From Personal to Collective: On the Role of Local and Global Knowledge in LLM Personalization

Zehong Wang[1], Junlin Wu[3], Zhaoxuan Tan[1], Bolian Li[4], Xianrui Zhong[5], Zheli Liu[2], Qingkai Zeng[2]*

[1]University of Notre Dame, [2]College of Computer Science, Nankai University,
[3]Washington University in St. Louis, [4]Purdue University,
[5]University of Illinois Urbana-Champaign
zwang43@nd.edu, qingkai.zeng@nankai.edu.cn

Abstract

Large language model (LLM) personalization typically relies on modeling each user in isolation, conditioning on their historical interactions to adapt model behavior. However, this user-centric formulation overlooks the *collective knowledge* shared across users, limiting generalization for users with sparse histories and amplifying overfitting for those with highly skewed behaviors. We argue that effective personalization requires leveraging both individual preferences and population-level patterns. To this end, we propose **LoGo**, a **Lo**cal–**Glo**bal knowledge framework that augments user-specific signals with a global knowledge encoding collective behavioral trends. LoGo models global knowledge through a temporally evolving process that captures how population-wide preferences change over time, and a community-aware structure that organizes users into coherent groups with shared interests. To balance potentially conflicting local and global signals, LoGo employs a mediator module that adaptively fuses the two knowledge sources. Experiments on five personalization benchmarks show that LoGo consistently enhances personalization quality, outperforming existing methods by improving generalization in users with limited histories and mitigating bias in users with abundant histories. These results demonstrate the central role of collective knowledge in advancing LLM personalization. Our code is publicly available at https://github.com/Zehong-Wang/LoGo.

1 Introduction

Large language models (LLMs) have become the backbone of modern AI systems, supporting a wide range of applications such as search, recommendation, education, productivity, and open-ended assistance (Zhao et al., 2023). Their strength lies in their ability to generalize across tasks. However, as LLMs are increasingly deployed in user-facing settings, general competence alone is often insufficient. Many applications require models to adapt their behavior to individual users in order to reflect personal preferences, goals, and interaction histories (Chen et al., 2024a; Raji et al., 2024; Chen et al., 2024b; Deldjoo et al., 2024). Consequently, personalization has emerged as a central requirement across domains, including e-commerce (AG et al., 2024), entertainment (Trifts and Aghakhani, 2019), education (Tetzlaff et al., 2021), and productivity tools (Kim et al., 2019).

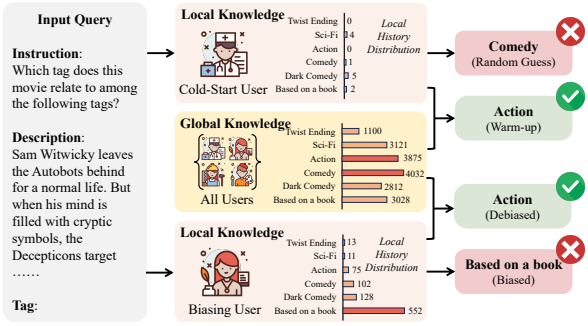


Figure 1: **Global knowledge improves personalization by addressing data sparsity and bias.** For cold-start users, it compensates for limited local history with population-wide knowledge. For users with skewed histories, it helps correct biased predictions by aligning to boarder user behaviors.

LLM personalization typically aims to adapt model behavior to a specific user by leveraging that user's past interactions as contextual signals. Most existing approaches adopt a single-user perspective, modeling each individual in isolation. One line of work treats a user's interaction history as textual evidence and personalizes the model through retrieval-augmented generation, where relevant past utterances are retrieved and fed to the LLM at inference time (Salemi et al., 2023, 2024; Richardson et al., 2023; Zhuang et al., 2024). In parallel, parameter-efficient fine-tuning (PEFT) en-

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 29555–29569

ables explicit user-level adaptation; for example, One PEFT Per User (OPPU) (Tan et al., 2024b) assigns each user a dedicated lightweight adapter to capture personalized preferences.

Although these personalization methods have proven effective, relying solely on the history of an individual user introduces two persistent challenges (Figure 1). First, users with limited interaction data suffer from a *cold-start* problem, as the model lacks sufficient signals to infer their preferences. Second, for users with rich but highly skewed histories, personalization can become *overly biased*, causing the model to overfit idiosyncratic behaviors and ignore generalizable patterns. We argue that these challenges arise from a common root: current approaches ignore to capture the *collective knowledge* shared across users. This collective knowledge includes social norms, shared preferences, and common behavioral patterns across the user population. It acts as a stable foundation that supports the model in two key ways: when a user has little interaction data, it helps fill in the gaps; when a user's history is extensive but unbalanced, it helps correct for potential biases. As illustrated in Figure 2, this collective knowledge (global knowledge) naturally complements individual interaction traces (local knowledge), enabling more robust, balanced, and principled personalization, even outperforming the state-of-the-art baselines.

