A Survey of Personalized Large Language Models: Progress and Future Directions

Anonymous EMNLP submission

Abstract

Large Language Models (LLMs) excel in handling general knowledge tasks, yet they struggle with user-specific personalization, such as understanding individual emotions, writing styles, and preferences. Personalized Large Language Models (PLLMs) tackle these challenges by leveraging individual user data, such as user profiles, historical dialogues, content, and interactions, to deliver responses that are contextually relevant and tailored to each user's specific needs. This is a highly valuable research topic, as PLLMs can significantly enhance user satisfaction and have broad applications in conversational agents, recommendation systems, emotion recognition, medical assistants, and more. This survey reviews recent advancements in PLLMs from three technical perspectives: prompting for personalized context (input level), finetuning for personalized adapters (model level), and alignment for personalized preferences (objective level). To provide deeper insights, we also discuss current limitations and outline several promising directions for future research. The papers are organized in an anonymous Github Repo.

1 Introduction

In recent years, substantial progress has been made in Large Language Models (LLMs) such as GPT, PaLM, LLaMA, DeepSeek, and their variants (Zhao et al., 2023). These models have demonstrated remarkable versatility, achieving state-of-the-art performance across various natural language processing (NLP) tasks, including question answering, logical reasoning, and machine translation (Chang et al., 2024; Hu et al., 2024; Zhang et al., 2024f,e; Zhu et al., 2024; Wang et al., 2023a, 2024a), with minimal task-specific adaptation.

The Necessity of Personalized LLMs (PLLMs) While LLMs excel in general knowledge and multi-domain reasoning, their lack of personalization creates challenges in situations where user-specific understanding is crucial. For instance, conversational agents need to adapt to a user's preferred tone and incorporate past interactions to deliver relevant, personalized responses. As LLMs evolve, integrating personalization capabilities has become a promising direction for advancing human-AI interaction across diverse domains such as education, healthcare, and finance (Hu et al., 2024; Zhang et al., 2024f,e; Zhu et al., 2024; Wang et al., 2023a, 2024a). Despite its promise, personalizing LLMs presents several challenges. These include efficiently representing and integrating diverse user data, addressing privacy concerns, managing long-term user memories, etc (Salemi et al., 2023). Moreover, achieving personalization often requires balancing accuracy and efficiency while addressing biases and maintaining fairness in the outputs.

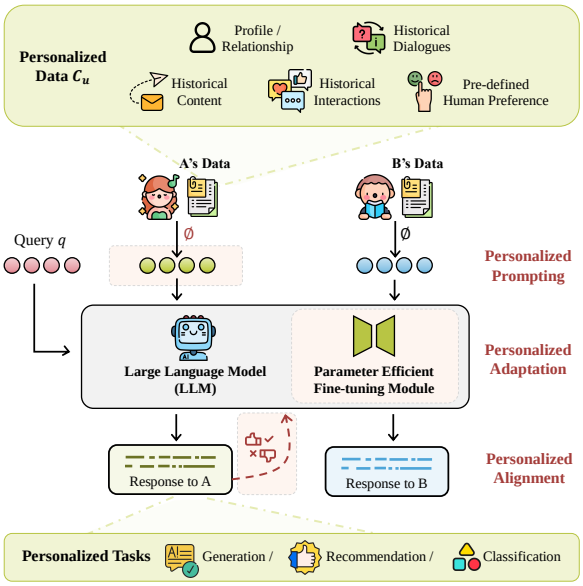

Figure 1: Illustration of PLLM techniques for generating personalized responses through three levels: prompting, adaptation, and alignment.

Contributions Despite growing interest, the field of PLLMs lacks a systematic review that consolidates recent advancements. This survey aims to bridge the gap by systematically organizing exist-

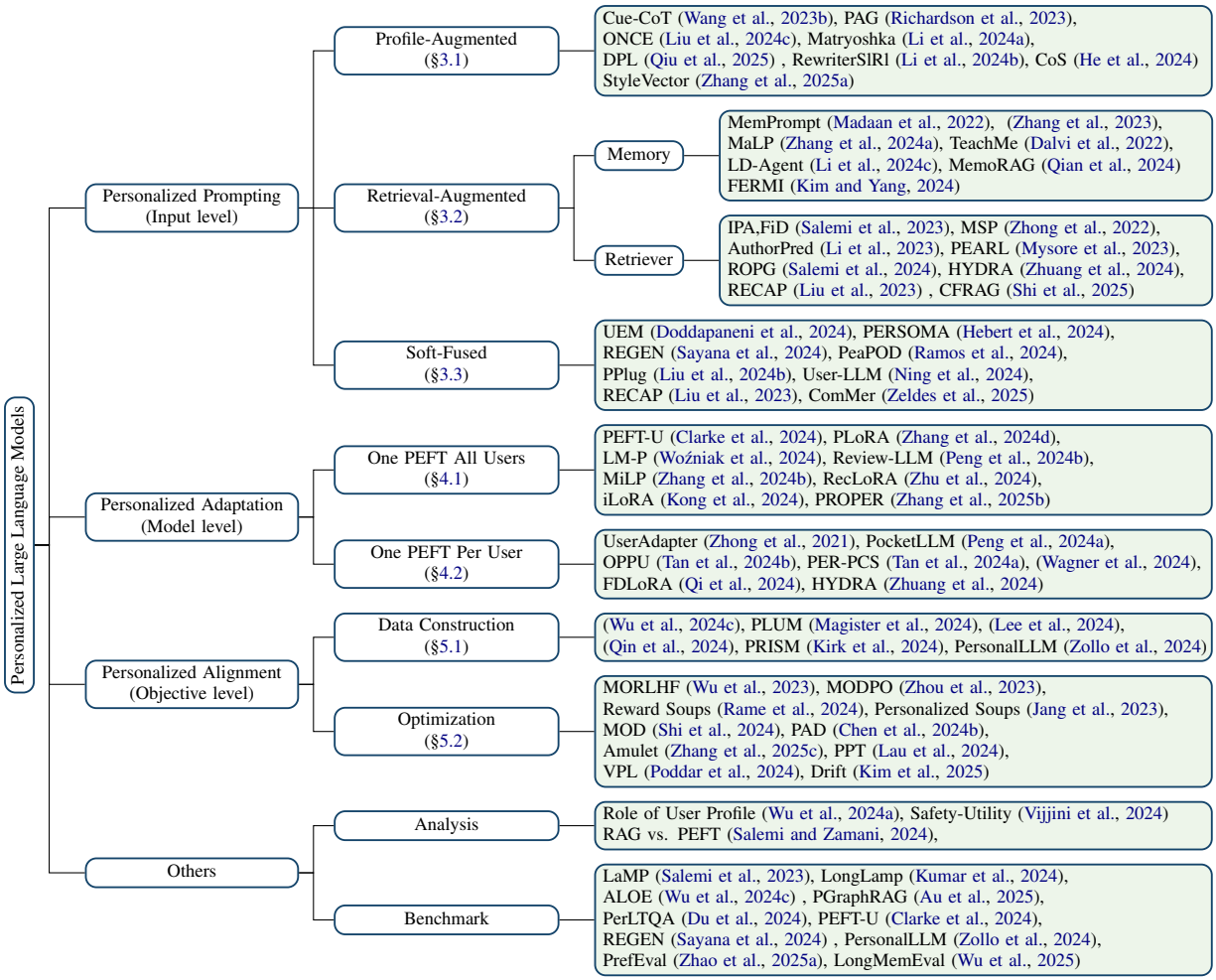

Figure 2: A taxonomy of PLLMs with representative examples.

ing research on PLLMs and offering insights into their methodologies and future directions. The contributions of this survey can be summarized as follows: *(1) A structured taxonomy:* We propose a comprehensive taxonomy, providing a technical perspective on the existing approaches to building PLLMs. *(2) A comprehensive review:* We systematically review state-of-the-art methods for PLLMs, analyzing fine-grained differences among the methods. *(3) Future directions:* We highlight current limitations, such as data privacy and bias, and outline promising avenues for future research, including multimodal personalization, edge computing, lifelong updating, trustworthiness, etc.

