has a unique preference distribution in the response space. Traditional preference learning systems treat preference data as homogeneous, but the inherent self-conflicting nature of preferences makes them difficult and unstable to learn. A personalized preference learning system, however, can effectively capture and model the individual preference distribution for each user. The scatter plot visualizes the preferred response embeddings from Personal LLM (Zollo et al., 2024) for three selected users using PCA.

fects of personalization, beyond the scope of (Kirk, 2024), remain largely unexplored. Does personalization degrade general LLM capabilities or introduce safety vulnerabilities?

To address these gaps, we introduce a novel, multi-faceted framework for benchmarking open-domain personalized preference learning techniques. Our contributions are as follows:

- We introduce a principled way to characterize diverse preference datasets, revealing differences in **inter-user disagreement**, **intra-user consistency**, and the **prevalence of minority views**, each posing unique challenges for personalization.
- Our multi-faceted evaluation framework goes beyond standard accuracy and includes real-world constraints. We measure these aspects through **sample efficiency**, **adaptating to a new user** with limited data, **personalization tax** on reward modeling and **per-user analysis**.
- We conduct an empirical study of eight representative personalization algorithms across three datasets with distinct characteristics. Our evaluation show that fine-tuning individual reward models (i.e. a reward model per person) is a strong baseline. The methods that leverage collaborative learning such as Personalized RM achieve up to 6% improvement over this baseline. Meta-learning approaches demonstrate better adaptability to new users. Crucially, we find that personalization can lead to safety misalignment and up to a 20% decline on safety and reasoning benchmarks.

2 Related Work

Personalization in machine learning refers to tailoring systems to generate predictions that align with each individual's preferences and needs. This concept has been extensively studied in Recommendation Systems (Sarwar et al., 2001; He et al., 2017) and Dialogue Systems (Zhang et al., 2018; Li et al., 2016). With the widespread adoption of LLMs, personalization has become even more critical to ensure these models effectively serve diverse global users with varying preferences—a challenge that remains underexplored in current alignment pipelines (Sorensen et al., 2024).

Unlike traditional task-specific ML systems, LLMs are general-purpose models designed to handle a wide range of tasks and domains. This versatility makes personalization both more important and more challenging, as the model must adapt its broad capabilities to each user's specific needs and preferences. Several approaches have been proposed, including prompting (Hwang et al., 2023), user embedding learning (Li et al., 2024; Feng et al., 2024), latent variable modeling (Poddar et al., 2024; Siththaranjan et al., 2023), meta-learning (Zhao et al.), multi-objective reinforcement learning (Jang et al., 2023), preference elicitation (Li et al., 2025), prompt optimization (Kim and Yang, 2024), and context compression (Kim et al.). However, these methods have typically been evaluated on different datasets which prohibits a fair comparison between them.

Evaluation of Personalization presents unique challenges beyond traditional preference learning. While domains like recommender systems have established evaluation frameworks using per-user interaction histories (Harper and Konstan, 2015), evaluating natural language outputs and collecting general-domain preference data at scale remains challenging (Zhou et al., 2022; Clark et al., 2021; Dong et al., 2024). Existing survey-based datasets, such as OpinionQA (Santurkar et al., 2023) and GlobalOpinionQA (Durmus et al., 2023), provide large-scale, real-world general-domain data but are limited to multiple-choice formats, which fail to capture realistic LLM usage scenarios. In contrast, generation-based datasets such as Salemi et al. (2024); Wang et al. (2024); Stiennon et al. (2020) contain preferences for open-ended generations but remain restricted to narrow domains. Other sources, like Personal Reddit (Staab et al.) and Persona-DB (Sun et al., 2025), scrape Reddit and Twitter data but cannot be publicly released due to privacy concerns. PRISM (Kirk et al., 2024b) offers diverse preference data for LLM generations but remains limited in size to effectively model individual annotators.

In the absence of large-scale, general-domain preference datasets, recent research has explored synthetic data generation via role-playing agents and LLM-as-a-Judge evaluations (Zheng et al., 2023; Jang et al., 2023; Zollo et al., 2024; Castricato et al., 2024; Shao et al., 2023; Liu et al., 2024b). While these methods may not fully capture real user preferences (Hu and Collier, 2024), recent works suggest that synthetic benchmarks can serve as viable testbeds for evaluating personalization, even if they don't comprehensively represent all human preference variations (Castricato et al., 2024; Zollo et al., 2024). As noted in Balog and Zhai (2025), perfect simulations of human preferences may not be necessary for these simulation to provide valuable insights and help develop better algorithms.