In this work, we investigate how collective knowledge, derived from the interactions of many users, can be leveraged to enhance LLM personalization. While population-level signals naturally complement individual user histories, incorporating such knowledge introduces two key design challenges. *(1) Modeling collective knowledge.* Collective knowledge naturally reflects both its temporal evolution and user-level structural patterns. Treating it as a single static representation fails to capture two important aspects: the temporal dynamics of changing population preferences over time(Wu et al., 2017), and the latent structural patterns shared across users. To be effective, global knowledge should reflect both temporal changes and user relationships. *(2) Balancing global and local signals.* Although collective knowledge provides priors at the population level, personalization requires preserving individual preferences that may differ from dominant trends (Zheng et al., 2021). Overemphasizing global patterns dilutes personalization, while relying solely on local histories exacerbates cold-start and overfitting issues (Hechter and Opp, 2001). Therefore, collective knowledge should inform personalization without overriding individual preferences.

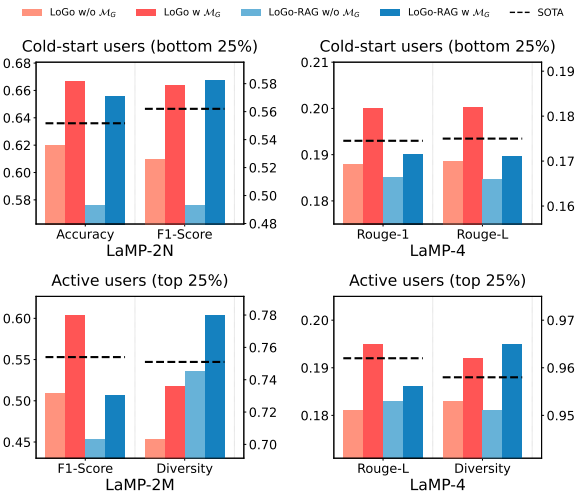


Figure 2: **Empirical effect of global knowledge $\mathcal{M}_G$ on personalization.** Adding collective knowledge improves prediction quality for cold-start users (top row) and increases output diversity for highly active users (bottom row), demonstrating that global knowledge benefits both underrepresented and over-specialized users. These benefits enable our model achieve better performance than the state-of-the-art baseline. We provide the detailed analysis in Appendix A.

To address these challenges, we introduce **LoGo**, a **Lo**cal–**Glo**bal knowledge framework designed to jointly model individual preferences and population-level knowledge for LLM personalization. In particular, LoGo constructs a *local knowledge* that encodes user-specific behaviors alongside a complementary *global knowledge* that captures collective knowledge distilled from many users. In modeling global knowledge, LoGo incorporates both the temporal evolution of collective behavior and the structural relations among users, allowing the framework to capture how shared patterns emerge, propagate, and organize across the user population. To manage the inherent tension between local and global memories, LoGo incorporates a *mediator* that dynamically balances local and global signals, ensuring that global knowledge provides principled guidance without overwhelming personalization. As a general framework, LoGo can be instantiated in both parametric (white-box) and non-parametric (black-box) personalization settings. Experiments on five personalization benchmarks demonstrate that incorporating collective knowledge yields consistent gains. Furthermore, as shown in Figure 2, it improves generalization

for cold-start users, reduces overfitting for highly active users, ultimately enabling more robust and nuanced personalization.

2 Related Work

Personalization of LLMs. Methods for LLM personalization can be broadly categorized into *non-parametric* and *parametric* approaches. Non-parametric methods personalize outputs by conditioning on user-specific information without modifying model parameters. Prior work has shown that treating a user's interaction history as in-context examples enables LLMs to produce personalized outputs across diverse tasks (Zhiyuli et al., 2023; Kang et al., 2023; Wang et al., 2023b; Kim and Yang, 2025). Retrieval-augmented prompting (Salemi et al., 2023, 2024; Mysore et al., 2023; Li et al., 2023) and profile-augmented prompting (Richardson et al., 2023; Sun et al., 2024, 2025; Tan et al., 2025a; Liu et al., 2025) further improve personalization by selecting or summarizing relevant past interactions. Beyond retrieval, several works design long-term knowledge systems (Packer et al., 2023; Wang et al., 2023a; Wei et al., 2025; Xu et al., 2025) to better capture and organize user-specific context over extended interaction sessions. Parametric approaches explicitly adapt model weights to encode individual preferences. OPPU (Tan et al., 2024b) fine-tunes per-user adapters using LoRA (Hu et al., 2021), enabling lightweight, user-specific customization. Other lines of work explore model merging for personalized alignment (Jang et al., 2023), personalized RLHF (Park et al., 2024; Li et al., 2024), and user-specific reward modeling (Cheng et al., 2023). While these methods capture fine-grained user preferences, they treat each user in isolation and overlook the collective knowledge shared across users.