2 Preliminary

2.1 Large Language Models

Large Language Models (LLMs) generally refer to models that utilize the Transformer architecture and are equipped with billions of parameters trained on trillions of text tokens. These models have demonstrated substantial improvements in a myriad of tasks related to natural language understanding and generation, increasingly proving beneficial in assisting human activities. In this work, we mainly focus on autoregressive LLMs, which are based on two main architectures: decoder-only models and encoder-decoder models. Encoder-decoder models such as Flan-T5 (Chung et al., 2022) and ChatGLM (Zeng et al., 2022) analyze input through the encoder for semantic representations, making them effective in language understanding in addition to generation. Decoder-only LLMs focus on left-to-right generation by predicting the next token in a sequence, with numerous instances (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023; Guo et al., 2025) under this paradigm achieving breakthroughs in advanced capabilities.

However, these models are typically pre-trained on general-purpose data and **lack an understanding of specific user information**. As a result, they are unable to generate responses tailored to a user's

unique tastes and expectations, limiting their effectiveness in personalized applications where user-specific adaptation is critical.

2.2 Problem Statement

Personalized Large Language Models (PLLMs) generate responses that align with the user's style and expectations, offering diverse answers to the same query for different users (Clarke et al., 2024). A PLLM is defined as an LLM that generates responses conditioned not only on an input query q, but also on a user u's personalized data $\mathcal{C}_u$. It aims to predict the most probable response sequence y given a query q and the personalized context $\mathcal{C}_u$, such that: $y = \text{argmax}_y P(y \mid q, \mathcal{C}_u)$. The personalized data $\mathcal{C}_u$ may encapsulate information about the user's preferences, history, context, and other user-specific attributes. These can include profile/relationship, historical dialogues, historical content, and predefined human preference (Figure 1). The goal of the PLLM is to utilize some techniques to let the LLM-generated response y align with the users' preference and expectations $\hat{y}$. More details are shown in Appendix A.

By incorporating personalized data, PLLMs enhance traditional LLMs, improving response generation, recommendation, and classification tasks. **Note that** our survey differs significantly from role-play related LLM personalization (Tseng et al., 2024; Chen et al., 2024a; Zhang et al., 2024g). While role-play focuses on mimicking characters during conversations, PLLMs in this survey focus on understanding users' contexts and preferences to meet their specific needs. Compared to (Zhang et al., 2024g), which emphasizes broad categories, our work provides a systematic analysis of techniques to enhance PLLM efficiency and performance, with a detailed technical classification.

2.3 Proposed Taxonomy

We propose a taxonomy (as illustrated in Figure 1 and Figure 2) from technical perspectives, categorizing the methods for Personalized Large Language Models (PLLMs) into three major levels: (1) **Input level: Personalized Prompting** focuses on handling user-specific data outside the LLM and injecting it into the model. (2) **Model level: Personalized Adaptation** emphasizes designing a framework to efficiently fine-tune or adapt model parameters for personalization. **(3) Objective Level: Personalized Alignment** aims to refine model behavior to align with user preferences effectively.

3 Personalized Prompting

Prompt engineering acts as a bridge for interaction between users and LLMs. In this survey, prompting involves guiding an LLM to generate desired outputs using various techniques, from traditional text prompts to advanced methods like soft embedding. Soft embedding can be extended not only through input but also via cross-attention or by adjusting output logits, enabling more flexible and context-sensitive responses. For each user u, the framework can be expressed as

$$y = f_{\text{LLM}}\left(q \oplus \phi\left(\mathcal{C}_u\right)\right), \tag{1}$$

where, f_{LLM} is the LLM model that generates the response; ϕ is a function that extracts relevant context from the user's personal context $\mathcal{C}_u$; $\oplus$ represents the combination operator that fuses the query q and the relevant personalized context $\phi(\mathcal{C}_u)$, producing enriched information for the LLM.

3.1 Profile-Augmented Prompting

Profile-augmented prompting (Figure 3(a)) explicitly utilize summarized user preferences and profiles in natural language to augment LLMs' input at the token level (ϕ is the *summarizer* model).

Non-tuned Summarizer A frozen LLM can be directly used as the summarizer to summarize user profiles due to its strong language understanding capabilities, i.e., $\phi\left(\mathcal{C}_u\right) = f_{\text{LLM}}\left(\mathcal{C}_u\right)$. For instance, *Cue-CoT* (Wang et al., 2023b) employs chain-of-thought prompting for personalized profile augmentation, using LLMs to extract and summarize user status (e.g., emotion, personality, and psychology) from historical dialogues. *PAG* (Richardson et al., 2023) leverages instruction-tuned LLMs to pre-summarize user profiles based on historical content. The summaries are stored offline, enabling efficient personalized response generation while meeting runtime constraints. *ONCE* (Liu et al., 2024c) prompts closed-source LLMs to summarize topics and regions of interest from users' browsing history, enhancing personalized recommendations.

Tuned Summarizer Black-box LLMs are sensitive to input noise, like off-topic summaries, and struggle to extract relevant information. Thus, training the summarizer to adapt to user preferences and style is essential. *Matryoshka* (Li et al., 2024a) uses a white-box LLM to summarize user histories, similar to PAG, but fine-tunes the summarizer instead of the generator LLM. *RewriterSlRl* (Li

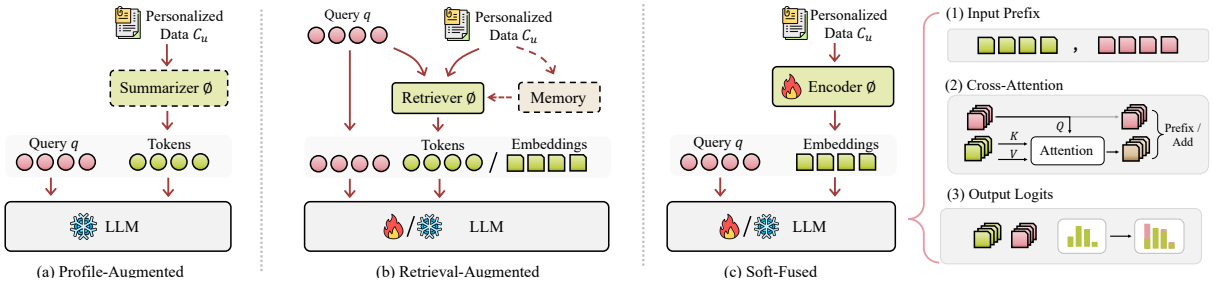

Figure 3: The illustration of personalized prompting approaches: **a) Profile-Augmented, b) Retrieval-Augmented, c) Soft-Fused.**

et al., 2024b) rewrites the query q instead of concatenating summaries, optimized with supervised and reinforcement learning.

CoS (He et al., 2024) is a special case that assumes a brief user profile $\phi(\mathcal{C}_u)$ and amplifies its influence in LLM response generation by comparing output probabilities with and without the profile, adjusting personalization without fine-tuning.

3.2 Retrieval-Augmented Prompting

Retrieval-augmented prompting (Gao et al., 2023; Fan et al., 2024; Qiu et al., 2024) excels at extracting the most relevant records from user data to enhance PLLMs (See Figure 3(b)). Due to the complexity and volume of user data, many methods use an additional *memory* for more effective retrieval. Common retrievers including sparse (e.g., BM25 (Robertson et al., 1995)), and dense retrievers (e.g., Faiss (Johnson et al., 2019), Contriever (Izacard et al., 2021)). These methods effectively manage the increasing volume of user data within the LLM's context limit, improving relevance and personalization by integrating key evidence from the user's personalized data.

3.2.1 Personalized Memory Construction

This part designs mechanisms for retaining and updating memory to enable efficient retrieval of relevant information.