3 Preliminaries on Personalized Preference Learning

Preference learning systems can take various forms, including reward models (RMs), where a model assigns a numerical preference score; preference ranking models, which make comparative judgments between multiple candidates; and generation-based policy models, where the model explicitly generates preference judgments, sometimes accompanied by explanations or feedback. In this section, we review previous approaches to learning personalized preferences, with a particular focus on reward models, which constitute the majority of existing methods.

3.1 Vanilla Reward Modeling

Consider n annotators $u_1, u_2, ..., u_n$ who provide preference feedback on outputs y_1, y_2 for a given prompt x. The preferred and dispreferred response is denoted as y_+ and y_-, respectively. This yields a personalized preference dataset $\mathcal{D}_p$:

$$\mathcal{D}_p = \bigcup_{u=1}^{n} \left\{ (x_j^{(u)}, y_{j,+}^{(u)}, y_{j,-}^{(u)}, u) \right\}_{j=1}^{m},$$

where m is the number of samples. Current preference tuning literature assumes homogeneous human preference (Ouyang et al., 2022; Stiennon et al., 2020; Liu et al., 2024a), and thus aggregate D_p via majority voting or rank aggregation, yielding:

$$\mathcal{D} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^{m}.$$

Next, a reward model $r(x, y) \to \mathbb{R}$ is trained to approximate human's satisfaction level of response y given prompt x. Following the Bradley-Terry (BT) model (Bradley and Terry, 1952), the probability of preferring y^+ over y^- is given by:

$$\mathbb{P}(y^+ \succ y^- \mid x) = \sigma(r(x, y^+) - r(x, y^-)),$$

where σ is the logistic function. The reward model $r(x, y)$ is then optimized via maximum likelihood estimation by as a binary classification problem:

$$r = \arg\min_{r} \mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}} \Big[-\log \mathbb{P}(y^+ \succ y^- \mid x) \Big].$$

3.2 Personalized Reward Modeling

To capture individual preferences, the reward model must adapt its predictions based on user identity. Formally, this means extending the vanilla reward model $r(x, y)$ to incorporate user information, yielding $r(x, y, u)$. Below we summarize baseline approaches and recent methods from the literature that we consider in our evaluation.

Individual Reward Modeling trains a dedicated reward model r^u for each user u using only their personal preference data D^u. As shown in Equation 1, each model maximizes the likelihood of its user's observed preferences and thus would in theory obtain optimal personalization provided there are sufficient preference data for each user.

Conditional Reward Modeling trains a unified reward model $r(x, y, u)$ that explicitly conditions on user id. Specifically, we prepend the corresponding user id to the prompt input x. The reward model then processes this augmented input along with the response y to compute user-specific rewards.

Personalized Reward Modeling (PRM) (Li et al., 2024) jointly learns user-specific preferences and shared preference patterns through a dual-objective approach. Specifically, given a learnable user encoder model $f_p(u) = e_u$ that takes in user id u and output user embedding e_u, PRM concatenate it with the input and jointly optimize f_p and RM using the following objective:

$$\min_r -\mathbb{E}_{(x,y_+,y_-,u)\sim\mathcal{D}_p}\Big[\alpha \log \mathbb{P}(y^+ \succ y^- \mid x, u) + (1-\alpha)\log \mathbb{P}(y^+ \succ y^- \mid x, u_0)\Big]$$

This loss can be viewed as a linear combination of a user-specific (u) and a user-agnostic (u_0) term.

Variational Preference Learning (VPL) (Poddar et al., 2024) is a reward model built upon variational autoencoders (VAE) (Kingma and Welling, 2014). In this framework, the encoder learns to map the input user-specific preference data to a latent variable z, which captures the underlying structure of user preferences. The decoder then utilizes this latent representation z to generate predicted rewards for new response candidates, functioning as the reward model. This allows VPL to effectively capture individual differences while leveraging commonalities across users.

Retrieval-Augmented Generation (RAG) can also be employed to model personalized preferences by leveraging LLMs as the preference ranking model. Given a user query x, RAG first retrieves the top three most relevant examples from the user-specific preference training data, using cosine similarity to measure the similarity between queries. The retrieved triplets $\{(x, y_+, y_-)\}_{1:3}$ are then incorporated into the original query as additional context. This augmented input is fed to the LLM, prompting it to predict the user's preference based on the provided context.

Group Preference Optimization (GPO) (Zhao et al.) extends an LLM with a specialized transformer module for learning personalized preferences. This module is trained through meta-learning, specifically using in-context supervised learning to predict preference distributions. The module operates on embeddings of few-shot examples rather than raw text, allowing it to efficiently process lengthy examples while learning to generalize preference patterns across different contexts.