Collective knowledge. Collective knowledge has long been recognized as a powerful signal in personalized systems. For example, collaborative filtering (Schafer et al., 2007; He et al., 2017) exploits shared behavioral structure to improve recommendation quality. Yet, despite its success in traditional personalization domains, collective knowledge remains relatively underexplored in LLM personalization. Recent work has begun to incorporate shared components across users as a proxy for collective knowledge. PER-PCS (Tan et al., 2024a) aggregates adapters from related users to enrich user representations, but it struggles when population-level patterns conflict with individual preferences. HYDRA (Zhuang et al., 2024) fine-tunes a shared backbone to capture global trends and uses user-specific heads for personalization, while P2P (Tan et al., 2025b) trains hypernetworks on high-resource users to generate personalized adapters for others. However, these two approaches typically treat collective knowledge as static and homogeneous, without accounting for how population behaviors evolve over time or how users cluster into coherent communities.

Our LoGo addresses these limitations by modeling evolving and community-aware global knowledge, and by introducing a mediator that reconciles global trends with individual user preferences.

3 LoGo for LLM Personalization

LLM personalization aims to adapt model outputs to the unique behaviors and preferences of a user based on their past interactions. In this section, we propose LoGo that incorporates not only a user's individual history but also the collective knowledge shared across the user population, addressing the cold-start and biasing issues inherent in purely user-centric personalization. An overview of the framework is shown in Figure 3.

Problem Definition. For user $u \in \mathcal{U}$, we define the user's interaction history as $\mathcal{H}_u = \{h_i\}_{i=1}^n$, where each record $h_i = (q_i, r_i, t_i)$ is a timestamped query-response pair. The goal of personalization is to condition the model on both the current input q and the user's history $\mathcal{H}_u$ to produce response r. In LoGo, we parameterize this conditioning through a user-specific adapter Φ_u composed of two components: a *local knowledge* $\mathcal{M}_L$ derived from the user's own history, and a *global knowledge* $\mathcal{M}_G$ that encodes collective knowledge across users.

3.1 Modeling Local Knowledge

Local knowledge in LoGo captures the fine-grained behavioral patterns that are specific to an individual user. For a user u, we define the local knowledge at time t as the set of all interactions occurring prior to the current query:

$$\mathcal{M}_L(t) = \mathcal{H}_u^{<t} = \{(q_i, r_i, t_i) \in \mathcal{H}_u \mid t_i < t\}. \quad (1)$$

This knowledge forms the foundational signal for personalization, as it reflects the user's historically expressed intentions, writing style, preferences, and temporal behaviors.

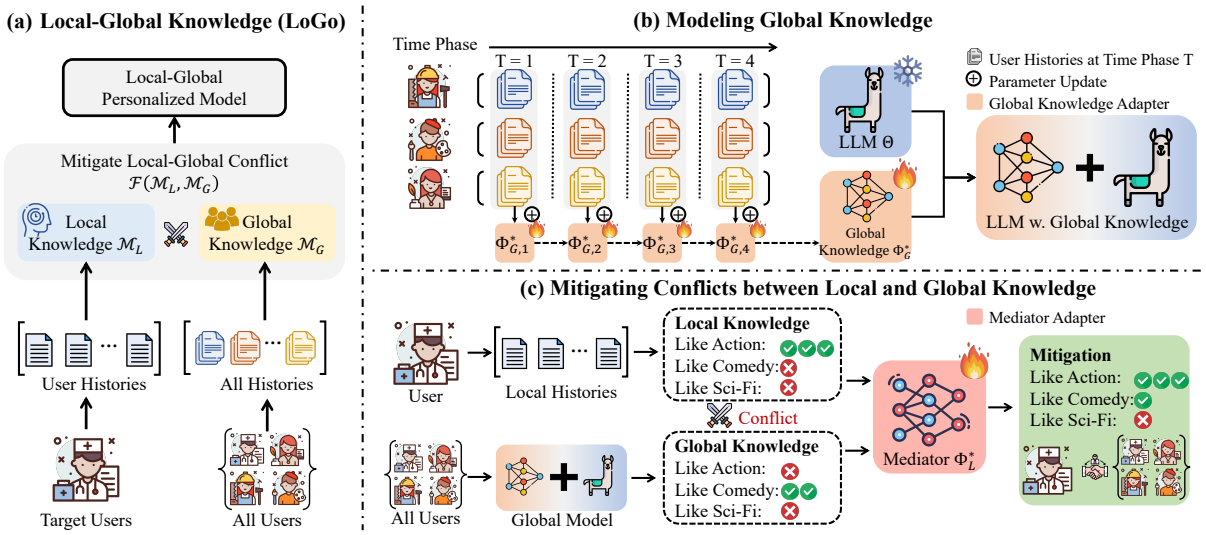


Figure 3: **Overview of the LoGo framework for LLM personalization.** (a) The model integrates local knowledge, which captures user-specific history, with global knowledge that encodes shared knowledge across users. (b) Global knowledge is updated over time through temporal phases to reflect evolving user interests. (c) A mediator module resolves conflicts between local and global signals, enabling balanced personalization and generalization.