Non-Parametric Memory This category maintains a token-based database, storing and retrieving information in its original tokenized form without using parameterized vector representations. For example, *MemPrompt* (Madaan et al., 2022) and *TeachMe* (Dalvi et al., 2022) maintain a dictionary-based feedback memory (key-value pairs of mistakes and user feedback). MemPrompt focuses on prompt-based improvements, whereas TeachMe emphasizes continual learning via dynamic memory that adapts over time. *MaLP* (Zhang et al., 2024a) further integrates multiple memory types, leveraging working memory for immediate processing, short-term memory (STM) for quick access, and long-term memory (LTM) for key knowledge.

Parametric Memory Recent studies parameterize and project personalized user data into a learnable space, with parametric memory filtering out redundant context to reduce noise. For instance, *LD-Agent* (Li et al., 2024c) maintains memory with separate short-term and long-term banks, encoding long-term events as parametric vector representations refined by a tunable module and retrieved via an embedding-based mechanism. *MemoRAG* (Qian et al., 2024), in contrast, adopts a different approach by utilizing a lightweight LLM as memory to learn user-personalized data. Instead of maintaining a vector database for retrieval, it generates a series of tokens as a draft to further guide the retriever, offering a more dynamic and flexible method for retrieval augmentation.

3.2.2 Personalized Memory Retrieval

The key challenge in the personalized retriever design lies in selecting not only relevant but also representative personalized data for downstream tasks. *LaMP* (Salemi et al., 2023) investigates how retrieved personalized information affects the responses of large language models (LLMs) through two mechanisms: in-prompt augmentation (IPA) and fusion-in-decoder (FiD). *PEARL* (Mysore et al., 2023) and *ROPG* (Salemi et al., 2024) similarly aim to enhance the retriever using personalized generation-calibrated metrics, improving both the personalization and text quality of retrieved documents. Meanwhile, *HYDRA* (Zhuang et al., 2024) trains a reranker to prioritize the most relevant information additionally from top-retrieved historical records for enhanced personalization.

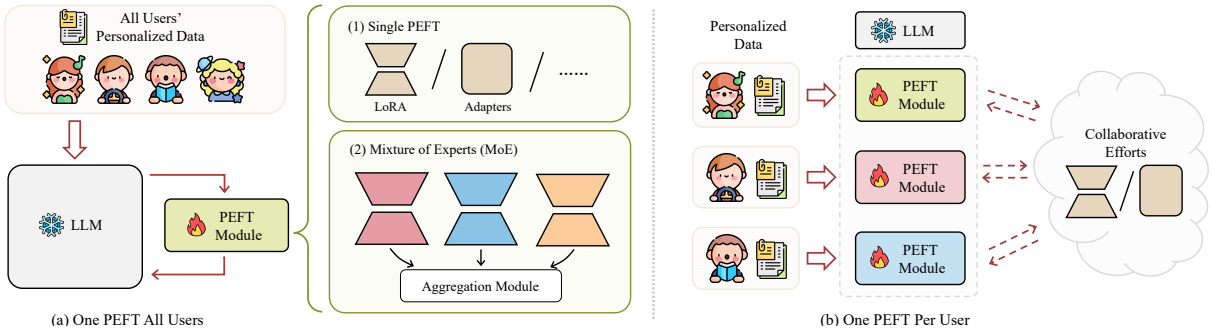

Figure 4: The illustration of personalized adaptation approaches: **a) One PEFT for all users, b) One PEFT per user.**

3.3 Soft-Fused Prompting

Soft prompting differs from profile-augmented prompting by compressing personalized data into soft embeddings, rather than summarizing it into discrete tokens. These embeddings are generated by a user feature *encoder* ϕ.

In this survey, we generalize the concept of soft prompting, showing that soft embeddings can be integrated (combination operator $\oplus$) not only through the input but also via cross-attention or by adjusting output logits, allowing for more flexible and context-sensitive responses (See Figure 3(c)).

Input Prefix Soft prompting, used as an input prefix, focuses on the embedding level by concatenating the query embedding with the soft embedding, and is commonly applied in recommendation tasks. *PPlug* (Liu et al., 2024b) constructs a user-specific embedding for each individual by modeling their historical contexts using a lightweight plug-in user embedder module. This embedding is then attached to the task input. *UEM* (Doddapaneni et al., 2024) is a user embedding module (transformer network) that generates a soft prompt conditioned on the user's personalized data. *PERSOMA* (Hebert et al., 2024) enhances UEM by employing resampling, selectively choosing a subset of user interactions based on relevance and importance. *REGEN* (Sayana et al., 2024) combines item embeddings from user-item interactions via collaborative filtering and item descriptions using a soft prompt adapter to generate contextually personalized responses. *PeaPOD* (Ramos et al., 2024) personalizes soft prompts by distilling user preferences into a limited set of learnable, dynamically weighted prompts. Unlike previously mentioned methods, which focus on directly embedding user interactions or resampling relevant data, PeaPOD adapts to user interests by weighting a shared set of prompts.

Cross-Attention Cross-attention enables the model to process and integrate multiple input sources by allowing it to attend to personalized data and the query. *User-LLM* (Ning et al., 2024) uses an autoregressive user encoder to convert historical interactions into embeddings through self-supervised learning, which are then integrated via cross-attention. The system employs joint training to optimize both the retriever and generator for better performance. *RECAP* (Liu et al., 2023) utilizes a hierarchical transformer retriever designed for dialogue domains to fetch personalized information. This information is integrated into response generation via a context-aware prefix encoder, improving the model's ability to generate personalized, contextually relevant responses.

Output Logits *GSMN* (Wu et al., 2021) retrieves relevant information from personalized data, encodes it into soft embeddings, and uses them in attention with the query vector. Afterward, the resulting embeddings are concatenated with the LLM-generated embeddings, modifying the final logits to produce more personalized and contextually relevant responses.

3.4 Discussions

While prompting methods are efficient and adaptable, enabling dynamic personalization with minimal computational overhead, they fall short in deeper personalization analysis and global knowledge access due to their reliance on predefined prompt structures (Appendix C).

4 Personalized Adaptation

PLLMs require balancing fine-tuning's deep adaptability with the efficiency of prompting. Therefore, specialized methods need to be specifically designed for PLLMs to address these challenges utilizing parameter-efficient fine-tuning methods

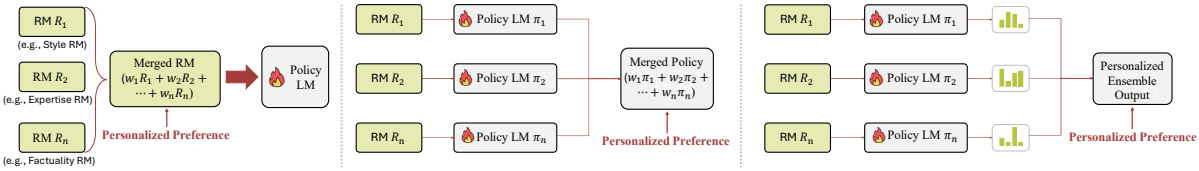

Figure 5: The illustration of personalized alignment method under the multi-objective reinforcement learning paradigm.

(PEFT), such as LoRA (Hu et al., 2021; Yang et al., 2024), prefix-tuning (Li and Liang, 2021), MeZo (Malladi et al., 2023), etc. (See Figure 4).

4.1 One PEFT All Users

This method trains on all users' data using a *shared PEFT module*, eliminating the need for separate modules per user. The shared module's architecture can be further categorized.

Single PEFT *PLoRA* (Zhang et al., 2024d) and *LM-P* (Woźniak et al., 2024) utilize LoRA for PEFT of LLM, injecting personalized information via user embeddings and user IDs, respectively. PLoRA is further extended and supports online training and prediction for cold-start scenarios. *UserIdentifier* (Mireshghallah et al., 2021) uses a static, non-trainable user identifier to condition the model on user-specific information, avoiding the need for trainable user-specific parameters and reducing training costs. *Review-LLM* (Peng et al., 2024b) aggregates users' historical behaviors and ratings into prompts to guide sentiment and leverages LoRA for efficient fine-tuning. However, these methods rely on a single architecture with fixed configurations (e.g., hidden size, insertion layers), making them unable to store and activate diverse information for personalization (Zhou et al., 2024). To solve this problem, *MiLP* (Zhang et al., 2024b) utilizes a Bayesian optimization strategy to automatically identify the optimal configuration for applying multiple LoRA modules, enabling efficient and flexible personalization.

