A Survey on Personalized and Pluralistic Preference Alignment in Large Language Models

Zhouhang Xie[1], **Junda Wu**[1], **Yiran Shen**[1], **Raghav Jain**[1], **Yu Xia**[1], **Xintong Li**[1], **Aaron Chang**[2], **Ryan Rossi**[3], **Tong Yu**[3], **Sachin Kumar**[4], **Bodhisattwa Prasad Majumder**[5] **Jingbo Shang**[1], **Prithviraj Ammanabrolu**[1], **Julian McAuley**[1]

[1]University of California, San Diego [2]University of California, Los Angeles
[3]Adobe Research [4]The Ohio State University [5]Allen Institute for AI
{zhx022,jw069,jes038,r6jain,yux078,xil240,prithvi,jmcauley}@ucsd.edu
aaronchang21@g.ucla.edu, {ryrossi, tyu}@adobe.com
kumar.1145@osu.edu, bodhisattwam@allenai.org

Abstract

Personalized preference alignment for large language models (LLMs), the process of tailoring LLMs to individual users' preferences, is an emerging research direction spanning the area of NLP and personalization. In this survey, we present an analysis of works on personalized alignment and modeling for LLMs. We introduce a taxonomy of preference alignment techniques, including training time, inference time, and additionally, user-modeling based methods. We provide analysis and discussion on the strengths and limitations of each group of techniques and then cover evaluation, benchmarks, as well as open problems in the field.

1 Introduction

Recently, significant progress has been made in aligning LLMs to the *overall* preferences of users (Zhao et al., 2024; Shen et al., 2023). However, prior works show that there is no one-size-fits-all solution for preference alignment (Sorensen et al., 2024; Kirk et al., 2024a). Intuitively, while there are universal preferences that are shared across users, such as "it is good to respond in the tone of a friendly and helpful assistant", user preferences are often also individualized and use-case dependent (i.e., contextual). For example, users might have different preferences towards the tone and style of responses (Jang et al., 2023), and even for the same user, the preferred style of response would change depending on the context of interaction, akin to prior works in contextualized personalization (Meng et al., 2023). For example, an expert might have different preferences from a beginner when asking the same question about a concept, and even the same person might have different preferences depending on when they interact with the system.

Along with the emergence of these challenges, the notion of personalization in the context of NLP and LLM research has recently attracted increasing attention within the NLP and machine learning communities (Sorensen et al., 2024; Kirk et al., 2024a; Flek, 2020). At first glance, personalization seems to be an intuitive solution for catering LLMs to individualized and contextual preferences described above. However, in practice, the notion of personalization for LLMs is convoluted, ranging from applications of LLMs to classical personalization tasks (Tan & Jiang, 2023; Chen et al., 2024c) to role-playing and simulation of individual human behaviors (Chen et al., 2024d; Mou et al., 2024). As we shall discuss later (Section 7), not all types of personalization benefit individual users in a conversation setting.

In this work, we focus on personalized and pluralistic preference alignment, a specific notion of personalization that aims at adapting an LLM's behavior to dynamic user *preferences* across individuals, groups, and contexts to enhance user satisfaction. We start by defining the problem formulation for personalized preference alignment (Section 3), then introduce an intuitive technique taxonomy covering training and test-time methods for preference

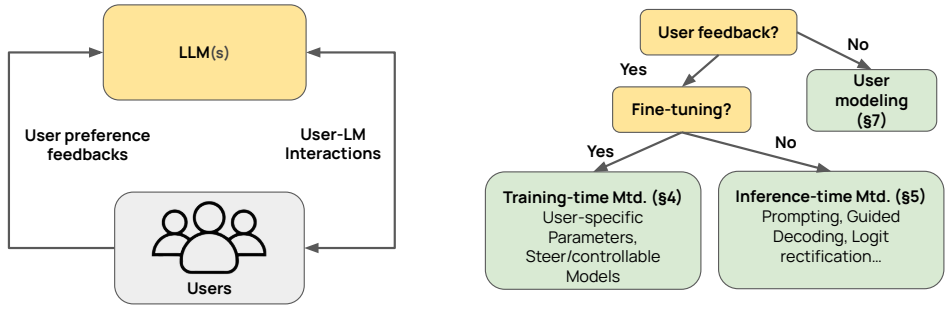

Figure 1: (1a) Overview on personalized preference alignment for LLMs. This includes training (Section 4) and test-time (Section 5) methods, leveraging various feedbacks such as verbal feedback and choices. (1b) An over-simplified decision tree for determining the class of method to use for personalized preference alignment.

alignment (Section 4, Section 5). We then discuss the connection between user modeling and personalized preference alignment (Section 7), and discuss benchmark and evaluation (Section 6) as well as open problems in the field (Section 8).

What's covered? Broadly speaking, as shown in Figure 1, personalization in the context of preference alignment can be categorized into two classes of methods: Training-time (Section 4) and inference-time (Section 5) personalization that leverages personalized user feedback at different stages of LLMs' life cycle. Additionally, when user preference feedback is absent, personalized adaption of LLMs and LLM-based systems can sometimes also be achieved by modeling the user, which we discuss in Section 7.

What's not covered. Under the theme of LLMs and personalization, there are other relevant areas such as user-behavior modeling (i.e., systems that predict the behavior of users) (Tan & Jiang, 2023; Wu et al., 2024b; Chen et al., 2024c), user-group-behavior modeling (i.e., LLM role-playing) (Chen et al., 2024d;d; Tseng et al., 2024), and LLMs *for* personalized recommender systems (e.g., LLMs as a component in product recommender systems) (Tan & Jiang, 2023; Wu et al., 2024b; Chen et al., 2024c). However, these lines of work does not involve directly catering the behavior of LLMs or LLM-based conversational systems to user preferences. We point interested readers to related surveys (Section 2) on these topics.

Contribution statement. Despite numerous recent efforts on consolidating works that involve LLM and personalization (Sorensen et al., 2024; Kirk et al., 2024a; Zhang et al., 2024b; Wu et al., 2024a; Tan & Jiang, 2023; Wu et al., 2024b; Chen et al., 2024c), we show personalized alignment is an emerging valuable research direction, with its own methods and evaluation paradigm. To this end, this survey provides a comprehensive overview of personalized and pluralist alignment while establishing its differences and connections to adjacent domains. Further, personalized preference alignment is an emerging domain with no universally acknowledged evaluation and benchmarks. To make it easier for future research to access existing evaluation schemes, we provide an overview of the current state of evaluation methods. Overall, this survey serves as an up-to-date resource for practitioners and researchers working on personalized and pluralistic alignment and, more broadly, LLM personalization and alignment.