A straightforward approach to leveraging local knowledge is to adapt the model directly on the user's past interactions. Following OPPU-style personalization (Tan et al., 2024b), we finetune a user-specific adapter Φ^u_L on the full history $\mathcal{H}^{<t}_u$ as the local knowledge by minimizing the task loss:

$$\Phi^u_L = \arg\min_{\Phi} \sum_{(q,r,t') \in \mathcal{H}^{<t}_u} \mathcal{L}(q, r; \Theta, \Phi), \quad (2)$$

where $\mathcal{L}$ is the task-specific objective (e.g., next-token prediction) and Θ is the frozen parameters of the base LLM. The module can be instantiated using adapters such as LoRA (Hu et al., 2021), prefix tuning (Li and Liang, 2021), or any other lightweight modules. This baseline exploits the entire local knowledge but ignores query-specific relevance: all past interactions contribute equally regardless of their utility for the current query.

Augmenting Local Knowledge via Retrieval and Profiling. Beyond the full-history finetuning baseline, LoGo enhances local knowledge with retrieval-augmented mechanisms that tailor the conditioning signal to the current query. In the basic retrieval-augmented variant, ϕ_{RAG} selects the top-k most relevant interactions from $\mathcal{H}^{<t}_u$, yielding a query-aware knowledge $\mathcal{M}^{\text{RAG}}_L(q) = \mathcal{R}(q, \mathcal{H}^{<t}_u, k)$ that captures fine-grained behavioral patterns most predictive of the user's desired response. Building on this, the profile-augmented variant (Richardson et al., 2023) incorporates an additional long-term summary of user behavior by generating a profile representation $p_u = \text{LLM}(\mathcal{H}^{<t}_u)$, which complements the retrieved examples with stable, high-level preference signals. The resulting local knowledge becomes $\mathcal{M}^{\text{PAG}}_L(q) = \mathcal{R}(q, \mathcal{H}^{<t}_u, k) \cup \{p_u\}$, making PAG a natural extension of RAG that integrates both query-specific relevance and global summaries of the user's historical patterns.

3.2 Modeling Global Knowledge

While local knowledge captures user-specific behaviors, effective personalization also requires access to population-level regularities that generalize across users. We refer to this collective signal as the *global knowledge* $\mathcal{M}_G$, which encodes patterns distilled from interaction histories of the entire user base. A natural way to instantiate global knowledge is to pool all user interactions into a single dataset $\mathcal{H}^{\text{all}} = \bigcup_{u \in \mathcal{U}} \mathcal{H}_u$ and train a global adapter on this aggregate supervision. Concretely, we optimize

$$\Phi^*_G = \arg\min_{\Phi_G} \sum_{(q,r) \in \mathcal{H}^{\text{all}}} \mathcal{L}(q, r; \Theta, \Phi_G), \quad (3)$$

where Θ denotes the frozen base model parameters. Although this pooled approach provides a simple and implementation-agnostic construction of global knowledge, it overlooks two fundamental properties of real-world user populations. First, collective behavior evolves over time, making a static global representation insufficient to reflect

changing trends. Second, users naturally form latent communities with distinct behavioral patterns, and collapsing all users into a single population obscures this structural diversity. These limitations motivate modeling global knowledge in a manner that is temporally adaptive and community-aware.

Temporal Dynamics. Population behavior shifts with external events, collective preferences, and long-term trends, making a static global knowledge insufficient. To model these temporal patterns, we divide the full interactions into T non-overlapping time periods, $\{\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \ldots, \mathcal{H}^{(T)}\}$, where $\mathcal{H}^{(t)}$ contains data from all users during period t. For each period, we learn a global knowledge module $\Phi_{G,t}$, warm-started from the previous one:

$$\Phi^*_{G,t} = \arg \min_{\Phi_{G,t} \leftarrow \Phi^*_{G,t-1}} \sum_{(q,r) \in \mathcal{H}^{(t)}} \mathcal{L}(q, r; \Theta, \Phi_{G,t}). \tag{4}$$

This sequential optimization naturally incorporates recency, smooths out noise in short-term fluctuations, and enables the global knowledge to adapt as collective behavior evolves. The final temporal global knowledge is given by $\Phi^*_{G,T}$.

Structural Diversity. Beyond temporal change, user populations are heterogeneous: different subgroups exhibit distinct interests and interaction patterns. To capture this structural organization, LoGo introduces *community-aware* global knowledge. We compute a profile vector for each user,

$$\boldsymbol{\rho}_u = \frac{1}{|\mathcal{H}_u|} \sum_{(q,r) \in \mathcal{H}_u} (\mathrm{emb}(q) \oplus \mathrm{emb}(r)), \tag{5}$$

collect the set $\mathcal{P} = \{\boldsymbol{\rho}_u\}_{u \in \mathcal{U}}$, and apply k-means clustering to partition users into K communities $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$. For each community, we learn a community-specific global knowledge module $\Phi^k_{G,t}$ using only the interactions within that subgroup. These modules capture finer-grained collective behavior that would otherwise be obscured by treating the population as homogeneous.