Mixture of Experts (MoE) Several methods use the LoRA module, but with a static configuration for all users. This lack of parameter personalization limits adaptability to user dynamics and preference shifts, potentially resulting in suboptimal performance (Cai et al., 2024). *RecLoRA* (Zhu et al., 2024) addresses this limitation by maintaining a set of parallel, independent LoRA weights and employing a soft routing method to aggregate meta-LoRA weights, enabling more personalized and adaptive results. Similarly, *iLoRA* (Kong et al., 2024) creates a diverse set of experts (LoRA) to capture specific aspects of user preferences and generates dynamic expert participation weights to adapt to user-specific behaviors.

Shared PEFT methods rely on a centralized approach, where user-specific data is encoded into a shared adapter by centralized LLMs. This limits the model's ability to provide deeply personalized experiences tailored to individual users. Furthermore, using a centralized model often requires users to share personal data with service providers, raising concerns about the storage, usage, and protection of this data.

4.2 One PEFT Per User

Equipping *a user-specific PEFT module* makes LLM deployment more personalized while preserving data privacy. However, the challenge lies in ensuring efficient operation in resource-limited environments, as users may lack sufficient local resources to perform fine tuning.

No Collaboration There is no collaboration or coordination between adapters or during the learning process for each use in this category. *UserAdapter* (Zhong et al., 2021) personalizes models through prefix-tuning, fine-tuning a unique prefix vector for each user while keeping the underlying transformer model shared and frozen. *PocketLLM* (Peng et al., 2024a) utilizes a derivative-free optimization approach, based on MeZo (Malladi et al., 2023), to fine-tune LLMs on memory-constrained mobile devices. *OPPU* (Tan et al., 2024b) equips each user with a LoRA module.

Collaborative Efforts The "one-PEFT-per-user" paradigm without collaboration is computationally and storage-intensive, particularly for large user bases. Additionally, individually owned PEFTs hinder community value, as personal models cannot easily share knowledge or benefit from collaborative improvements. *PER-PCS* (Tan et al., 2024a) enables efficient and collaborative PLLMs by sharing a small fraction of PEFT parameters across users. It first divides PEFT parameters into

reusable pieces with routing gates and stores them in a shared pool. For each target user, pieces are autoregressively selected from other users, ensuring scalability, efficiency, and personalized adaptation without additional training.

Another efficient collaborative strategy is based on the federated learning (FL) framework. For example, (Wagner et al., 2024) introduces a FL framework for on-device LLM fine-tuning, using strategies to aggregate LoRA model parameters and handle data heterogeneity efficiently, outperforming purely local fine-tuning. *FDLoRA* (Qi et al., 2024) introduces a personalized FL framework using dual LoRA modules to capture personalized and global knowledge. It shares only global LoRA parameters with a central server and combines them via adaptive fusion, enhancing performance while minimizing communication and computing costs.

There are other frameworks that can be explored, such as *HYDRA* (Zhuang et al., 2024), which also employs a base model to learn shared knowledge. However, in contrast to federated learning, it assigns distinct heads to each individual user to extract personalized information.

4.3 Discussions

PEFT techniques reduce computational costs and memory usage while maintaining high personalization. It has the risk of overfitting with limited or noisy data, which can hinder generalization for new or diverse users (Appendix C).

5 Personalized Alignment

Alignment techniques (Bai et al., 2022; Rafailov et al., 2024) typically optimize LLMs to match the generic preferences of humans. However, in reality, individuals may exhibit significant variations in their preferences for LLM responses across different dimensions like language style, knowledge depth, and values. Personalized alignment seeks to further align with individual users' unique preferences beyond generic preferences. A significant challenge in personalized alignment is creating high-quality user-specific preference datasets, which are more complex than general alignment datasets due to data sparsity. The second challenge arises from the need to refine the canonical RLHF framework (Ouyang et al., 2022) to handle the diversification of user preferences, which is essential for integrating personalized preferences without compromising efficiency and performance.

5.1 Data Construction

High-quality data construction is critical for learning PLLMs, primarily involving self-generated data through interactions with the LLM. Wu et al.(Wu et al., 2024c) constructs a dataset for aligning LLMs with individual preferences by initially creating a diverse pool of 3,310 user personas, which are expanded through iterative self-generation and filtering. This method is similar to *PLUM* (Magister et al., 2024) that both simulate dynamic interactions through multi-turn conversation trees, allowing LLMs to infer and adapt to user preferences. To enable LLMs to adapt to individual user preferences without re-training, Lee et al. (Lee et al., 2024) utilizes diverse system messages as meta-instructions to guide the models' behavior. To support this, the MULTIFACETED COLLECTION dataset is created, comprising 197k system messages that represent a wide range of user values. To facilitate real-time, privacy-preserving personalization on edge devices while addressing data privacy, limited storage, and minimal user disruption, Qin et al. (Qin et al., 2024) introduces a self-supervised method that efficiently selects and synthesizes essential user data, improving model adaptation with minimal user interaction.

Research efforts are also increasingly concentrating on developing datasets that assess models' comprehension of personalized preferences. Kirk et al. (Kirk et al., 2024) introduces *PRISM Alignment Dataset* that maps the sociodemographics and preferences of 1,500 participants from 75 countries to their feedback in live interactions with 21 LLMs, focusing on subjective and multicultural perspectives on controversial topics. *PersonalLLM* (Zollo et al., 2024) introduces a novel personalized testdb, which curates open-ended prompts and multiple high-quality responses to simulate diverse latent preferences among users. It generates simulated user bases with varied preferences from pre-trained reward models, addressing the challenge of data sparsity in personalization.

5.2 Personalized Alignment Optimization

Personalized preference alignment is usually modeled as a multi-objective reinforcement learning (MORL) problem, where personalized preference is determined as the user-specific combination of multi-preference dimensions. Based on this, a typical alignment paradigm involves using a personalized reward derived from multiple reward

models to guide during the training phase of policy LLMs, aiming for personalization (Figure 5). *MORLHF* (Wu et al., 2023) separately trains reward models for each dimension and retrains the policy language models using proximal policy optimization, guided by a linear combination of these multiple reward models. This approach allows for the reuse of the standard RLHF pipeline. *MODPO* (Zhou et al., 2023) introduces a novel RL-free algorithm extending Direct Preference Optimization (DPO) for managing multiple alignment objectives. It integrates linear scalarization into the reward modeling process, enabling the training of LMs using a margin-based cross-entropy loss as implicit collective reward functions.

Another strategy for MORL is to consider ad-hoc combinations of multiple trained policy LLMs during the decoding phase to achieve personalization. *Personalized Soups* (Jang et al., 2023) and *Reward Soups* (Rame et al., 2024) address the challenge of RL from personalized human feedback by first training multiple policy models with distinct preferences independently and then merging their parameters post-hoc during inference. Both methods allow for dynamic weighting of the networks based on user preferences, enhancing model alignment and reducing reward misspecification. Also, the personalized fusion of policy LLMs can be achieved not only through parameter merging but also through model ensembling. *MOD* (Shi et al., 2024) outputs the next token from a linear combination of all base models, allowing for precise control over different objectives by combining their predictions without the need for retraining. The method demonstrates significant effectiveness when compared to the parameter-merging baseline. *PAD* (Chen et al., 2024b) leverages a personalized reward modeling strategy to generate token-level rewards that guide the decoding process, enabling the dynamic adaptation of the base model's predictions to individual preferences.

There are some other emerging personalized alignment studies beyond the "multi-objective" paradigm. *PPT* (Lau et al., 2024) unlocks the potential of in-context learning for scalable and efficient personalization by generating two potential responses for each user prompt, asking the user to rank them, and incorporating this feedback into the model's context to dyanmic adapt to individual preferences over time. *VPL* (Poddar et al., 2024) employs a variational inference framework to capture diverse human preferences via user-specific latent variables. By inferring these latent distributions from limited preference annotations, it enhances the accuracy and personalization of reward modeling while improving data efficiency.