2 Related Surveys and Key Differences

User modeling and personalizing LLMs to simulate user behavior. Recently, a few position papers have discussed the concept of personalized preference alignment (Sorensen et al., 2024; Kirk et al., 2024a). However, these works focus on discussing the implications without summarizing current progress. There have also been surveys on personalizing LLMs (Zhang et al., 2024b; Wu et al., 2024a), but they do not differentiate between the replicating user

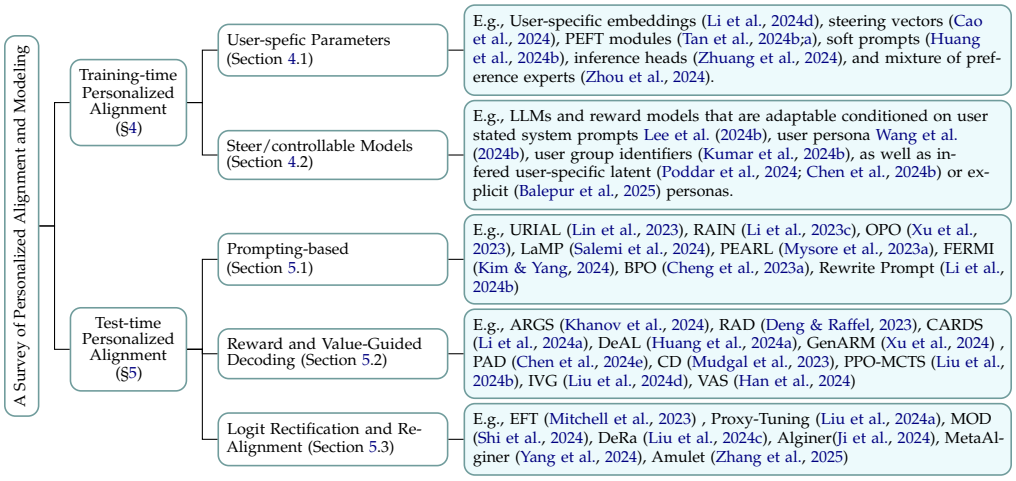

Figure 2: Technique taxonomy on personalized and pluralistic preference alignment.

behavior to catering for user preferences. Finally, another stream of recent works focuses on LLM-based role-laying (Tseng et al., 2024). However, their focus is still on empowering LLMs to replicate certain user groups' behavior. In contrast, our work focuses on summarizing the progress with respect to personalized preference alignment.

Application of LLMs to downstream personalization tasks. There are also a few surveys on the *application* of LLMs to other tasks related to personalization, such as recommender systems and user modeling (Tan & Jiang, 2023; Wu et al., 2024b; Chen et al., 2024c), personalized wearable devices (Li et al., 2024e), personalization on the web (Chen et al., 2024c), narrative-drive recommendation (Mysore et al., 2023b), and LLM-based role-laying (Chen et al., 2024d), which are different from the scope covered in this work since the goal of these lines of research is not to build an optimal LLM-based dialogue system. We provide further clarification of our scope in Section 3, and discuss in more detail the relationship between (personalized) user modeling and preference alignment in Section 7.

3 Problem Formulation and Techniques Taxonomy

3.1 Problem Statement

We begin by introducing the notion of personalization in the context of preference alignment, as illustrated in Figure 1-a. In this setting, system deployers aim to build an LLM (or LLM-based system) that caters specifically to each individual user's unique preferences.

For clarity, let $\mathcal{X}$ denote the input space (e.g., user queries or contextual prompts), and $\mathcal{Y}$ denote the output space (e.g., responses generated by the LLM). We formalize the personalized reward function as

$$r : \mathcal{X} \times \mathcal{Y} \times \mathcal{U} \to \mathbb{R},$$

which measures how well an LLM's response y to an input x satisfies the unique preferences of user u. Our goal is to learn a set of individualized policies $\{\pi_u\}_{u \in \mathcal{U}}$, where each policy π_u is tailored exclusively to user u. Such an objective can be expressed as:

$$\pi_u^* = \arg\max_{\pi_u} \mathbb{E}_{x \sim \mathcal{X}, y \sim \pi_u(\cdot|x,u)} \left[r(x, y, u) \right] \quad \forall u \in \mathcal{U} \tag{1}$$

In other words, for each user u, the optimal policy π_u^* is one that, in expectation over the distribution of inputs, generates responses that best satisfy that user's individual preferences. However, compared to personalization which emphasizes individualized prediction for each user, approaches for achieving pluralistic alignment are more diverse, sometimes involving systems that produce sets of outputs catering for different preferences. Meanwhile, the LLM-based policies can often have shared parameters, or are built from a single model conditioned on different user information, as we shall discuss in Section 4 and Section 5.

3.2 Technique Overview

The problem formulation discussed above yields a straight-forward way to partition existing methods, as shown in Figure 2. Specifically, in order to obtain a LLM-based policy that cater to each individual's preferences, it is crucial to be able to adapt the LLM itself depending on the interacting user. Naturally, this can be achieved both by *training* models with user-specific parameters (Section 4.1) or making the model steerable with respect to user inputs (Section 4.2), as shown in Figure 1-b. On the other hand, fine-tuning or adapting LLMs is frequently costly. This challenge motivates another category of works that aims at influencing a pre-trained LLMs' behavior at inference time, with methods such as prompting (Section 5.1), controlled-decoding (Section 5.2), and logit manipulation (Section 5.3). Additionally, we note that there are personalization techniques that nevertheless improve base LLMs towards the goal of catering to individual preferences without using user feedback, such as building user-specific memories (e.g., (Yuan et al., 2025; Zhang et al., 2022)) following the assumption that users generally likes to be remembered. We provide discussion of this complementary class of valuable techniques in Section 7.

3.3 Granularity of Personalization

While personalized preference alignment aims to cater to, ultimately, individualized preferences, personalization typically suffers from the issue of sparse feedback (Li et al., 2023b). Specifically, an individual user's interactions with the system are frequently too few to allow meaningful learning to happen. To this end, similar to works in adjacent personalization areas such as recommender systems (Li et al., 2023b), it is often helpful to exploit the fact that there can be *groups* of users that shares similar preferences, thus bypassing the feedback sparsity issue. Similar to these prior works, personalized LLM alignment can also happen on different granularity: individual users' levels and user groups based on user profiles and social relationships (Sorensen et al., 2024; Kumar et al., 2024b). We note that there is also a special case of "contextual" shared preference between user groups, where a set of users momentarily shares preferences based on their purpose of interacting with LLMs. For example, a group of users that are seeking help from an LLM-based therapy chatbot may collectively wish the LLM to act as a helpful therapist Stade et al. (2024). We provide discussions on this special case in Section 7.