Incorporating temporal dynamics and structural diversity yields a global knowledge that is both time-sensitive and community-aware. Unless otherwise noted, LoGo uses this global knowledge as the default and denotes it simply as $\mathcal{M}_G$, with each user assigned to its corresponding community-specific module. This enriched representation complements local knowledge and provides reliable population-level guidance during personalization.

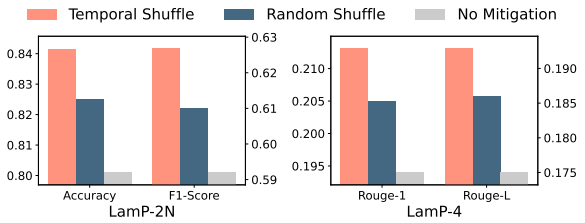


Figure 4: **Temporally aware finetuning outperforms random shuffling, and both outperform no mitigation.** This shows that respecting temporal structure better aligns local adaptation with global knowledge and prevents instability from unordered histories.

3.3 Mitigating Local-Global Conflicts

Local knowledge $\mathcal{M}_L$ encodes highly personalized behavioral signals, while global knowledge $\mathcal{M}_G$ captures broad population-level patterns. Although both are essential for effective personalization, they may provide contradictory guidance—particularly when a user's behavior deviates from dominant trends. Moreover, the non-stationarity of population behavior motivates a mediator that is both hierarchical and temporally aligned with the chronology used to learn $\mathcal{M}_G$. A naïve approach that simply applies local and global adapters jointly fails to resolve such conflicts: the two sets of parameters may interfere with each other, leading to unstable or biased personalization, as shown in Figure 4.

To mitigate this issue, LoGo adopts a hierarchical adaptation strategy. We first construct a global model by combining the frozen base model Θ with the learned global knowledge parameters Φ^*_G. This global model serves as the initialization for user-specific personalization. For each user u, we then learn a lightweight local adapter Φ^u_L on top of the global model rather than on the base model alone. Formally, the personalized parameters for user u are obtained by optimizing

$$\Phi^{u*}_L = \arg\min_{\Phi^u_L} \sum_{(q,r,t) \in \mathcal{H}_u} \mathcal{L}\big(q, r; \Theta, \Phi^*_G, \Phi^u_L\big), \tag{6}$$

which ensures that local adaptation is always grounded in the global representation. This formulation naturally reconciles conflicts: global knowledge acts as a stabilizing prior, while the user-specific adapter captures fine-grained deviations without distorting population-level knowledge.

At inference time, LoGo applies the personalized model composed of the base parameters Θ, the global knowledge Φ^*_G, and the learned local adapter Φ^{u*}_L. Given a query q from user u, the

Method	LaMP 2N		LaMP 2M		LaMP 3		LaMP 4		LaMP 5		A.R.
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	MAE ↓	RMSE ↓	R-1 ↑	R-L ↑	R-1 ↑	R-L ↑	
Base Model	0.791	0.536	0.542	0.503	0.277	0.543	0.203	0.183	0.522	0.457	11.1
Base Model + Rand.	0.802	0.577	0.569	0.506	0.232	0.530	0.208	0.186	0.520	0.465	9.3
OPPU	0.807	0.606	0.702	0.612	0.218	0.467	0.216	0.195	0.519	0.465	6.3
OPPU + RAG	0.817	0.580	0.588	0.524	0.223	0.509	0.205	0.185	0.521	0.468	8.3
OPPU + PAG	0.814	0.608	0.581	0.528	0.250	0.582	0.208	0.187	0.513	0.453	9.5
HYDRA	0.780	0.401	0.540	0.458	0.400	0.747	0.178	0.169	0.434	0.372	13.5
PER-PCS	0.804	0.539	0.679	0.583	0.251	0.494	0.205	0.185	0.520	0.471	8.7
P2P	0.716	0.613	0.442	0.408	0.383	0.670	0.160	0.145	0.490	0.431	12.3
LoGo (c = 1)	0.824	0.611	0.710	0.630	0.188	0.453	**0.216**	**0.196**	0.516	0.470	5.0
LoGo + RAG (c = 1)	0.828	0.614	0.724	0.642	**0.161**	**0.443**	0.214	0.194	0.526	0.477	3.6
LoGo + PAG (c = 1)	0.842	0.627	0.724	0.644	0.196	0.518	0.213	0.193	0.527	0.479	4.5
LoGo + PAG (c = 5)	**0.854**	**0.650**	0.739	0.668	0.186	0.488	0.212	0.193	0.536	0.481	3.0
LoGo + PAG (c = 10)	0.852	0.649	**0.741**	**0.689**	0.193	0.448	0.199	0.184	**0.538**	**0.483**	3.8
LoGo + PAG (c = 20)	0.849	0.640	0.740	0.673	0.248	0.487	0.211	0.192	0.516	0.472	5.2
Improvement	4.53% ↑	6.91% ↑	5.56% ↑	12.58% ↑	26.15% ↑	5.14% ↑	0.00% ↑	0.51% ↑	3.07% ↑	3.21% ↑	-