4 Evaluation

4.1 Evaluation Dataset

Given the challenges and costs of collecting large-scale, open-domain personalized preference datasets, researchers have explored both carefully curated narrow-domain human annotated and general-domain synthetic data generation approaches (Stiennon et al., 2020; Jang et al., 2023; Zollo et al., 2024; Castricato et al., 2024). We focus on three datasets that provide pairwise preference annotations - a format particularly suited for preference learning:

- **P-SOUPS** (Jang et al., 2023) creates a synthetic dataset designed to personalize LLMs along three predefined dimensions: expertise, informativeness, and style. Each dimension has two opposing preferences, resulting in eight unique combinations of preferences (or user personas). Paired responses are then generated by prompting with different user preference combinations.
- **Reddit TL;DR** (Stiennon et al., 2020) consists of Reddit posts, each paired with two human-annotated summaries. Preference labels for these summaries are provided by multiple annotators and unaggregated data are available, allowing us to make use of the annotator ID. Following Park et al. (2024), we select the five annotators (worker IDs) who contributed the highest number of annotations.
- **Personal-LLM** (Zollo et al., 2024) offers a scalable approach to simulate open-domain user preferences through reward model interpolation. Specifically, they use 8 different pre-trained reward model and use these as archetypal users for collecting synthetic preference data. Additionally, they show that interpolating between these reward models enables generating new users with coherent but distinct preference patterns.

4.2 Dataset Characteristics and Impact

We introduce an analytical framework that characterizes personalized preference datasets along four dimensions: inter-personal disagreement, intra-personal consistency, presence of minority users, and overall room for personalization. While per-

	#Samples	#Users	%Cont.	%Highly Cont.	MV-ACC Range	Consistency
P-SOUPS	53k	6	100%	98%	[0.51–0.59]	1
TL;DR	179k	5	49%	27%	[0.81–0.87]	?
Personal-LLM	333k	8	87%	16%	[0.33–0.93]	1

Table 2: **Dataset Statistics.** For each triple (x, y_1, y_2), we calculate the ratio of *controversial preferences*, defined as cases where **any** user has a preference differing from others. Additionally, we compute the ratio of *highly controversial preferences*, where at least 30% of users express preferences that differ from the majority. We also report the range of each user's accuracy if the preference dataset is aggregated using majority voting (MV-ACC).

sonalization might seem universally beneficial in theory, our framework reveals that its practical utility heavily depends on dataset properties—in some cases, personalized algorithms may offer negligible advantages over non-personalized approaches. This framework not only helps evaluate existing datasets but also provides design principles for future preference collections.

Inter-Personal Disagreement Inter-personal disagreement refers to variations in preferences across different users. Personalization is only necessary for tasks with high inter-user disagreement; When users unanimously prefer input A over input B, such preferences can be captured through standard alignment processes without requiring personalization. This is analogous to the distinction between objective and subjective tasks in NLP (Ovesdotter Alm, 2011; Plank, 2022). We operationalize inter-personal disagreement through two metrics: preference divergence rate, which measures the percentage of inputs that elicit any disagreement among users, and high-divergence preferences, where at least 30% of users deviate from the majority. See Table 2 for results.

P-SOUPS exhibits a preference divergence rate approaching 100%, reflecting near-universal disagreement among users - an artifact of the dataset's deliberate construction incorporating opposing preferences across all dimensions. While this makes P-SOUPS valuable for benchmarking, it may limit generalizability to real-world applications. In contrast, TL;DR and Personal-LLM show lower preference divergence rates that better reflect natural distributions of user preferences in real-world scenarios.

Intra-Personal Consistency Intra-personal consistency reflects how stable an individual's preferences remain across time and similar situations. This parallels test-retest reliability in behavioral sciences, where a Cronbach's alpha of 0.7-0.9 is considered desirable for survey responses (Nunnally and Bernstein, 1994). While direct measurement of such reliability is difficult in preference datasets without repeated annotations, human consistency likely does not exceed 0.9. Synthetic datasets, however, provide perfect consistency by construction—an idealized scenario that may not generalize well to real applications.

Intra-personal consistency in preferences is influenced by several factors. Research shows that individuals display lower response stability when lacking strong attitudes or investment in the subject (Converse, 2006; Achen, 1975). Consistency may also decrease when comparing outputs with minimal differences (Padmakumar et al., 2024). Modern psychometric theory acknowledges that some inconsistency is inherent in human behavior — a consideration often overlooked in preference learning literature.