4 Training-time Personalized Alignment

To effectively adapt LLMs and LLM-based systems to personalized user preferences, a straight-forward solution is to develop specialized models via training. In this section, we introduce two popular classes of techniques: building models with user-specific parameters and training models that are sensitive to input user preferences.

4.1 Learning from Feedbacks with User-specific Parameters

Motivation. One of the most straightforward ways to build LLM that caters to individualized user preference is by keeping separate parameters for each user (or user group), effectively learning a set of policies whose behavior slightly deviates from each other without explicitly specified user preference under the standard RLHF setting.

Comparative Analysis. Addressing the challenge of inferring user-specific preferences without requiring explicit specification, (Li et al., 2024d) employs a lightweight user model to learn from human feedback, while (Cao et al., 2024) follows a similar setting, but learn per-user steering-vectors to achieve personalization. Following similar intuition, (Tan et al., 2024b) and (Tan et al., 2024a) advocates for a dedicated per-user PEFT module that stores individual behavior patterns. Other than adapter modules, (Huang et al., 2024b) explores a complimentary parameter-efficient personalization approach via soft prompts. (Zhuang et al., 2024) introduces HYDRA, a factorization framework that couples a shared base model with user-specific heads, yielding notable improvements over prompt-based methods. (Zhou et al., 2024) introduces RLPHF, which merges outputs from specialized

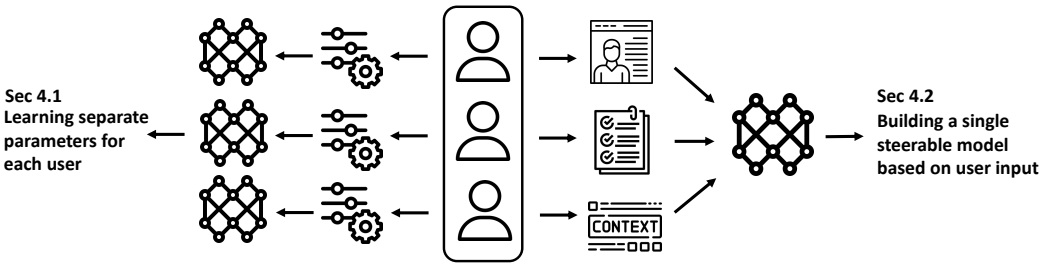

Figure 3: Personalized language model alignment during training time.

expert LLMs using a lightweight Preference Control Model (PCM) that dynamically adjusts token predictions based on user context; i.e., mixture of preference experts. Finally, in contrast to learning user-specific parameters in LLMs, (Park et al., 2024a) explored learning multiple reward models to handle the trade-off between bias and variance in personalized preference alignment.

Limitations. Despite these advances, existing personalized RLHF frameworks often struggle to simultaneously ensure personalized adaptation and global model performance, with many approaches relying on complex multi-stage processes or additional components that may hinder model scalability (Park et al., 2024b; Han et al., 2024; Lee et al., 2024a). Furthermore, challenges remain in robustly handling heterogeneous and strategic feedback while integrating efficient privacy-preserving techniques, pointing to the need for more streamlined and resilient solutions such as federated personalized alignment (Zhang et al., 2024a; Jiang et al., 2024).

4.2 Building Steerable Model that Adapts Responses Based on User Input

Motivation. Another popular choice for personalization is building a single base model that's steerable (Sorensen et al., 2024), where are *single* models' behavior changes based on the user it is interacting with, similar to prior research for controlled text generation (Hu et al., 2017). This line of research often assumes user inputs relevant to individualized preferences are available, such as explicitly stated preferences or personas that imply user preferences are available at inference time, bypassing the need for user-specific parameters.

Comparative Analysis. Following the assumptions that user can provide their own preferences as prompts to LLM, (Lee et al., 2024b) train an instruction-following LLM that can adapt to user-written values in system prompt via data synthesis, (Wang et al., 2024b) builds base models geared towards llm-as-a-judge (Zheng et al., 2023) use-cases that can adapt to explicitly stated user personas. Aside from building steerable LLM policies, such adaptability can also be built into reward models. For example, (Pitis et al., 2024) builds context-conditioned reward models, which are then used for personalized (i.e., context-conditioned) alignment. Similarly, (Kumar et al., 2024b) builds reward models conditioned on user group identifiers on Reddit to achieve personalization. We note that there are works that attempt to infer latent user representations when users don't explicitly state their preference, but still build steerable models based on these latent representations. For example, (Balepur et al., 2025) infer user persona from choices using LLMs, and train model to adapt to those personas. In contrast, (Chen et al., 2024b) adopts a plurality-based approach, using ideal point and mixture modeling to learn a common latent preference space that generalizes to new users, Finally, (Poddar et al., 2024) also learns user latent factors, but instead opts to use user-conditional reward models to achieve personalization.

Limitations. While alleviating the need for maintaining user-specific parameters, this class of methods often depends on user such as personas and verbal preferences, which are unavailable in standard RLHF settings. Further, collecting datasets with diverse user personas itself is poses a challenge, which we discuss extensively in Section 6.

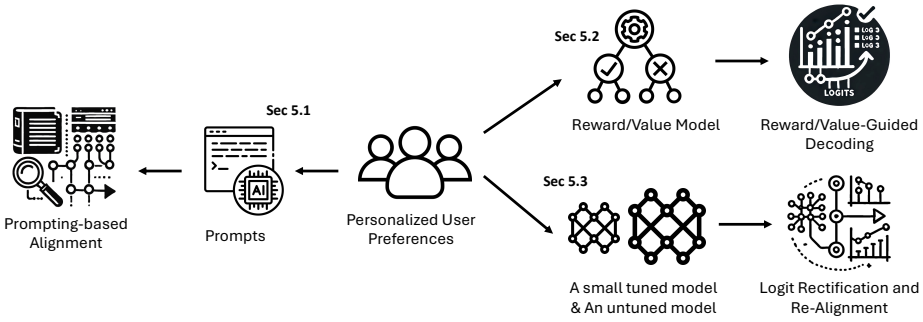

Figure 4: Personalized language model alignment during test time.