Table 1: **Experimental results of LoGo using LLaMA 3.1–8B.** R-1 and R-L denote ROUGE-1 and ROUGE-L. ↑ indicates that higher values are better, while ↓ indicates that lower values are preferred. **Bold** marks the best performance, and green denotes the improvement of the best LoGo result over the strongest baseline. A.R. represents the average ranking across all metrics.

model generates the response using

$$\hat{r} = f\left(q; \Theta, \Phi_G^*, \Phi_L^{u*}\right), \tag{7}$$

where the global knowledge provides population-level priors and the local adapter injects user-specific adjustments. This combined parameterization ensures that inference faithfully reflects both general trends and individual preferences.

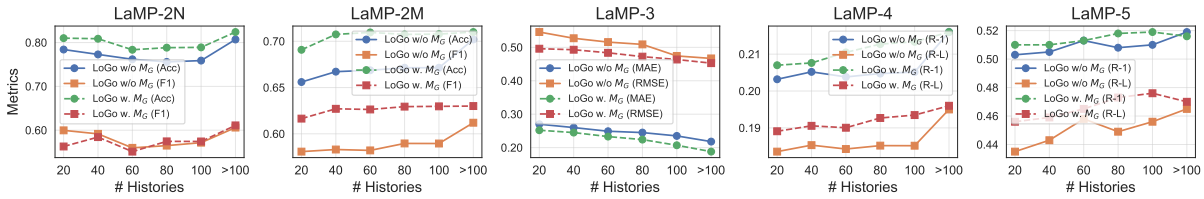
Temporal-Aware Mediator. Because the global knowledge Φ_G^* is learned through a temporally structured procedure, local finetuning should respect the same chronology. Directly training the local adapter on a shuffled or unordered history can misalign personalization with the temporal patterns already encoded in the global knowledge and overweight recent idiosyncrasies. To maintain coherence, we adopt a temporally aware finetuning strategy in which the user's interaction history is processed in chronological order. Let $(q_1, r_1, t_1) \prec (q_2, r_2, t_2) \prec \cdots \prec (q_n, r_n, t_n)$ denote the user's ordered history. We initialize the local adapter from the global model and update it sequentially. As shown in Figure 4, this chronological finetuning ensures that long-term behavioral patterns are learned first, while later updates capture more recent preferences without overriding globally consistent behaviors.

4 Experiments

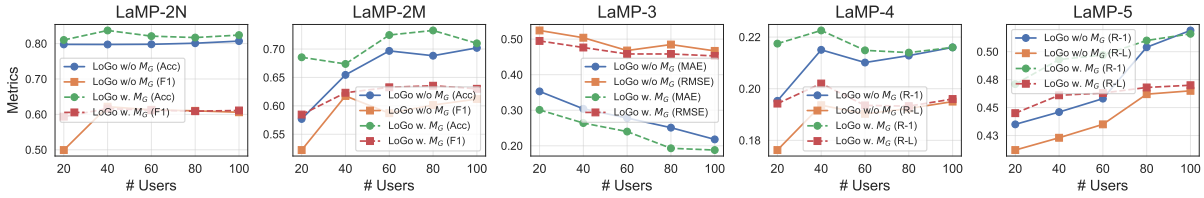
We evaluate our LoGo using the LaMP benchmark (Salemi et al., 2023), covering classification, regression, and text generation tasks, using task-specific metrics: Accuracy/F1 for classification, MAE/RMSE for regression, and ROUGE-1/L for text generation. We consider both white-box and black-box settings using LLaMA 3.1 8B (Dubey et al., 2024) and Claude 3.7 as the respective backbone models. Following Tan et al. (2024b), we evaluate performance on the top 100 most active users, identified based on the length of their interaction histories, while the remaining users are used to train the base LLM. For retrieval-based methods, we use BM25 (Robertson et al., 2009) with one retrieved item by default for consistency.

4.1 Main Results

To demonstrate the effectiveness of LoGo, we conduct experiments across a broad spectrum of baselines, including: the base model and its variant with randomly retrieved items; OPPU (Tan et al., 2024b), as well as retrieval-augmented variants such as RAG (Gupta et al., 2024) and PAG (Richardson et al., 2023); PER-PCS (Tan et al., 2024a), HYDRA (Zhuang et al., 2024), and P2P (Tan et al., 2025b) that leverage shared knowledge to improve the generalization. For our LoGo, we consider two settings: the *base version* and the *community version*. In the base version, we evaluate three variants: the base model, a RAG version, and a PAG version. In the community version, we use the PAG variant as the base and vary the number of clusters to assess performance under dif-



(a) Performance when limiting the maximum number of histories per user.



(b) Performance when varying the number of users used to construct the global knowledge.