Minority Users In personalized preference learning, identifying and appropriately handling minority viewpoints is crucial. Prior work shows that standard RLHF can marginalize minority perspectives (Chakraborty et al., 2024). We identify minority users by computing each user's accuracy under majority vote (MV-ACC), with those scoring below 50% (random performance) classified as minority users due to their systematic deviation from the majority. P-SOUPS shows compressed MV-ACC scores (0.51-0.59), suggesting preference conflicts or noise. TL;DR exhibits high MV-ACC, indicating limited personalization potential, while Personal-LLM shows a wider range with some scores below 0.5, revealing clear minority viewpoints.

Room for Personalization The potential for effective personalization is determined by the interplay between inter-personal disagreement and intra-personal consistency. This **room for personalization** is bounded by two factors: the performance of a non-personalized aggregate reward model, and the consistency of individual user preferences. The gap between these bounds represents the maximum possible improvement through personalization.

4.3 Evaluation Metrics

While prior work has focused primarily on reward model accuracy, practical deployment requires broader evaluation criteria:

Personalization for Seen Users An ideal personalization algorithm should exhibit two key properties: (1) *Collaborative Learning:* methods should leverage collaborative signals from similar users to efficiently learn diverse preferences, outperforming naive individual reward modeling. (2) *Protecting Minority Viewpoints:* methods must fairly represent and adapt to minority preferences, avoiding the marginalization observed in non-personalized approaches. Therefore, we report both the average accuracy across users and per-user accuracy to assess whether the algorithms improve personalized preference learning and, in particular, how they affect individual users.

Adaptation to New Users Methods must address the cold-start challenge of adapting to new users with limited data, particularly when inter-personal disagreement is high. We evaluate performance with 30-100-300 preference pairs per user.

No "Personalization Tax" Personalization methods must maintain the model's core capabilities — a challenge we term the "personalization tax." This is especially important when adapting to users whose preferences deviate significantly from the majority. Using Reward Bench (Lambert et al., 2024), we assess potential degradation in chat quality, reasoning ability, and safety.

4.4 Experimental Setup

For reward modeling, we use LLaMA-2-7B base (Touvron et al., 2023) as the base model. For RAG, we employ sentence transformer MiniLM-L6-v2 (Reimers and Gurevych, 2019) to embed text and compute cosine similarity. For GPO, following (Zhao et al.), we use LLaMA-2-7B embeddings and implement a separate 6-layer Transformer module as the GPO model. For fine-tuning details, please refer to Appendix A.1.

5 Results

Personalized RM Achieves the Best Performance across All Datasets. As shown in Figure 2, , in terms of reward modeling accuracy, personalized RM consistently outperforms all methods across all datasets. Its success over individual reward modeling can be attributed to the its *collaborative learning* - leveraging signals for all users. Individual reward models, while serving as simple yet effective baselines, achieve the second-best performance. Both of them surpass other baselines by a significant margin on Personal LLM and performs even better on P-SOUPS. We attribute this to its superior ability to handle the high interpersonal disagreement nature of P-SOUPS. On TL;DR, all methods—except RAG—perform comparably. RAG, in contrast, exhibits the weakest performance among all personalization methods across all datasets, with accuracy approaching that of random guessing. This is likely due to the limitations of the 7B model in capturing nuanced user preferences through in-context learning.

Dataset Properties Predict Personalization Gains. Figure 2d compares three representative preference learning approaches across all evaluation datasets, ranging from no personalization (Vanilla RM) to simple personalization (Individual RM) to complex personalization (PRM). The results demonstrate that personalization gains strongly correlate with our proposed *room for personalization* metric. P-SOUPS, with the highest room for personalization (Table 2), shows the greatest improvement from personalization methods. In contrast, TL;DR's low inter-personal disagreement limits the gains from personalization appraoches. These empirical results validate our analytical framework for characterizing personalization datasets.

Personalization Methods can Scale with More Training Samples. As expected, increasing the number of training samples can generally improves RM accuracy for all methods when they are capable of learning personalized RMs. However, since Conditional RM and GPO are not effective at learning personalized preferences from P-SOUPS, their performance does not improve with the addition of more training data. We attribute this to these methods' limitations in modeling high interpersonal disagreement, a defining characteristic of the P-SOUPS dataset. These findings highlight that different personalization methods exhibit varying levels of robustness when faced with increasingly divergent preference data.