5 Test-time Personalized Alignment

The growing need for language models that can adapt on the fly to diverse user preferences—while keeping compute costs manageable and simplifying deployment—has spurred a variety of test-time alignment techniques. These methods adjust the decoding process, without altering the underlying model parameters, to steer outputs toward desired behaviors and make debugging easier. We broadly categorize these approaches into three groups: prompting and context optimization, reward- and value-guided decoding methods, and rectification-based decoding and correction.

5.1 Prompting-based Alignment Methods

Motivation. Prompting-based approaches modify the input context—via in-context examples, retrieval, or prompt rewriting—to elicit behavior that mirrors specific user preferences. This training-free strategy is particularly attractive for personalization, as it can quickly adapt to new or evolving user profiles (see Figure 4).

Comparative Analysis. A central challenge for prompting-based methods is achieving personalized alignment without the need for full-scale fine-tuning. (Lin et al., 2023) and (Li et al., 2023c) propose methods that harness a small number of stylistic examples or self-correction mechanisms to nudge LLMs toward outputs that align with individual user styles. Similarly, (Xu et al., 2023) dynamically incorporate external memories to retrieve rules tailored to diverse social norms, while (Salemi et al., 2024) and (Mysore et al., 2023a) further personalize responses by retrieving user-specific items to augment the prompt. In parallel, prompt optimization techniques by (Kim & Yang, 2024), (Cheng et al., 2023a), and (Li et al., 2024b) iteratively refine prompts based on misaligned outputs, thereby progressively incorporating user feedback.

Limitations. Despite their promise for personalization, these methods are limited by the fixed context length and increased computational overhead when handling complex, dynamic user profiles. Additionally, the reliance on relatively static representations of user preferences can hinder real-time adaptation to rapidly changing individual needs.

5.2 Reward and Value-Guided Decoding Methods

Motivation. Reward and value-guided decoding methods integrate personalized alignment objectives directly into the LLM decoding by adjusting token probabilities based on reward signals or value functions. This approach supports fine-grained, token-level personalization and enables LLMs to adapt their outputs on the fly to meet individual user preferences.

Comparative Analysis. One of the major challenges is balancing high-reward personalized output with maintaining natural language fluency. (Khanov et al., 2024) and (Deng & Raffel, 2023) propose frameworks that incorporate reward signals into the decoding process, which can be tuned to reflect personalized quality metrics. Building on this, (Li et al., 2024a) and (Huang et al., 2024a) introduce strategies that break the generation process

into segments, allowing for iterative refinement toward personalized and fluent outputs. Complementing these reward-guided methods, (Xu et al., 2024) and (Chen et al., 2024e) focus on scalable, test-time alignment by employing autoregressive reward models and personalized reward modeling, respectively. In parallel, value-guided techniques offer another path to personalization: (Mudgal et al., 2023) and (Liu et al., 2024b) integrate auxiliary value functions during decoding, whereas (Liu et al., 2024d) and (Han et al., 2024) demonstrate that combining implicit and explicit value guidance can further enhance the model's responsiveness with respect to users' personalized preferences and values.

Limitations. A key limitation of reward and value-guided methods is their reliance on pre-trained reward or value models, which introduces challenges such as limited transparency and sensitivity to adversarial inputs. Additionally, the need for real-time evaluation of reward signals and value functions often results in significant computational overhead, making it challenging to scale these methods for high-throughput, low-latency applications.

5.3 Logit Rectification and Re-Alignment Approaches

Motivation. Logit rectification approaches modify the internal decision process of LLMs by integrating corrective signals—often from smaller, fine-tuned models or auxiliary modules—directly into the decoding stage. This strategy enables personalized re-alignment without retraining the entire model, thereby offering a lightweight path to tailor outputs to individual user needs.

Comparative Analysis. A key challenge in this category is to harness the complementary strengths of large pretrained models and smaller, aligned models. (Mitchell et al., 2023) introduce emulated fine-tuning, which decouples the contributions of pre-training and fine-tuning by combining the knowledge of a large model with the behavioral adjustments of a small model; a special case, LM up-scaling, demonstrates that ensembling can emulate the effects of full fine-tuning without additional training. In a similar vein, (Liu et al., 2024a) propose proxy-tuning, a lightweight algorithm that steers the output distribution of a large black-box model using the difference between the predictions of a tuned small model and its untuned counterpart. Addressing the need for multi-objective personalization, (Shi et al., 2024) develop a method that computes a linear combination of predictions from several base models, enabling flexible adjustment of competing objectives, while (Liu et al., 2024c) propose decoding-time re-alignment to dynamically control the degree of alignment without retraining. Complementing these strategies, (Ji et al., 2024) offers a model-agnostic correction approach that learns residuals to refine outputs, and (Yang et al., 2024) extends this idea with a generalizable framework that supports multi-objective and personalized alignment by adjusting target objectives via prompt updates. Finally, (Zhang et al., 2025) frame re-alignment as an online per-token optimization problem with a closed-form solution, achieving real-time adaptation to diverse and evolving user preferences.

Limitations. Rectification-based approaches are highly dependent on the quality and consistency of the auxiliary correction signals. Their effectiveness can be limited if the small, tuned models do not capture the full complexity of user-specific nuances, which may lead to a mismatch between the corrective adjustments and the large model's inherent output tendencies. Additionally, ensuring robust performance across a wide range of personalization scenarios without introducing extra latency remains a practical challenge.

6 Dataset & Evaluation

Datasets. Early attempts on constructing datasets for personalized preference alignment typically rely on data synthesization, with some recent attempts in collecting real-world data for preference alignment. We provide an overview of relevant benchmarks and evaluations in Table 1. For example, (Cheng et al., 2023b) introduce the Domain-Specific Preference (DSP) dataset by augmenting the Alpaca instruction corpus (Taori et al., 2023) with answers targeting different domain (e.g., Academia or Entertainment). (Jang et al., 2023) introduce Personalized Soups (P-SOUP), where Alpaca instruction prompts (Taori et al., 2023) are answered by an LLM in contrasting styles along three dimensions (expertise level, ver-