Figure 5: **Generalization performance under varying data availability.** LoGo demonstrates strong generalization even with limited histories (a), and maintains robust with fewer users contributing to the global knowledge (b).

ferent community sizes. The experimental results are shown in Table 1.

Effectiveness of LoGo. Overall, personalized methods consistently outperform non-personalized ones, demonstrating the effectiveness of personalization. Our LoGo further improves performance over existing personalization baselines by leveraging global knowledge that captures shared knowledge across users as well as mitigates the conflict between the personalized and collective interests.

Effect of Community Clusters. To evaluate the impact of community-aware global knowledge, we vary the number of user clusters $c \in \{5, 10, 20\}$. Performance improves significantly when increasing c from 1 to 5 or 10, suggesting that finer-grained user grouping helps capture more meaningful shared knowledge. However, setting $c = 20$ leads to diminishing or inconsistent gains, likely due to over-fragmentation and reduced generalization. These results indicate that incorporating community-level global knowledge enhances performance by effectively modeling fine-grained shared knowledge, but an optimal cluster size is critical to balance specificity and generalizability.

4.2 Generalization Analysis

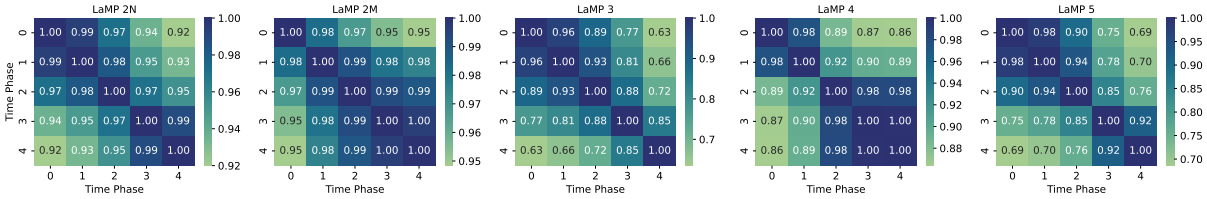
The effectiveness of global knowledge in mitigating cold-start issues and reducing bias for high-activity users suggests that it captures generalizable patterns beyond individual user histories. To further evaluate this generalization capability, we examine performance under two constrained settings: (1) limiting the number of historical interactions per user, and (2) restricting the number of users contributing to global knowledge. The results are presented in Figure 5.

Limited History Setting. We evaluate model performance when restricting each user to a maximum of n past interactions. Unlike the cold-start scenario, this setting introduces a temporal distribution shift, as earlier interactions may not align well with the current query context. As shown in Figure 5a, global knowledge significantly enhances performance when history is sparse (e.g., $n = 20$ or $n = 40$), effectively compensating for the lack of local signals. These results underscore the generalization capability of global knowledge in underdetermined personalization settings.

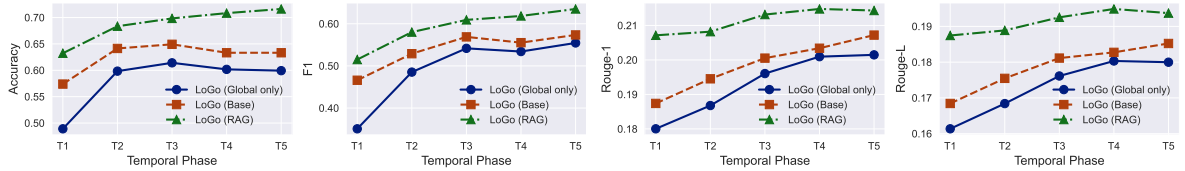
Limited User Setting. We assess model robustness when limiting the number of users contributing to global knowledge construction. As shown in Figure 5b, the performance gains from incorporating global knowledge are most pronounced when user coverage is low. This suggests that the learned global knowledge remains transferable even with limited population diversity. However, as the number of contributing users increases, the performance improvements begin to plateau, indicating that overly broad aggregation may introduce noise or dilute informative signals. These findings highlight the importance of designing fine-grained strategies for global modeling.

4.3 Temporal Distribution Shift Analysis

The global knowledge captures temporal information, which introduces temporal distribution shift.



(a) Cosine similarity between global memories across different time phases.



(b) Model performance on LaMP-2M.

(c) Model performance on LaMP-4.

Figure 6: **Performance under temporal distribution shift.**

Setting	Variant	LaMP-2N		LaMP-2M		LaMP-3		LaMP-4		LaMP-5	
		Acc ↑	F1 ↑	Acc ↑	F1 ↑	MAE ↓	RMSE ↓	R-1 ↑	R-L ↑	R-1 ↑	R-L ↑
Time Splits	T = 1	0.807	0.602	0.702	0.619	0.297	0.523	0.201	0.182	0.504	0.465
	T = 5	0.842	0.627	0.724	0.644	**0.196**	**0.518**	0.213	0.193	**0.527**	**0.479**
	T = 10	**0.859**	**0.638**	**0.737**	**0.656**	0.204	0.534	**0.214**	**0.196**	0.524	0.478
Retrieved Items	k = 1	0.842	0.627	0.724	0.644	0.196	0.518	0.213	0.193	0.527	0.479
	k = 2	0.850	0.636	0.727	0.651	**0.192**	**0.515**	**0.221**	**0.197**	0.537	0.484
	k = 4	**0.856**	**0.642**	**0.733**	**0.654**	0.207	0.530	0.206	0.189	**0.541**	**0.488**

Table 2: **Hyper-parameter analysis on key design choices in the LoGo.** We evaluate (1) the number of temporal splits used to update global knowledge (default T=5), and (2) the number of retrieved history items (default k=1).