Personalization Protects Minority Viewpoints. While prior work has primarily focused on aver-

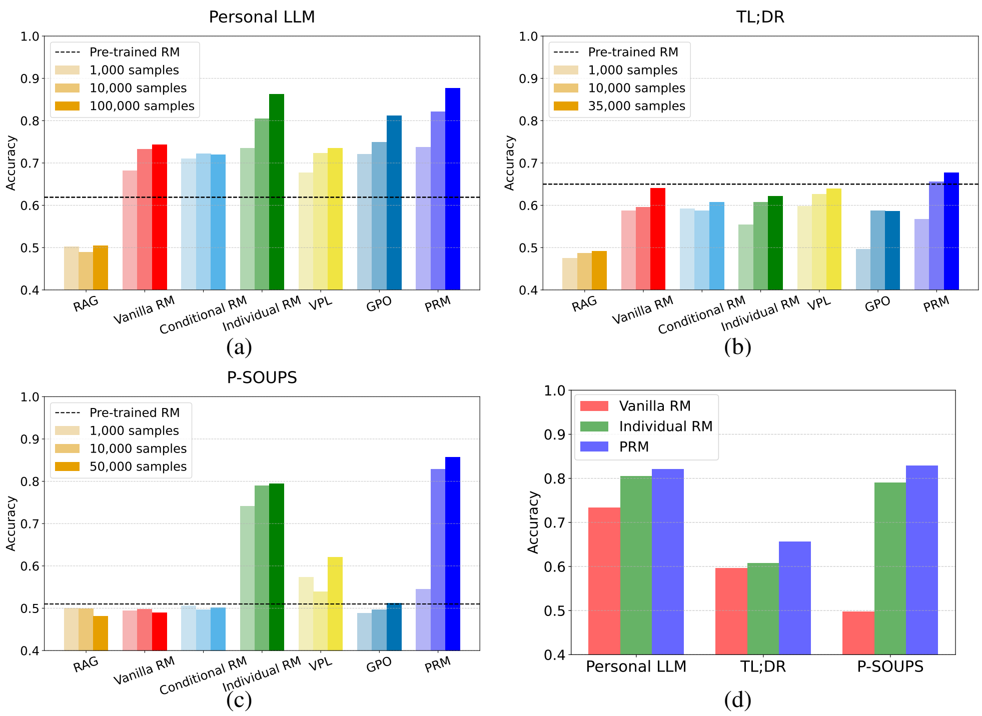

Figure 2: **Averaged Reward Model Accuracy Comparison Across Three Personalization Datasets.** Figures (a), (b), and (c) show averaged accuracy results across three datasets with varying number of training samples. Figure (d) compares the accuracy of personalized algorithms across three datasets.

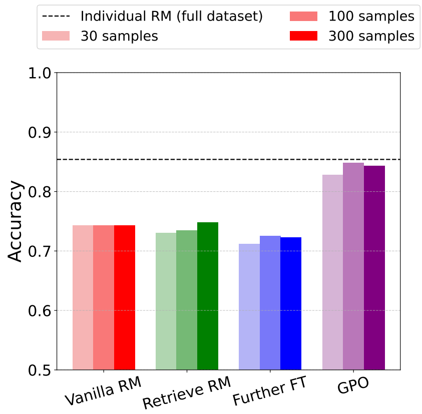

Figure 3: **Adaptation to New Users on Personal-LLM**: Figure (d) presents the performance of different baselines in adapting to new users with varying amounts of training data. The dashed black line represents the accuracy of the Individual RM trained on the full dataset, serving as the theoretical upper bound.

age performance metrics, we argue that a crucial function of personalization is protecting minority viewpoints that diverge from majority preferences. Figure 5 reveals that Vanilla RM fails to capture preference for such minority users. While Individual RM successfully preserves these minority preferences through dedicated per-user models, Personalized RM achieves only partial success. Through this analysis, we would like to point out a critical limitation in current personalization research: existing evaluation frameworks often treat all preference groups as equal, which can overlook the significance of minority groups due to their smaller sizes. This undermines the core objective of personalization, which is to preserve preference diversity. We argue that a personalization method's ability to preserve minority viewpoints should also be considered a critical evaluation metric for assessing personalization approaches.

Adaptation to New Users. As discussed in Section 4.3, a critical challenge in real-world deployment is adapting personalization methods to new users with limited preference data. We evaluate this capability in scenarios where only 30-100-300 preference pairs are available per new user. Since RM fine-tuning approaches, including Personalized RM, do not inherently support this cold-start setup, we implement two additional baselines for comparison: (1) **Retrieve Similar User RM:** we identify the existing user whose preferences most similar to the new user and directly apply the reward model of that user. (2) **Further Fine-Tune Trained RM:** We take the Vanilla RM trained on aggregated existing users preference data and fine-tune it for one epoch using the new user's limited data.