Reference	Task/Data	Metric (Dimension)	Notes
DSP (Cheng et al., 2023b)	Prompt selection	PPA (Domain: Acad., Bus., Ent., Lit. & Art)	Tailored prompts
P-SOUP (Jang et al., 2023)	Pairwise feedback	PWR (Expertise, Info., Friendliness)	GPT-4 simulated
MULTIFACETED (Lee et al., 2024b)	Preference alignment	AAS (Human pref.)	Diverse preferences
HH-RLHF (Bai et al., 2022)	RLHF dataset	CRS (Helpfulness, Harmlessness)	Personalized dimensions
HelpSteer2 (Wang et al., 2024d)	RLHF feedback	MFS (Helpfulness, Harmlessness, Humor)	Humor-enhanced
PRISM (Kirk et al., 2024b)	Demographic alignment	AFI (Fairness across demographics)	Participatory alignment
LaMP (Salemi et al., 2024)	Personalized generation	POQ (User-profile retrieval)	Retrieval-augmented
LongLaMP (Kumar et al., 2024a)	Personalized long-text generation	POQ (User-profile retrieval)	Retrieval-augmented
PersonalLLM (Zollo et al., 2024)	Personalization benchmark	PAS (Individual prefs.)	Beyond uniform alignment
ALOE (Wu et al., 2024c)	Multi-turn dialogues	PCS (Persona consistency)	Persona-specific dataset
PersoBench (Afzoon et al., 2024)	Persona-aware dialogue	PAA (Persona awareness)	Zero-shot evaluation

Table 1: Compressed overview of datasets and benchmarks for personalized alignment and generation.

bosity, and tone) and ranked based on predefined style preferences. On a much larger scale, (Lee et al., 2024b) constructed the Multifaceted Collection dataset, containing about 65k instructions, each paired with three LLM-generated responses that vary along thousands of dimensions. Finally, (Zollo et al., 2024) present PersonalLLM, an open benchmark that generate preference rankings from a set of simulated synthetic "users". In parallel, other recent works have assembled personalized human preference datasets to study personalization with real users . The PRISM Alignment dataset (Kirk et al., 2024b) is a notable example, recording 8,011 live chat interactions between 1,500 users across 75 countries, with each user's persona profile and preference feedbacks (ratings), providing a complementary setting for evaluating personalized preference alignment.

Evaluation in Personalized Alignment. Evaluating personalization in LLM alignment requires measuring how well a model's output matches the particular preferences of a user. When the evaluation dataset includes response variations across known dimensions (e.g., style, tone), evaluation often uses multi-dimensional scoring or pairwise comparisons. For example, (Jang et al., 2023) assess their proposed methods by scoring responses along each dimension separately, but assumes different users place different value on the set of dimensions. Another popular evaluation approach is to employ LLM-as-a-judge Zheng et al. (2023), but prompt the judge LLMs with pre-defined user personas (Wu et al., 2024c; Lee et al., 2024b). This evaluation formulation leads to the creation of specialized evaluation models such as PerSE (Wang et al., 2024b), a 13B Llama-2 based evaluator fine-tuned to judge alignment with personal profiles. Overall, a unique open challenge in personalized alignment evaluation is realistically simulating personalized preference judgment (Zheng et al., 2023; Wu et al., 2024c; Lee et al., 2024b; Wang et al., 2024b), calling for further investigations.

Limitations. Despite progress on model evaluations catered for personalized preference alignment, most assessments rely heavily on rule-based metrics or persona-based LLM-as-a-judge Zheng et al. (2023), yielding simulations of user satisfaction. Since these heuristics and personas for evaluations are use-case specific, there are currently no unified evaluation across studies, posing a challenge for systematic evaluation and progress tracking. As the research community moves toward widely accepted multidimensional preference benchmarks and the development of publicly available evaluators, more consistent and comparable metrics are expected to emerge.

7 On Personalized User Modeling and Personalized Preference Alignment

Why this section? As discussed previously in Section 2, personalized preference alignment is a subset of personalization research for LLMs. Specifically, there is another emerging research area that focuses on LLM-based user modeling Tan & Jiang (2023), building LLM-based *simulations* for individual users, which most commonly manifest as directly predicting users' responses. While prior surveys do not differentiate between personalized user modeling and preference alignment (Zhang et al., 2024b; Wu et al., 2024a; Tseng et al., 2024), we note that these are two distinct and complementary research directions. To this end, this section serves two purposes. First, we clearly distinguish between the two

research directions currently studied under LLMs and personalization, making it easier for future researchers to sift through relevant literature. Second, we discuss various ways user modeling can enable better-personalized preference alignment.

Personalized preference alignment research benefits from personalized user modeling. Due to the difficulty in collecting large-scale feedbacks from real users, current research for personalized preference alignment relies heavily on user simulation. For example, various recently proposed datasets rely on persona-grounded simulation with LLMs to build benchmarks (Cheng et al., 2023b; Jang et al., 2023; Lee et al., 2024b; Zollo et al., 2024) or evaluations (Zheng et al., 2023; Wu et al., 2024c; Lee et al., 2024b; Wang et al., 2024b) for personalized preference methods. To this end, better simulation of diverse, realistic user behaviors naturally improves the development of better personalized alignment methods.

Modeling individual user (sometimes) enables preference-free personalized alignment. Meanwhile, by modeling individual users, system deployers can still increase user satisfaction, even without directly modeling user preferences. For example, in areas such as personalized review prediction (Xie et al., 2023; Ni et al., 2019) and recommender system (Wu et al., 2024b), user behavior happens to strongly correlate with user preferences, and thus better prediction of user behavior directly improves preference alignment. Similarly, related research directions such as LLM-based chit-chat dialogue systems with personalized user memory (e.g., (Yuan et al., 2025)) also better caters to the preference of individual users by remembering personal facts, even when there are no explicit modeling of user preferences. Finally, simulating specific desired personas such as teachers (Wang et al., 2024c), therapists (Stade et al., 2024), and travel-planners (Chen et al., 2024a) also enables LLMs to better cater to the corresponding user groups, such as students, patients, and travelers.

8 Future Works and Emerging Directions

Online and Continuous Personalized Alignment While existing work on personalized preference alignment primarily explores the setting of learning user preference from offline data or given explicitly stated user preference, another complementary setting is personalized LLM alignment in an online setting (Chen et al., 2024f). Additionally, given prior success in adjacent research (such as user modeling) on continuous personalization over multiple dialogue sessions (Li et al., 2024c; Zhang et al., 2023; Zhong et al., 2024; Qian et al., 2024), personalized preference alignment in a multi-session dialogue setting is a natural extension, which typically models turn-wise or user-provided preference feedback. Nevertheless, recent work Zhao et al. (2025) show large language models frequently fail to recall user preference in continuous personalization setting, calling for more elaborate solution for personalization of LLMs over time.