5.3 Discussions

Personalized alignment technologies model personalization as multi-objective reinforcement learning, incorporating user preferences during training with RLHF or in decoding via parameter merging. They often use a limited set of predefined preference dimensions, while real-world scenarios involve many users with unknown preferences based solely on interaction history (Appendix C).

6 Future Directions

Despite advances in Personalized Large Language Models (PLLMs), significant challenges persist, particularly in **technical improvements**. Current methods effectively handle basic user preferences but struggle with complex, multi-source data, especially in multimodal contexts like images and audio. Efficiently updating models on resource-constrained edge devices is also crucial. Fine-tuning enhances personalization but can be resource-intensive and difficult to scale. Developing small, personalized models through techniques like quantization could address these issues.

Trustworthiness remains a critical concern, particularly regarding user privacy when generating personalized responses. As LLMs are not typically deployed locally, risks of privacy leakage arise. Future research should focus on privacy-preserving methods, such as federated learning and differential privacy, to protect user data effectively while leveraging the model's capabilities. Please check Appendix D for more explanations.

7 Conclusions

This survey offers a comprehensive overview of PLLMs, focusing on personalized responses to individual user data. It presents a taxonomy categorizing approaches into three key perspectives: Personalized Prompting (Input Level), Personalized Adaptation (Model Level), and Personalized Alignment (Objective Level), with further subdivisions. A detailed method summarization is shown in Table 1. We highlight current limitations and suggest future research directions, providing valuable insights to advance PLLM development.

8 Limitations

In this paper, we present a detailed survey of personalized large language models. However, the fast-paced advancement of this field poses challenges in covering all research efforts, as new methods, datasets, and evaluation metrics constantly emerge, necessitating ongoing updates to our taxonomy. Additionally, developing more effective and universally accepted benchmarks for different personalized tasks is an ongoing challenge.

References

Steven Au, Cameron J Dimacali, Ojasmitha Pedirappagari, Namyong Park, Franck Dernoncourt, Yu Wang, Nikos Kanakaris, Hanieh Deilamsalehy, Ryan A Rossi, and Nesreen K Ahmed. 2025. Personalized graph-based retrieval for large language models. *arXiv:2501.02157*.

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Proc. of NeurIPS*.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2024. A survey on mixture of experts. *arXiv:2407.06204*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

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

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv:2210.11416*.

Christopher Clarke, Yuzhao Heng, Lingjia Tang, and Jason Mars. 2024. Peft-u: Parameter-efficient finetuning for user personalization. *arXiv:2407.18078*.

Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. 2022. Towards teachable reasoning systems: Using a dynamic memory of user feedback for continual system improvement. In *Proc. of EMNLP*.

Sumanth Doddapaneni, Krishna Sayana, Ambarish Jash, Sukhdeep Sodhi, and Dima Kuzmin. 2024. User embedding model for personalized language prompting. *arXiv:2401.04858*.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. 2024. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and synthesis in question answering. *arXiv:2402.16288*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proc. of KDD*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv:2312.10997*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv:2501.12948*.

Jerry Zhi-Yang He, Sashrika Pandey, Mariah L Schrum, and Anca Dragan. 2024. Cos: Enhancing personalization and mitigating bias with context steering. *arXiv:2405.01768*.

Liam Hebert, Krishna Sayana, Ambarish Jash, Alexandros Karatzoglou, Sukhdeep S Sodhi, Sumanth Doddapaneni, Yanli Cai, and Dima Kuzmin. 2024. Persoma: Personalized soft prompt adapter architecture for personalized language prompting. *CoRR*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*.

Minda Hu, Licheng Zong, Hongru Wang, Jingyan Zhou, Jingjing Li, Yichen Gao, Kam-Fai Wong, Yu Li, and Irwin King. 2024. SeRTS: Self-rewarding tree search for biomedical retrieval-augmented generation. In *Proc. of EMNLP Findings*.

Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. Concept–an evaluation protocol on conversational recommender systems with system-centric and user-centric factors. *arXiv preprint arXiv:2404.03304*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv:2112.09118*.

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. *arXiv:2310.11564*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

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

Minbeom Kim, Kang-il Lee, Seongho Joo, Hwaran Lee, and Kyomin Jung. 2025. Drift: Decoding-time personalized alignments with implicit user preferences. *arXiv preprint arXiv:2502.14289*.

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

Xiaoyu Kong, Jiancan Wu, An Zhang, Leheng Sheng, Hui Lin, Xiang Wang, and Xiangnan He. 2024. Customizing language models with instance-wise lora for sequential recommendation. In *Proc. of NeurIPS*.

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

Allison Lau, Younwoo Choi, Vahid Balazadeh, Keertana Chidambaram, Vasilis Syrgkanis, and Rahul G Krishnan. 2024. Personalized adaptation via in-context preference learning. *arXiv:2410.14001*.

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

Changhao Li, Yuchen Zhuang, Rushi Qiang, Haotian Sun, Hanjun Dai, Chao Zhang, and Bo Dai. 2024a. Matryoshka: Learning to drive black-box llms with llms. *arXiv:2410.20749*.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024b. Learning to rewrite prompts for personalized text generation. In *Proc. of Web Conference*.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023. Teach llms to personalize–an approach inspired by writing education. *arXiv:2308.07968*.

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv:2101.00190*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jiahong Liu, Xinyu Fu, Menglin Yang, Weixi Zhang, Rex Ying, and Irwin King. 2024a. Client-specific hyperbolic federated learning. In *FedKDD@KDD*.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2024b. Llms+ persona-plug= personalized llms. *arXiv:2409.11901*.

Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024c. ONCE: boosting content-based recommendation with both open- and closed-source large language models. In *Proc. of WSDM*.

Shuai Liu, Hyundong J Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023. Recap: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. *arXiv:2306.07206*.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. 2024. Small language models: Survey, measurements, and insights. *arXiv:2409.15790*.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve gpt-3 after deployment. In *Proc. of EMNLP*.

Lucie Charlotte Magister, Katherine Metcalf, Yizhe Zhang, and Maartje ter Hoeve. 2024. On the way to llm personalization: Learning to remember user conversations. *arXiv:2411.13405*.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. 2023. Fine-tuning language models with just forward passes. *Proc. of NeurIPS*.

Fatemehsadat Mireshghallah, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim, and Dimitrios Dimitriadis. 2021. Useridentifier: implicit user representations for simple and effective personalized sentiment analysis. *arXiv:2110.00135*.

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

Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. 2024. User-llm: Efficient llm contextualization with user embeddings. *arXiv:2402.13598*.

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

Dan Peng, Zhihui Fu, and Jun Wang. 2024a. Pocketllm: Enabling on-device fine-tuning for personalized llms. *arXiv:2407.01031*.

Qiyao Peng, Hongtao Liu, Hongyan Xu, Qing Yang, Minglai Shao, and Wenjun Wang. 2024b. Llm: Harnessing large language models for personalized review generation. *arXiv:2407.07487*.

Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. 2024. Personalized visual instruction tuning. *arXiv:2410.07113*.

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

Jiaxing Qi, Zhongzhi Luan, Shaohan Huang, Carol Fung, Hailong Yang, and Depei Qian. 2024. Fdlora: Personalized federated learning of large language model via dual lora tuning. *arXiv:2406.07925*.

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

Ruiyang Qin, Jun Xia, Zhenge Jia, Meng Jiang, Ahmed Abbasi, Peipei Zhou, Jingtong Hu, and Yiyu Shi. 2024. Enabling on-device large language model personalization with self-supervised data selection and synthesis. In *Proc. of DAC*.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. 2025. Measuring what makes you unique: Difference-aware user modeling for enhancing llm personalization. *arXiv preprint arXiv:2503.02450*.

Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. 2024. Entropy-based decoding for retrieval-augmented large language models. *arXiv:2406.17519*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Proc. of NuerIPS*, 36.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2024. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Proc. of NeurIPS*.

Jerome Ramos, Bin Wu, and Aldo Lipani. 2024. Preference distillation for personalized generative recommendation. *arXiv:2407.05033*.

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

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*.

Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization methods for personalizing large language models through retrieval augmentation. In *Proc. of SIGIR*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv:2304.11406*.

Alireza Salemi and Hamed Zamani. 2024. Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. *arXiv:2409.09510*.

Krishna Sayana, Raghavendra Vasudeva, Yuri Vasilevski, Kun Su, Liam Hebert, Hubert Pham, Ambarish Jash, and Sukhdeep Sodhi. 2024. Beyond retrieval: Generating narratives in conversational recommender systems. *arXiv:2410.16780*.

Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM on Web Conference 2024*.

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

Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Han Li. 2025. Retrieval augmented generation with collaborative filtering for personalized text generation. *arXiv preprint arXiv:2504.05731*.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024a. Personalized pieces: Efficient personalized large language models through collaborative efforts. *arXiv:2406.10471*.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024b. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proc. of EMNLP*.

Yuqing Tian, Zhaoyang Zhang, Yuzhi Yang, Zirui Chen, Zhaohui Yang, Richeng Jin, Tony QS Quek, and Kai-Kit Wong. 2024. An edge-cloud collaboration framework for generative ai service provision with synergetic big cloud model and small edge models. *arXiv:2401.01666*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv:2302.13971*.

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

Anvesh Rao Vijjini, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2024. Exploring safety-utility trade-offs in personalized language models. *arXiv:2406.11107*.

Nicolas Wagner, Dongyang Fan, and Martin Jaggi. 2024. Personalized collaborative fine-tuning for on-device large language models. *arXiv:2404.09753*.

Hongru Wang, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. 2023a. Large language models as source planner for personalized knowledge-grounded dialogue. *arXiv:2310.08840*.

Hongru Wang, Huimin Wang, Lingzhi Wang, Minda Hu, Rui Wang, Boyang Xue, Yongfeng Huang, and Kam-Fai Wong. 2024a. Tpe: Towards better compositional reasoning over cognitive tools via multi-persona collaboration. In *NLPCC*, pages 281–294. Springer.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023b. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. In *Proc. of EMNLP Findings*.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024b. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. *arXiv:2402.09269*.

Bin Wu, Zhengyan Shi, Hossein A Rahmani, Varsha Ramineni, and Emine Yilmaz. 2024a. Understanding the role of user profile in the personalization of large language models. *arXiv:2406.17803*.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *Proc. of ICLR*.

Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A Rossi, Franck Dernoncourt, et al. 2024b. Personalized multimodal large language models: A survey. *arXiv:2412.02142*.

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

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024d. Continual learning for large language models: A survey. *arXiv:2402.01364*.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proc. of NAACL*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *Proc. of NeurIPS*.

Menglin Yang, Jialin Chen, Yifei Zhang, Jiahong Liu, Jiasheng Zhang, Qiyao Ma, Harshit Verma, Qianru Zhang, Min Zhou, Irwin King, et al. 2024. Low-rank adaptation for foundation models: A comprehensive review. *arXiv:2501.00365*.

Yuhang Yao, Jianyi Zhang, Junda Wu, Chengkai Huang, Yu Xia, Tong Yu, Ruiyi Zhang, Sungchul Kim, Ryan Rossi, Ang Li, et al. 2024. Federated large language models: Current progress and future directions. *arXiv:2409.15723*.

Yoel Zeldes, Amir Zait, Ilia Labzovsky, Danny Karmon, and Efrat Farkash. 2025. Commer: a framework for compressing and merging user data for personalization. *arXiv preprint arXiv:2501.03276*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv:2210.02414*.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025a. Personalized text generation with contrastive activation steering. *arXiv preprint arXiv:2503.05213*.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2024a. Llm-based medical assistant personalization with short-and long-term memory coordination. In *Proc. of NAACL*.

Kai Zhang, Lizhi Qing, Yangyang Kang, and Xiaozhong Liu. 2024b. Personalized llm response generation with parameterized memory injection. *arXiv:2404.03565*.

Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. 2025b. Proper: A progressive learning framework for personalized large language models with group-level adaptation. *arXiv preprint arXiv:2503.01303*.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024c. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Yi Zhang, Zhongyang Yu, Wanqi Jiang, Yufeng Shen, and Jin Li. 2023. Long-term memory for large language models through topic-based vector database. In *Proc. of IALP*.

You Zhang, Jin Wang, Liang-Chih Yu, Dan Xu, and Xuejie Zhang. 2024d. Personalized lora for human-centered text understanding. In *Proc. of AAAI*.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024e. A survey on the memory mechanism of large language model based agents. *arXiv:2404.13501*.

Zeyu Zhang, Quanyu Dai, Luyu Chen, Zeren Jiang, Rui Li, Jieming Zhu, Xu Chen, Yi Xie, Zhenhua Dong, and Ji-Rong Wen. 2024f. Memsim: A bayesian simulator for evaluating memory of llm-based personal assistants. *arXiv:2409.20163*.

Zhaowei Zhang, Fengshuo Bai, Qizhi Chen, Chengdong Ma, Mingzhi Wang, Haoran Sun, Zilong Zheng, and Yaodong Yang. 2025c. Amulet: Realignment during test time for personalized preference adaptation of llms. In *Proc. of ICLR*.

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

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025a. Do llms recognize your preferences? evaluating personalized preference following in llms. *Proc. of ICLR*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv:2303.18223*.

Xiaoyan Zhao, Yang Deng, Wenjie Wang, Hong Cheng, Rui Zhang, See-Kiong Ng, Tat-Seng Chua, et al. 2025b. Exploring the impact of personality traits on conversational recommender systems: A simulation with large language models. *arXiv preprint arXiv:2504.12313*.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. *arXiv:2204.08128*.

Wanjun Zhong, Duyu Tang, Jiahai Wang, Jian Yin, and Nan Duan. 2021. Useradapter: Few-shot user learning in sentiment analysis. In *Proc. of ACL Findings*.

Han Zhou, Xingchen Wan, Ivan Vulić, and Anna Korhonen. 2024. Autopeft: Automatic configuration search for parameter-efficient fine-tuning. *Proc. of ACL*.

Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, and Yu Qiao. 2023. Beyond one-preference-for-all: Multi-objective direct preference optimization. *arXiv:2310.03708*.

Jiachen Zhu, Jianghao Lin, Xinyi Dai, Bo Chen, Rong Shan, Jieming Zhu, Ruiming Tang, Yong Yu, and Weinan Zhang. 2024. Lifelong personalized low-rank adaptation of large language models for recommendation. *arXiv:2408.03533*.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. Hydra: Model factorization framework for black-box llm personalization. *arXiv:2406.02888*.

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

A Supplementary for Problem Statement

A.1 Input Description: Personalized Data

A detailed description, including examples, of the classification of input data types is provided:

- **Profile/Relationship:** User profile, including attributes (e.g., name, gender, occupation), and relationships (e.g., friends, family members), such as $\mathcal{C}_u = \{A, 18, \text{student}, \text{friends:}\{B, C, D\} \dots\}$.
- **Historical Dialogues:** Historical dialogues, such as question-answer pairs that user u interacts with the LLM (e.g., $\mathcal{C}_u = \{(q_0, a_0), (q_1, a_1), \dots, (q_i, a_i)\}$), where each q_i is a query and a_i is the answer.
- **Historical Content:** Includes documents, previous reviews, comments or feedback from user u. For example, $\mathcal{C}_u = \{\text{I like } \textit{Avtar} \text{ because } \dots, \dots\}$.
- **Historical Interactions:** Includes historical interactions, preferences, ratings from user u. For example, $\mathcal{C}_u = \{\textit{The Lord of the Rings} : 5, \textit{Interstellar} : 3 \dots\}$.
- **Pre-defined Human Preference**: Define a set $S = \{d_k\}_{k=1}^{K}$ containing of K preference dimensions such as "Helpfulness". Choose various combinations of these dimensions, form individual preferences, and incorporate them as the instruction. For example, a preference prompt could be "Be harmless and helpful".