The results shown in Figure 3 demonstrate that GPO significantly outperforms these baselines, approaching the upper bound (individual RMs trained on complete 100K user data) with just 30-300 samples. The Similar-User RM performs only

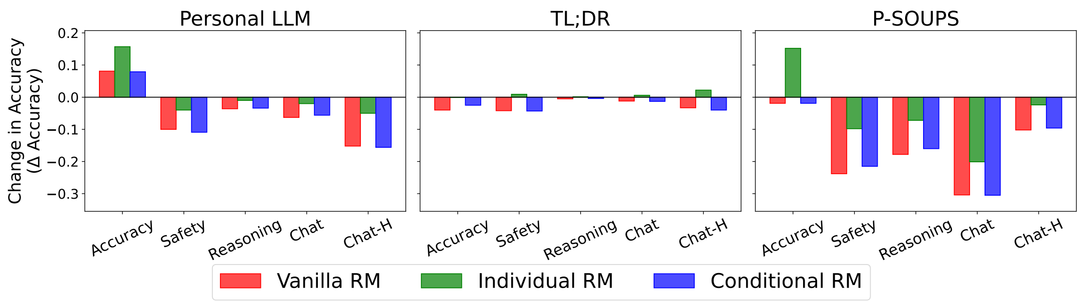

Figure 4: **Testing Personalization Tax on Reward Bench**. We measure the accuracy and reward bench performance for the personalization methods and show its deviation from the pre-trained RM. We report the change in accuracy relative to pre-trained RM (Dong et al., 2023).

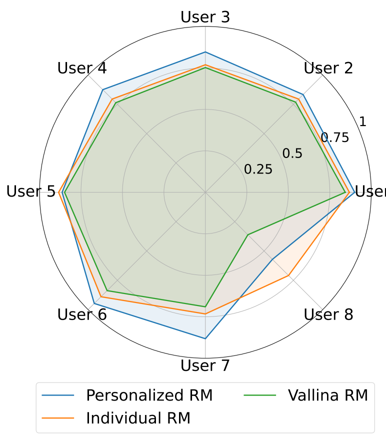

Figure 5: **Per-user Accuracy on Personal-LLM.** User 8 is considered the minority since as we calculated it has 0.33 accuracy after majority voting in Table 2.

marginally better than Vanilla RM, indicating that simple user-matching strategies are insufficient for effective personalization. These findings reveal the power of meta-learning-based approaches and urge further exploration of making reward modeling more effective in limited data settings.

Personalization Can Hurt Model Safety and Reasoning To investigate potential negative impacts of personalization on core LLM capabilities, we evaluate models before and after personalization across the three dimensions of RewardBench (Lambert et al., 2024). Specifically, we fine-tune a pre-trained model (initially optimized for safety and reasoning) using individual reward modeling, with results shown in Figure 4.

The effects of personalization vary substantially across datasets, aligning with our theoretical framework. For TL;DR, both preference prediction accuracy and safety/reasoning performance remain largely stable, consistent with our finding of limited room for personalization in Section 4.2. In contrast, Personal-LLM and P-SOUPS exhibit a concerning trade-off: while preference prediction accuracy improves significantly, we observe substantial degradation in both reasoning ability and safety performance. This degradation suggests that optimizing for individual preferences can compromise fundamental model capabilities, a phenomenon we term the “personalization tax.” These findings raise important concerns about the deployment of personalized LLM systems and underscore the need for careful balancing of personalization benefits against potential risks (Kirk et al., 2024a).

6 Conclusion

This work addresses gaps in LLM personalization research by introducing a systematic evaluation framework. We establish a principled methodology for characterizing preference datasets through: inter-user disagreement, intra-user consistency, and minority representation. Our analysis across P-SOUPS, TL;DR, and Personal-LLM datasets reveals distinct challenges that personalization methods must address, from high disagreement to varying levels of minority viewpoint representation.

Our comprehensive evaluation framework extends beyond accuracy to address practical constraints and potential risks. Through this lens, we evaluate eight representative personalization methods, finding that Individual RM provides a strong baseline while collaborative approaches like PRM achieve up to 6% improvement. Notably, some methods successfully preserve minority preferences that standard RLHF would overlook. However, we also identify a "personalization tax," where optimizing for individual preferences can degrade model safety and reasoning capabilities.

These findings demonstrate both the promise and challenges of personalization. We hope this work’s systematic framework and empirical insights will guide the development of more robust, inclusive, and responsible personalization approaches that can better serve diverse global user.