Addressing long and complex user-generated value statements As discussed in prior sections, personalized preference alignment in LLMs frequently relies on instruction following ability of LLMs as building blocks for alignment methods, both at training time and inference time. However, recent works show long and complex instruction-following is still an open challenge (Wu et al., 2024d; Gavin et al., 2024). Given the prevalence of personalized alignment methods that rely on explicit verbal preference statements (Section 4.2 and Section 5.1), it is still unclear whether existing methods can support complex and long user value statements. To this end, developing benchmarks and methods to further research the instruction-following ability of LLMs on complex user preference value statements can help LLM-based dialogue systems better handle rich, multifaceted user preferences.

9 Conclusions

In this survey, we perform a comprehensive analysis of existing methods, datasets, and benchmarks for personalized preference alignment in LLMs and LLM-based dialogue systems. We discuss various classes of methods and their advantages and drawbacks, covering both training and inference-time, as well as user-knowledge-based personalized preference alignment methods. We also discuss limitations and future directions for LLMs to cater to diverse and individualistic preferences.

References

Saleh Afzoon, Usman Naseem, Amin Beheshti, and Zahra Jamali. Persobench: Benchmarking personalized response generation in large language models. *arXiv preprint arXiv:2410.03198*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Nishant Balepur, Matthew Shu, Alexander Hoyle, Alison Robey, Shi Feng, Seraphina Goldfarb-Tarrant, and Jordan Lee Boyd-Graber. A SMART mnemonic sounds like "glue tonic": Mixing LLMs with student feedback to make mnemonic learning stick. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14202–14225, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.786. URL https://aclanthology.org/2024.emnlp-main.786/.

Nishant Balepur, Vishakh Padmakumar, Fumeng Yang, Shi Feng, Rachel Rudinger, and Jordan Lee Boyd-Graber. Whose boat does it float? improving personalization in preference tuning via inferred user personas. *arXiv preprint arXiv:2501.11549*, 2025.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization. *arXiv preprint arXiv:2406.00045*, 2024.

Aili Chen, Xuyang Ge, Ziquan Fu, Yanghua Xiao, and Jiangjie Chen. Travelagent: An ai assistant for personalized travel planning, 2024a. URL https://arxiv.org/abs/2409.08069.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint arXiv:2406.08469*, 2024b.

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, 2024c.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. The oscars of ai theater: A survey on role-playing with language models, 2024d. URL https://arxiv.org/abs/2407.11484.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. Pad: Personalized alignment at decoding-time. *arXiv preprint arXiv:2410.04070*, 2024e.

Zekai Chen, Weeden Daniel, Po yu Chen, and Francois Buet-Golfouse. Online personalizing white-box llms generation with neural bandits, 2024f. URL https://arxiv.org/abs/2404.16115.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. Black-box prompt optimization: Aligning large language models without model training. *arXiv preprint arXiv:2311.04155*, 2023a.

Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. Everyone deserves a reward: Learning customized human preferences. *arXiv preprint arXiv:2309.03126*, 2023b.

Haikang Deng and Colin Raffel. Reward-augmented decoding: Efficient controlled text generation with a unidirectional reward model. *arXiv preprint arXiv:2310.09520*, 2023.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W Black, Alexander Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. The second conversational intelligence challenge (convai2), 2019. URL https://arxiv.org/abs/1902.00098.

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. Towards measuring the representation of subjective global opinions in language models, 2024. URL https://arxiv.org/abs/2306.16388.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

Lucie Flek. Returning the N to NLP: Towards contextually personalized classification models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7828–7838, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.700. URL https://aclanthology.org/2020.acl-main.700/.

Songyang Gao, Qiming Ge, Wei Shen, Shihan Dou, Junjie Ye, Xiao Wang, Rui Zheng, Yicheng Zou, Zhi Chen, Hang Yan, et al. Linear alignment: A closed-form solution for aligning human preferences without tuning and feedback. *arXiv preprint arXiv:2401.11458*, 2024.

Shawn Gavin, Tuney Zheng, Jiaheng Liu, Quehry Que, Noah Wang, Jian Yang, Chenchen Zhang, Wenhao Huang, Wenhu Chen, and Ge Zhang. Longins: A challenging long-context instruction-based exam for llms, 2024. URL https://arxiv.org/abs/2406.17588.

Anmol Goel, Yaxi Hu, Iryna Gurevych, and Amartya Sanyal. Differentially private steering for large language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lLkgj7FEtZ.

Seungwook Han, Idan Shenfeld, Akash Srivastava, Yoon Kim, and Pulkit Agrawal. Value augmented sampling for language model alignment and personalization. *arXiv preprint arXiv:2405.06639*, 2024.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Controllable text generation. *CoRR*, abs/1703.00955, 2017. URL http://arxiv.org/abs/1703.00955.

James Y Huang, Sailik Sengupta, Daniele Bonadiman, Yi-an Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchhoff, and Dan Roth. Deal: Decoding-time alignment for large language models. *arXiv preprint arXiv:2402.06147*, 2024a.

Qiushi Huang, Xubo Liu, Tom Ko, Bo Wu, Wenwu Wang, Yu Zhang, and Lilian Tang. Selective prompting tuning for personalized conversations with llms. *arXiv preprint arXiv:2406.18187*, 2024b.

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.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023. URL https://arxiv.org/abs/2307.04657.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Alex Qiu, Juntao Dai, and Yaodong Yang. Aligner: Efficient alignment by learning to correct. *Advances in Neural Information Processing Systems*, 37:90853–90890, 2024.

Feibo Jiang, Li Dong, Siwei Tu, Yubo Peng, Kezhi Wang, Kun Yang, Cunhua Pan, and Dusit Niyato. Personalized wireless federated learning for large language models, 2024. URL https://arxiv.org/abs/2404.13238.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search. *arXiv preprint arXiv:2402.01694*, 2024.

Jaehyung Kim and Yiming Yang. Few-shot personalization of llms with mis-aligned responses. *arXiv preprint arXiv:2406.18678*, 2024.

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

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024b.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. URL https://arxiv.org/abs/2305.01569.

Lingkai Kong, Haorui Wang, Wenhao Mu, Yuanqi Du, Yuchen Zhuang, Yifei Zhou, Yue Song, Rongzhi Zhang, Kai Wang, and Chao Zhang. Aligning large language models with representation editing: A control perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=yTTomSJsSW.

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

Sachin Kumar, Chan Young Park, Yulia Tsvetkov, Noah A Smith, and Hannaneh Hajishirzi. Compo: Community preferences for language model personalization. *arXiv preprint arXiv:2410.16027*, 2024b.