A.2 Task Description

We divide personalized tasks from two perspectives: one is from the viewpoint of downstream tasks, and the other is from the classification of problems addressed by PLLMs. Different types of problems may require distinct technical approaches tailored to their specific characteristics.

A.2.1 Downstream Task Perspective

A detailed description, including examples, of the generated response y for downstream tasks.

- **Generation:** Generation tasks typically involve y representing a sequence of strings, such as generating answers for users based on their personalized data $\mathcal{C}_u$ and questions or generating content according to the user's writing style to assist their writing, and so forth (Salemi et al., 2023; Kumar et al., 2024; Zhao et al., 2025a; Au et al., 2025).
- **Recommendation:** The major difference between recommendation and generation is that recommendation requires suggesting specific items based on the user's historical interaction data, and it can provide reasons and explanations for the recommendations (Sayana et al., 2024).
- **Classification:** Classification tasks, including sentiment classification, involve labeling a particular entity (such as a movie, item, or description) based on the user's preferences to assist the user in categorization or summarization (Salemi et al., 2023; Au et al., 2025; Zhao et al., 2025a).

A.2.2 Personalized Query Types

The types of problems are mainly categorized based on the user's query q. Different queries q have different focal points regarding the expected answers $\hat{y}$.

- **Fact-based Queries (explicit):** These queries seek to retrieve or display **specific factual information**. Examples include questions like *"What time should I go out to play?"* or requests for *"passport information"* and similar concrete factual details (Du et al., 2024; Wu et al., 2025).
- **Personalized Associative Queries (explicit / implicit):** Some implicitly associated factual information that summarizes the user's preferences is contained within the user's personalized data $\mathcal{C}_u$. This type of query does not explicitly express the factual personalized information, but requires the LLM to generalize and answer related questions based on the user's implicit interests. For example, *"Can you recommend a good movie to watch this weekend?"* or *"Can you recommend a restaurant that suits my taste?"* (Zhao et al., 2025a; Wu et al., 2025).
- **Style-related Queries (implicit):** These queries need LLM to focus on summarizing the user's preferences or style, such as their writing style, preferred tone, or taste in tagging movies. For example, *"Help me write an*

email in my writing style" or *"Help me categorize the following movies"* (Salemi et al., 2023; Au et al., 2025; Zhao et al., 2025a).

B Evaluation Metrics

The evaluation metrics differ for different tasks.

B.1 Generation Task

Conventional Evaluation: BLEU, ROUGE-1 (Lin, 2004), ROUGE-L and METEOR (Banerjee and Lavie, 2005) to measure lexical overlap between the generated y and ground-truth $\hat{y}$ responses. These metrics provide surface-level comparisons via n-gram matching and semantic alignment techniques.

LLM-based Evaluation: The "LLM-as-a-judge" framework (Gu et al., 2024) uses LLM to automatically evaluate the quality and relevance of generated text. By prompting LLMs to score or compare outputs, it offers scalable, context-aware, and semantically rich assessments with minimal human input. This approach surpasses traditional metrics in flexibility but faces challenges like model bias and consistency. It represents a promising method for automated, nuanced evaluation.

B.2 Recommendation Task

Item Recommendation: Traditional recommendation tasks typically use (Hit Ratio) HR, Recall and Discounted Cumulative Gain (NDCG) (Ning et al., 2024; Ramos et al., 2024) as standard evaluation metrics to measure the effectiveness of top-K recommendation and preference ranking.

Conversational Recommendation: Conversational recommendation systems commonly use Recall and NDCG as evaluation metrics to measure coverage and ranking quality, employing "LLMs-as-a-judger" framework (Zhao et al., 2025b; Huang et al., 2024; Sayana et al., 2024). Additionally, an LLM-based user simulator—creating unique personas via zero-shot ChatGPT prompting and defining preferences using dataset attributes—is also used to assess whether outputs align with user preferences (Huang et al., 2024).

B.3 Classification Task

Multi-class Classification: In multi-class classification, where labels are categorical without inherent order, standard evaluation metrics such as Accuracy and F1 score (Salemi et al., 2023) are commonly employed to assess the model's ability to correctly predict class membership.

Ordinal Classification: For ordinal multi-class classification, where labels possess a natural order or ranking, performance metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) (Salemi et al., 2023) are preferred, as they account for the magnitude of prediction errors relative to the true order, providing a more nuanced evaluation of model quality.

C Supplementary for Discussions

Discussion for Prompting The three prompting methods have distinct pros and cons:

- Profile-augmented prompting improves efficiency by compressing historical data but risks information loss and reduced personalization.
- Retrieval-augmented prompting offers rich, context-aware inputs and scales well for long-term memory but can suffer from computational limits and irrelevant data retrieval.
- Soft prompting efficiently embeds user-specific info, capturing semantic nuances without redundancy, but is limited to black-box models and lacks explicit user preference analysis.

Overall, prompting-based methods are efficient and adaptable, and enable dynamic personalization with minimal computational overhead. These methods are more suitable for *fact-based queries (explicit)* that need to answer factual information, mentioned in Appendix A.2.2. However, they lack deeper personalization analysis as they rely on predefined prompt structures to inject user-specific information and are limited in accessing global knowledge due to the narrow scope of prompts, which fail in tasks with *style-related queries (implicit)*.

Discussion for Adaptation Fine-tuning methods enable deep personalization by modifying a large set of model parameters, and parameter-efficient fine-tuning methods (e.g., prefix vectors or adapters) reduce computational cost and memory requirements while maintaining high personalization levels. These methods improve task adaptation by tailoring models to specific user needs, enhancing performance in tasks like sentiment analysis and recommendations. They also offer flexibility, allowing user-specific adjustments while leveraging pre-trained knowledge. However, they still face

the risk of overfitting, particularly with limited or noisy user data, which can impact generalization and performance for new or diverse users.

Discussion for Alignment. Current mainstream personalized alignment technologies mainly model personalization as multi-objective reinforcement learning problems, where personalized user preferences are taken into account during the training phase of policy LLMs via canonical RLHF, or the decoding phase of policy LLM via parameter merging or model ensembling. Typically, these methods are limited to a small number (e.g., three) of predefined preference dimensions, represented through textual user preference prompts. However, in real-world scenarios, there could be a large number of personalized users, and their preference vectors may not be known, with only their interaction history accessed. Consequently, developing more realistic alignment benchmarks to effectively assess these techniques is a critical area for future research.

D Future Directions

Despite recent advances in PLLMs, challenges and opportunities remain. This section discusses key limitations and promising future directions.

Complex User Data While current approaches effectively handle basic user preferences, processing complex, multi-source user data remains a significant challenge. For example, methods that use user relationships in graph-like structures are still limited to retrieval augmentation (Du et al., 2024). How to effectively leverage this complex user information to fine-tune LLM parameters remains a significant challenge. Most methods focus on text data, while personalized foundation models for multimodal data (e.g., images, videos, audio) remain underexplored, despite their significance for real-world deployment and applications (Wu et al., 2024b; Pi et al., 2024; Shen et al., 2024).

Edge Computing A key challenge in edge computing is efficiently updating models on resource-constrained devices (e.g., phones), where storage and computational resources are limited. For example, fine-tuning offers deeper personalization but is resource-intensive and hard to scale, especially in real-time applications. Balancing resources with personalization needs is important. A potential solution is to build personalized small models (Lu et al., 2024) for edge devices, using techniques like quantization and distillation.

Edge-Cloud Collaboration The deployment of PLLMs in real-world scenarios encounters significant challenges in edge-cloud computing environments. Current collaborative approaches often lack efficient synchronization between cloud and edge devices, highlighting the need to balance local computation and cloud processing (Tian et al., 2024).