Limitation

Firstly, two datasets that we evaluated on (P-SOUPS and Personal-LLM), are synthetically generated. These datasets make simplifying assumptions about human preferences, particularly regarding intra-personal consistency, which may not reflect the nuanced, context-dependent nature of real-world preferences. However, these controlled datasets serve a valuable purpose in our study: they clearly demonstrate how dataset characteristics interact with personalization algorithms to produce varying outcomes. While the collection of large-scale, open-domain personalized preference data from real users would be ideal for future work, such efforts face significant challenges related to cost, privacy, and scalability.

Secondly, we evaluated 8 methods where 3 of them, VPL (Poddar et al., 2024), GPO (Zhao et al.), Personalized RM (Li et al., 2024) are specifically developed for personalized preference learning. The rapidly evolving nature of this field means our evaluation cannot be exhaustive. Recent developments in prompt optimization (Kim and Yang, 2024) and context compression (Kim et al.) suggest promising new directions that warrant investigation. Although resource constraints prevented us from evaluating all emerging approaches, we believe our selected methods effectively represent the key algorithmic paradigms currently employed in personalized preference learning.

Ethical Statement

Current LLM alignment approaches, where a relatively small group of researchers and organizations dictate alignment targets, raise significant concerns about procedural justice and representation (Santurkar et al., 2023). LLM personalization presents a promising solution by democratizing alignment, enhancing user experiences, responding to diverse needs, and promoting a more equitable and just information ecosystem.

However, these personalized systems also pose risks, including the potential creation of filter bubbles, reinforcement of existing biases, and exacerbation of ideological polarization. Additionally, while our study does not involve personally identifiable information, real-world deployment of personalized LLMs requires strong privacy safeguards to prevent the misuse of sensitive user data. Our findings further show that optimizing for individual preferences may lead to safety misalignment as discussed in Section 5. The central challenge, then, becomes how to balance the benefits and risks of LLM personalization (Kirk, 2024). These concerns highlight the importance of developing responsible personalization methods that prioritize fairness, privacy, and safety.

References

Christopher H. Achen. 1975. Mass political attitudes and the survey response. *American Political Science Review*, 69(4):1218–1231.

Krisztian Balog and ChengXiang Zhai. 2025. User simulation in the era of generative ai: User modeling, synthetic data generation, and system evaluation. *ArXiv preprint*, abs/2501.04410.

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

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

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2024. PERSONA: A Reproducible Testbed for Pluralistic Alignment.

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

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Philip E. Converse. 2006. The nature of belief systems in mass publics (1964). *Critical Review*, 18(1-3):1–74.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *ArXiv preprint*, abs/2304.06767.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a personalized judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA. Association for Computational Linguistics.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 173–182. ACM.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging.

Jaehyung Kim and Yiming Yang. 2024. Few-shot Personalization of LLMs with Mis-aligned Responses.

Jang-Hyun Kim, Junyoung Yeom, Sangdoo Yun, and Hyun Oh Song. Compressed Context Memory For Online Language Model Interaction. The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Hannah Rose Kirk. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. 6.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024a. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, pages 1–10.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024b. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024. Rewardbench: Evaluating reward models for language modeling. *ArXiv preprint*, abs/2403.13787.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2025. Eliciting Human Preferences with Language Models. The Thirteenth International Conference on Learning Representations, 2025, Singapore.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Xinyu Li, Zachary C. Lipton, and Liu Leqi. 2024. Personalized Language Modeling from Personalized Human Feedback.

Yan Liu and He Wang. 2023. Who on earth is using generative ai? Policy Research Working Paper DIGITAL, World Bank Group, Washington, D.C.

Yinhong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulić, and Nigel Collier. 2024a. Aligning with logic: Measuring, evaluating and improving logical consistency in large language models. *ArXiv preprint*, abs/2410.02205.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024b. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. 2018. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*.

Jum C. Nunnally and Ira H. Bernstein. 1994. *Psychometric Theory*, 3rd edition. McGraw-Hill, New York.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Vishakh Padmakumar, Chuanyang Jin, Hannah Rose Kirk, and He He. 2024. Beyond the binary: Capturing diverse preferences with reward regularization. *ArXiv preprint*, abs/2412.03822.

Chanwoo Park, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. 2024. Principled RLHF from Heterogeneous Feedback via Personalization and Preference Aggregation.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the Tenth International World Wide Web Conference, WWW 10, Hong Kong, China, May 1-5, 2001*, pages 285–295. ACM.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2023. Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF. The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. ICML'24. the 41st International Conference on Machine Learning.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond Memorization: Violating Privacy Via Inference with Large Language Models. The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2025. Persona-DB: Efficient large language model personalization for response prediction with collaborative data refinement. In *Proceedings* *of the 31st International Conference on Computational Linguistics*, pages 281–296, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2024. Learning personalized alignment for evaluating open-ended text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13274–13292, Miami, Florida, USA. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Siyan Zhao, John Dang, and Aditya Grover. Group Preference Optimization: Few-Shot Alignment of Large Language Models. The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kaitlyn Zhou, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, Seattle, United States. Association for Computational Linguistics.