Gihun Lee, Minchan Jeong, Yujin Kim, Hojung Jung, Jaehoon Oh, Sangmook Kim, and Se-Young Yun. Bapo: Base-anchored preference optimization for overcoming forgetting in large language models personalization, 2024a. URL https://arxiv.org/abs/2407.00693.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. Aligning to thousands of preferences via system message generalization. *arXiv preprint arXiv:2405.17977*, 2024b.

Bolian Li, Yifan Wang, Ananth Grama, and Ruqi Zhang. Cascade reward sampling for efficient decoding-time alignment. *arXiv preprint arXiv:2406.16306*, 2024a.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. Teach llms to personalize – an approach inspired by writing education, 2023a. URL https://arxiv.org/abs/2308.07968.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM Web Conference 2024*, pp. 3367–3378, 2024b.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. Hello again! llm-powered personalized agent for long-term dialogue. *arXiv preprint arXiv:2406.05925*, 2024c.

Xinyu Li, Zachary C Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*, 2024d.

Yang Li, Kangbo Liu, Ranjan Satapathy, Suhang Wang, and Erik Cambria. Recent developments in recommender systems: A survey, 2023b. URL https://arxiv.org/abs/2306.12680.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*, 2024e.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. Rain: Your language models can align themselves without finetuning. *arXiv preprint arXiv:2309.07124*, 2023c.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A Smith. Tuning language models by proxy. *arXiv preprint arXiv:2401.08565*, 2024a.

Jiacheng Liu, Andrew Cohen, Ramakanth Pasunuru, Yejin Choi, Hannaneh Hajishirzi, and Asli Celikyilmaz. Don't throw away your value model! generating more preferable text with value-guided monte-carlo tree search decoding. In *First Conference on Language Modeling*, 2024b.

Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. Decoding-time realignment of language models. *arXiv preprint arXiv:2402.02992*, 2024c.

Zhixuan Liu, Zhanhui Zhou, Yuanfu Wang, Chao Yang, and Yu Qiao. Inference-time language model alignment via integrated value guidance. *arXiv preprint arXiv:2409.17819*, 2024d.

Xiangwu Meng, Yulu Du, Yujie Zhang, and Xiaofeng Han. A survey of context-aware recommender systems: From an evaluation perspective. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):6575–6594, 2023. doi: 10.1109/TKDE.2022.3187434.

Eric Mitchell, Rafael Rafailov, Archit Sharma, Chelsea Finn, and Christopher D Manning. An emulator for fine-tuning large language models using small language models. *arXiv preprint arXiv:2310.12962*, 2023.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. From individual to society: A survey on social simulation driven by large language model-based agents, 2024. URL https://arxiv.org/abs/2412.03563.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. *arXiv preprint arXiv:2310.17022*, 2023.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. *arXiv preprint arXiv:2311.09180*, 2023a.

Sheshera Mysore, Andrew McCallum, and Hamed Zamani. Large language model augmented narrative driven recommendations, 2023b. URL https://arxiv.org/abs/2306.02250.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1018. URL https://aclanthology.org/D19-1018/.

Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman E Ozdaglar. Rlhf from heterogeneous feedback via personalization and preference aggregation. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*, 2024a.

Chanwoo Park, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. Principled rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024b.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847/.

Silviu Pitis, Ziang Xiao, Nicolas Le Roux, and Alessandro Sordoni. Improving context-aware preference modeling for language models. *arXiv preprint arXiv:2407.14916*, 2024.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *arXiv preprint arXiv:2408.10075*, 2024.

Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *arXiv preprint arXiv:2409.05591*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect?, 2023. URL https://arxiv.org/abs/2303.17548.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey, 2023. URL https://arxiv.org/abs/2309.15025.

Ruizhe Shi, Yifang Chen, Yushi Hu, ALisa Liu, Noah Smith, Hannaneh Hajishirzi, and Simon Du. Decoding-time language model alignment with multiple objectives. *arXiv preprint arXiv:2406.18853*, 2024.

Feifan Song, Yuxuan Fan, Xin Zhang, Peiyi Wang, and Houfeng Wang. Instantly learning preference alignment via in-context DPO. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 161–178, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.8. URL https://aclanthology.org/2025.naacl-long.8/.

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. A roadmap to pluralistic alignment, 2024. URL https://arxiv.org/abs/2402.05070.

Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle H. Ungar, Cody L. Boland, H. A. Schwartz, David Bryce Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3, 2024. URL https://api.semanticscholar.org/CorpusID:268881423.

Zhaoxuan Tan and Meng Jiang. User modeling in the era of large language models: Current research and future directions. *arXiv preprint arXiv:2312.11518*, 2023.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. Personalized pieces: Efficient personalized large language models through collaborative efforts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6459–6475, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.371. URL https://aclanthology.org/2024.emnlp-main.371/.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. *ArXiv*, abs/2402.04401, 2024b. URL https://api.semanticscholar.org/CorpusID:267523232.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*, 2024.

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

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. Learning personalized alignment for evaluating open-ended text generation, 2024b. URL https://arxiv.org/abs/2310.03304.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. Large language models for education: A survey and outlook, 2024c. URL https://arxiv.org/abs/2403.18105.

Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Trans. Inf. Syst.*, 41(3), February 2023. ISSN 1046-8188. doi: 10.1145/3547333. URL https://doi.org/10.1145/3547333.

Zekun Moore Wang, Shenzhi Wang, King Zhu, Jiaheng Liu, Ke Xu, Jie Fu, Wangchunshu Zhou, and Wenhao Huang. PopAlign: Diversifying contrasting patterns for a more comprehensive alignment. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28893–28921, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1403. URL https://aclanthology.org/2025.acl-long.1403/.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024d.

Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A. Rossi, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Namyong Park, Sungchul Kim, Huanrui Yang, Subrata Mitra, Zhengmian Hu, Nedim Lipka, Dang Nguyen, Yue Zhao, Jiebo Luo, and Julian McAuley. Personalized multimodal large language models: A survey, 2024a. URL https://arxiv.org/abs/2412.02142.

Junkang Wu, Kexin Huang, Xue Wang, Jinyang Gao, Bolin Ding, Jiancan Wu, Xiangnan He, and Xiang Wang. Repo: Relu-based preference optimization, 2025. URL https://arxiv.org/abs/2503.07426.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. A survey on large language models for recommendation, 2024b. URL https://arxiv.org/abs/2305.19860.