Efficient Adaptation to Model Updates Updating fine-tuned PEFT parameters for each user when base LLM parameters change poses a challenge due to high user data volume and limited resources. Retraining costs can be prohibitive. Future research should focus on efficient methods for updating user-specific parameters without complete retraining, such as incremental learning and transfer learning.

Lifelong Updating Given the large variety of user behaviors, a key challenge is preventing catastrophic forgetting while ensuring the efficient update of long-term and short-term of memory. Future research could explore continual learning (Wu et al., 2024d) and knowledge editing (Wang et al., 2024b; Zhang et al., 2024c) to facilitate dynamic updates of user-specific information.

Trustworthiness Ensuring user privacy is crucial, especially when summarized or retrieved data is used to generate personalized responses. Since LLMs cannot be deployed locally due to resource limits, there is a risk of privacy leakage. Future research could focus on privacy-preserving methods like federated learning, secure computation, and differential privacy to protect user data (Yao et al., 2024; Liu et al., 2024a).

Method	Personalized Data	LLM (Generator)	Retriever	Prompting	Fine-tuning	Task
§3 **Personalized Prompting**						
Cue-CoT (Wang et al., 2023b)	Dialogues	ChatGLM-6B, BELLE-LLAMA-7B-2M, ChatGPT, Alpaca-7B, Vicuna-7B-v1.1	✗	Token	✗	**G**
‡ PAG (Richardson et al., 2023)	Content	ChatGPT-3.5, FlanT5, Vicuna-13B	✗	Token	✗	**G** , **C**
‡ Matryoshka (Li et al., 2024a)	Content	gpt-4o-mini, gpt-3.5-turbo	✗	Token	✗	**G** , **C**
RewriterSIRl (Li et al., 2024b)	Content	PaLM2	✗	Token	✗	**G**
CoS (He et al., 2024)	User Profile	GPT-3.5, Llama2-7B-Chat, T0pp, Mistral-7B-Instruct	✗	Token	✗	**G** , **C**
MemPrompt (Madaan et al., 2022)	Content	GPT3	✓	Token	✗	**G**
TeachMe (Dalvi et al., 2022)	Content	GPT3	✓	Token	✗	
MaLP (Zhang et al., 2024a)	Dialogues	GPT3.5, LLaMA-7B, LLaMA-13B	✓	Token	✓(LoRA)	**G** , **C**
LD-Agent (Li et al., 2024c)	Dialogues	ChatGLM, ChatGPT, BlenderBot	✓	Token	✓	**G**
MemoRAG (Qian et al., 2024)	Dialogues	Qwen2-7B-Instruct, Mistral-7B-Instruct	✓	Token	✗	**G**
‡ IPA (Salemi et al., 2023)	Content	FlanT5-base	✓	Token	✗	**G** , **C**
‡ FiD (Salemi et al., 2023)	Content	FlanT5-base	✓	Token	✓	**G** , **C**
MSP (Zhong et al., 2022)	Dialogues	DialoGPT	✓	Token	✗	**G**
AuthorPred (Li et al., 2023)	Content	T5-11B	✓	Token	✓	**G** , **C**
PEARL (Mysore et al., 2023)	Content	davinci-003, gpt-35-turbo	✓	Token	✗	**G**
‡ ROPG (Salemi et al., 2024)	Content	FlanT5-XXL-11B	✓	Token	✗	**G** , **C**
‡ HYDRA (Zhuang et al., 2024)	Content	gpt-3.5-turbo	✓	Token	(Adaptor)	**G** , **C**
UEM (Doddapaneni et al., 2024)	Interactions	FlanT5-base, FlanT5-Large	✗	Embedding	✓	**C**
PERSOMA (Hebert et al., 2024)	Interactions	PaLM 2	✗	Embedding	✓(LoRA)	**G**
REGEN (Sayana et al., 2024)	Interactions	PaLM2	✗	Embedding	✗	**G**
PeaPOD (Ramos et al., 2024)	Interactions	T5-small	✗	Embedding	✓	**G** , **R**
‡ PPlug (Liu et al., 2024b)	Content	FlanT5-XXL-11B	✗	Embedding	✗	**G** , **C**
User-LLM (Ning et al., 2024)	Interactions	PaLM-2 XXS	✗	Embedding	✓	**G** , **R**
RECAP (Liu et al., 2023)	Dialogues	DialoGPT	✓	Embedding	✓	**G**
GSMN (Wu et al., 2021)	User profile Comment	DialoGPT	✓	Embedding	✓	**G**
§4 **Personalized Adaptation**						
PLoRA (Zhang et al., 2024d)	User profile (ID)	BERT, RoBERTa, Flan-T5	✗	✗	LoRA	**C**
UserIdentifier (Mireshghallah et al., 2021)	User profile	RoBERTa-base	✗	✗	-	**C**
LM-P (Woźniak et al., 2024)	User profile (ID)	Mistral 7B, Flan-T5, Phi-2, StableLM, GPT-3.5, GPT-4	✗	✗	LoRA	**G** , **C**
Review-LLM (Peng et al., 2024b)	Interactions	GPT-3.5-turbo, GPT-4o, Llama-3-8b	✗	✓	LoRA	**G**
MiLP (Zhang et al., 2024b)	Content Dialogues	DialoGPT, RoBERTa, LLaMA2-7B, LlaMA2-13B	✗	✗	LoRA	**G**
RecLoRA (Zhu et al., 2024)	Interactions	Vicuna-7B	✓	✓	LoRA	**R**
iLoRA (Kong et al., 2024)	Interactions	Llama2-7B	✗	✗	LoRA	**R**
UserAdapter (Zhong et al., 2021)	Interactions	RoBERTa-base	✗	✗	Prefix-tuning	**R**
PocketLLM (Peng et al., 2024a)	Content	RoBERTa-large, OPT-1.3B	✗	✗	MeZo	**C**
‡ OPPU (Tan et al., 2024b)	Content	Llama-2-7B	(o)	(o)	LoRA	**G** , **C**
‡ PER-PCS (Tan et al., 2024a)	Content	Llama-2-7B	(o)	(o)	LoRA	**G** , **C**
(Wagner et al., 2024)	Content	GPT2	✗	✗	LoRA	**C**
FDLoRA (Qi et al., 2024)	Content	LLaMA2-7B	✗	✗	LoRA	**C**
§5 **Personalized Alignment**						
(Wu et al., 2024c)	Dialogues	Qwen2-7B-Instruct, LLaMA-3-8B-Instruct, Mistral-7B-Instruct-v0.3	✗	✗	✓	**G**
PLUM (Magister et al., 2024)	Dialogues	LLaMA-3-8B-Instruct	✗	✗	LoRA	**G**
(Lee et al., 2024)	User Profile	Mistral-7B-v0.2	✗	✗	✓	**G**
MORLHF (Wu et al., 2023)	Preference	GPT2, T5-Large	✗	✗	✓	**G**
MODPO (Zhou et al., 2023)	Preference	LLaMA-7B	✗	✗	LoRA	**G**
Personalized Soups (Jang et al., 2023)	Preference	Tulu-7B	✗	✗	LoRA	**G**
Reward Soups (Rame et al., 2024)	Preference	LLaMA-7B	✗	✗	LoRA	**G**
MOD (Shi et al., 2024)	Preference	LLaMA-2-7B	✗	✗	✓	**G**
PAD (Chen et al., 2024b)	Preference	LLaMA-3-8B-Instruct, Mistral-7B-Instruct	✗	✗	LoRA	**G**
PPT (Lau et al., 2024)	Preference	Self-defined	✗	✗	✓	**G**
VPL (Poddar et al., 2024)	Preference	GPT-2, LLaMA-2-7B	✗	✗	LoRA	**G**

Table 1: A systematic categorization of personalization strategies for PLLMs. Methods marked with ‡ use the LaMP benchmark. (o) means optional. The overview presents four data categories (Historical Content, Dialogues, Interactions, User profile, Pre-defined Human Preference) and three task types (Generation **G** , Classification **C** , Recommendation **R**), along with fine-tuning requirements for generator LLMs.