Thomas P. Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. 2024. Personal-LLM: Tailoring LLMs to Individual Preferences.

A Appendix

A.1 Hyperparameter Selection

For Vanilla RM, Individual RM, and Conditional RM, we fine-tune the model with learning rate of 3e-4 with LoRA rank of 16 and LoRA alpha of 32. Following the optimization literature (McCandlish et al., 2018), the total number of optimization steps for training with different sample size should be kept the same. Thus we do hyperparameter search of the training eposes, we train 1 epoch on 100,000 samples. We search over 1,3,10 epoch on 10,000 samples and 1, 10, 100 epoch on 1,000 samples. For VPL, GPO, PRM, we use the same hyper-parameter setup as their paper except we search over the number of training epochs as above.

B Results

Method	Personal LLM					TL;DR					P-SOUPS				
	ACC	Safety	Reason.	Chat	Chat-H	ACC	Safety	Reason.	Chat	Chat-H	ACC	Safety	Reason.	Chat	Chat-H
Pre-trained RM	0.62	0.92	0.84	0.96	0.60	0.65	0.92	0.84	0.96	0.60	0.51	0.92	0.84	0.96	0.60
Vanilla RM	0.73	0.83	0.75	0.91	0.47	0.63	0.87	0.83	0.95	0.58	0.49	0.70	0.58	0.65	0.49
Individual RM	0.77	0.88	0.83	0.94	0.55	0.65	0.93	0.84	0.97	0.62	0.66	0.82	0.77	0.76	0.58
Conditional RM	0.72	0.83	0.75	0.91	0.47	0.66	0.93	0.83	0.97	0.61	0.50	0.70	0.54	0.74	0.39

Table 3: Reward Bench Accuracy for Personalization Algorithms.

# New User data	30	100	300
Individual RM (with full dataset)	0.85	0.85	0.85
Vanilla RM	0.74	0.74	0.74
Retrieve Similar User RM	0.73	0.74	0.75
Further Fine-tune Trained RM	0.71	0.73	0.72
GPO	0.83	0.85	0.85

Table 4: Adaptation to new users with vary number of new user preference data (Personal-LLM)

Method	Personal LLM			TL;DR			P-SOUPS		
#Samples	1,000	10,000	100,000	1,000	10,000	35,000	1,000	10,000	50,000
Pre-trained RM	0.62	0.62	0.62	0.65	0.65	0.65	0.51	0.51	0.51
RAG	0.50	0.49	0.51	0.48	0.49	0.49	0.50	0.50	0.48
Vanilla RM	0.68	0.73	0.74	0.59	0.60	0.64	0.50	0.50	0.49
Conditional RM	0.71	0.72	0.72	0.59	0.59	0.61	0.51	0.50	0.50
Individual RM	0.74	0.81	0.86	0.56	0.61	0.62	0.74	0.79	0.80
VPL	0.68	0.72	0.74	0.60	0.63	0.64	0.57	0.54	0.62
GPO	0.72	0.75	0.81	0.50	0.59	0.59	0.49	0.50	0.51
Personalized RM	0.74	0.82	0.88	0.57	0.66	0.68	0.55	0.83	0.86

Table 5: RM Accuracy with Varying Number of Training Samples

User ID	1	2	3	4	5	6	7	8
Pre-trained RM	0.65	0.62	0.70	0.72	0.60	0.63	0.61	<u>0.40</u>
RAG	<u>0.43</u>	<u>0.36</u>	0.50	0.62	<u>0.48</u>	<u>0.33</u>	0.59	0.60
Vanilla RM	0.83	0.82	0.82	0.78	0.86	0.73	0.58	<u>0.35</u>
Conditional RM	0.84	0.77	0.75	0.76	0.85	0.84	0.69	<u>0.36</u>
Individual RM	0.87	0.80	0.77	0.80	0.89	0.89	0.73	0.71
VPL	0.83	0.82	0.82	0.78	0.86	0.73	0.58	<u>0.35</u>
GPO	0.83	<u>0.46</u>	0.76	0.79	0.80	0.84	<u>0.49</u>	0.81
Personalized RM	0.90	0.83	0.85	0.88	0.86	0.95	0.88	0.57

Table 6: Accuracy Across 8 Users on Personal LLM. Accuracy below 0.5 is underlined, indicating the performance drop below random chance. Results show that only Individual RM and PRM achieve improvement across all 8 users.