Shujin Wu, May Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. Aligning llms with individual preferences via interaction. *arXiv preprint arXiv:2410.03642*, 2024c.

Xiaodong Wu, Minhao Wang, Yichen Liu, Xiaoming Shi, He Yan, Xiangju Lu, Junmin Zhu, and Wei Zhang. Lifbench: Evaluating the instruction following performance and stability of large language models in long-context scenarios, 2024d. URL https://arxiv.org/abs/2411.07037.

Zhouhang Xie, Sameer Singh, Julian McAuley, and Bodhisattwa Prasad Majumder. Factual and informative review generation for explainable recommendation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i11.26618. URL https://doi.org/10.1609/aaai.v37i11.26618.

Chunpu Xu, Steffi Chern, Ethan Chern, Ge Zhang, Zekun Wang, Ruibo Liu, Jing Li, Jie Fu, and Pengfei Liu. Align on the fly: Adapting chatbot behavior to established norms. *arXiv preprint arXiv:2312.15907*, 2023.

Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. *arXiv preprint arXiv:2410.08193*, 2024.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Tianlin Zhang, and Sophia Ananiadou. Metaaligner: Towards generalizable multi-objective alignment of language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ruifeng Yuan, Shichao Sun, Yongqi Li, Zili Wang, Ziqiang Cao, and Wenjie Li. Personalized large language model assistant with evolving conditional memory. In Owen Rambow,

Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3764–3777, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.254/.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. Llm-based medical assistant personalization with short-and long-term memory coordination. *arXiv preprint arXiv:2309.11696*, 2023.

Tong Zhang, Yong Liu, Boyang Li, Zhiwei Zeng, Pengwei Wang, Yuan You, Chunyan Miao, and Lizhen Cui. History-aware hierarchical transformer for multi-session open-domain dialogue system. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3395–3407, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.247. URL https://aclanthology.org/2022.findings-emnlp.247/.

Yicheng Zhang, Zhen Qin, Zhaomin Wu, and Shuiguang Deng. Personalized federated fine-tuning for llms via data-driven heterogeneous model architectures. *arXiv preprint arXiv:2411.19128*, 2024a.

Zhaowei Zhang, Fengshuo Bai, Qizhi Chen, Chengdong Ma, Mingzhi Wang, Haoran Sun, Zilong Zheng, and Yaodong Yang. Amulet: Realignment during test time for personalized preference adaptation of LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=f9w89OY2cp.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen Ahmed, and Yu Wang. Personalization of large language models: A survey, 2024b. URL https://arxiv.org/abs/2411.00027.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QWunLKbBGF.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2024. URL https://arxiv.org/abs/2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024.

Jin Peng Zhou, Katie Z Luo, Jingwen Gu, Jason Yuan, Kilian Q Weinberger, and Wen Sun. Orchestrating llms with different personalizations. *arXiv preprint arXiv:2407.04181*, 2024.

Hanlin Zhu, Andrew Cohen, Danqing Wang, Kevin Yang, Xiaomeng Yang, Jiantao Jiao, and Yuandong Tian. End-to-end story plot generator, 2023. URL https://arxiv.org/abs/2310.08796.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. Hydra: Model factorization framework for black-box llm personalization. *ArXiv*, abs/2406.02888, 2024. URL https://api.semanticscholar.org/CorpusID:270257981.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. Personalllm: Tailoring llms to individual preferences. *arXiv preprint arXiv:2409.20296*, 2024.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

A On Risk of Pluralistic Preference Alignment

We note that as discussed in prior works (Kirk et al., 2024a), naively personalizing LLMs to individualized preference can raise other issues such as amplifying bias or inducing polarization. To this end, similar to adjacent domain in personalization such as recommender system Wang et al. (2023), ensuring safety and fairness in pluralistic alignment remain an continuously important and open challenge. Meanwhile, pluralistic alignment generally requires collecting diverse preference signals from users, calling for privacy-aware method such as differentially private LLM alignment Goel et al. (2025).

B On Datasets and Benchmarks for Pluralistic Alignment

Unlike recent work on general preference alignment for LLMs—which spans training and inference time methods, *inter alia* (Rafailov et al., 2023; Wang et al., 2025; Wu et al., 2025; Kong et al., 2024; Song et al., 2025), a unique ongoing challenge in the field is the lack of commonly acknowledged benchmarks and evaluation methods. To this end, we show a few recent works and their evaluation datasets in Table 2, and hope this makes accessing relevant dataset easier and facilitate comparisons on shared benchmarks.

Method	Dataset
RAG/PAG (Salemi et al., 2024)	LAMP (Salemi et al., 2024)
OPPU (Tan et al., 2024b)	LAMP (Salemi et al., 2024)
Personalized-PCs (Tan et al., 2024a)	LAMP (Salemi et al., 2024)
HyDRA (Zhuang et al., 2024)	LAMP (Salemi et al., 2024)
P-DPO (Li et al., 2024d)	Prism Alignment Kirk et al. (2024b)
BiPO (Cao et al., 2024)	AI Persona (Perez et al., 2023), TruthfulQA (Lin et al., 2022), ADVBench (Zou et al., 2023)
SPT (Huang et al., 2024b)	ConvAI2 (Dinan et al., 2019)
System Message Gen. (Lee et al., 2024b)	Multi-faceted Benchmark and other helpfulness/harmlessness benchmarks (Lee et al., 2024b)
PerSE (Wang et al., 2024a)	Per-MPST (Wang et al., 2024a), Per-DOC (Zhu et al., 2023)
CARM (Pitis et al., 2024)	RPR Dataset (Pitis et al., 2024)
PIPT (Balepur et al., 2025)	Beavertails/SHP (Ji et al., 2023), Stanford Preferences (Ethayarajh et al., 2022), HH-RLHF (Bai et al., 2022), Mnemonic (Balepur et al., 2024)
PAL (Chen et al., 2024b)	Anthropic Persona Perez et al. (2023), Pick-a-Pic (Kirstain et al., 2023)
PEARL (Mysore et al., 2023a)	WORKSM, AITA, see PEARL paper
FERMI (Kim & Yang, 2024)	OpinionQA (Santurkar et al., 2023), GlobalOpinionQA Durmus et al. (2024), LAMP Salemi et al. (2024)
Prompt Rewrite (Li et al., 2024b)	FtPersLlm (Li et al., 2023a)
Linear Align. (Gao et al., 2024)	HH-RLHF Bai et al. (2022)

Table 2: Inexhaustive List of Recent Personalized/Pluralistic Alignment Methods and Datasets Used