We analyze the impact of such distribution shift on model performance.

Evolution of Global Knowledge. To evaluate the stability of global knowledge over time, we compute the pairwise cosine similarity between global knowledge parameters $\Phi^*_{G,t}$ across different temporal phases. High similarity indicates slow evolution and stable preferences, whereas low similarity reflects rapidly shifting global trends. As shown in Figure 6a, datasets such as LaMP-2N, LaMP-2M, and LaMP-4 exhibit stable dynamics, with similarity consistently above 0.85 across phases. In contrast, LaMP-3 and LaMP-5 show more rapid evolution, particularly in later phases, suggesting greater temporal drift in user behavior.

Robustness to Temporal Shift. We further assess model robustness by evaluating performance across temporal phases (T1–T5). As shown in Figures 6b and 6c, models relying solely on global knowledge exhibit performance degradation in earlier phases due to distributional mismatch. In contrast, combining global and local knowledge yields consistently strong results, even under severe temporal shifts. This indicates that local knowledge effectively grounds global knowledge, enabling the model to adapt to evolving user interests over time.

4.4 Hyper-parameter Analysis

Temporal Resolution. As shown in Table 2, we examine the impact of varying the number of time splits $T \in \{1, 5, 10\}$ used to update the global knowledge. Increasing T consistently improves performance by capturing more fine-grained temporal dynamics. Notably, setting $T = 1$ leads to a substantial performance drop, showing the importance of modeling evolving user preferences over time.

Number of Retrieved Items. As shown in Table 2, we examine the impact of varying the number of retrieved local history items, $k \in \{1, 2, 4\}$. We observe consistent performance improvements as k increases, indicating that incorporating more relevant local context enhances the model's ability to reconcile global and personal preferences. These results underscore the value of retrieval-based local

knowledge in supporting effective personalization.

5 Conclusion

In this paper, we investigate LLM personalization from a new perspective—bridging individual preferences with collective user knowledge. We introduce a local-global memory framework (LoGo) that combines user-specific local memory with a global memory capturing shared patterns across users. Extensive experiments across multiple benchmarks show that our framework effectively enhances personalization quality, particularly by addressing cold-start limitations and mitigating user-specific bias.

Limitations

Our study has two primary limitations stemming from the available data resources. First, we evaluate LoGo on a single model and a single personalization task. While our framework is inherently capable of supporting multi-task and multi-domain personalization—since both local and global knowledge modules are task-agnostic—the lack of diverse benchmarks restricts our empirical coverage. Future work can assess LoGo across a broader range of tasks, models, and evaluation settings to more fully characterize its generality. Second, our construction of global knowledge uses interaction histories from only 100 users, reflecting the scale constraints of the underlying dataset. Although LoGo is designed to benefit from substantially larger user populations, increasing the number of users would require only substituting a richer dataset without modifying the method itself. Despite this limitation, our experiments demonstrate that LoGo effectively captures meaningful temporal and structural patterns even in small-scale settings, serving as a strong proof of concept for population-level personalization.

Ethical Considerations

Data Privacy. Training the global knowledge requires aggregating user interaction histories at a central node, which may expose sensitive or personally identifiable information. Although our experiments rely on publicly available data, real-world deployments would need to address the risks associated with collecting, storing, and processing user-level histories. Techniques such as federated learning (Li et al., 2020), secure multiparty computation (Lindell, 2020), differential privacy (Dwork, 2006), or encrypted model updates (Li et al., 2015) could help mitigate these risks by enabling global model training without revealing raw user data. Ensuring transparent data governance and obtaining informed user consent remain crucial steps when deploying personalization systems at scale.

Accessibility and Computational Fairness. LoGo's design assumes that local knowledge is trained on user-side devices or local nodes. However, users vary widely in computational capacity, and personal devices may struggle to support fine-tuning—even lightweight adapters—when histories grow large. This disparity risks creating unequal access to personalization quality. To promote accessibility, future implementations may require more efficient on-device learning methods (Tan et al., 2025b) or server-assisted personalization protocols (Bicakci and Baykal, 2004; He et al., 2023) that preserve privacy while reducing computational demands.

Scalability and Responsible Deployment. Because each user maintains a personalized local adapter, large-scale deployment could introduce challenges in managing, storing, and updating many user-specific models over time. Without careful design, this may increase system complexity and exacerbate maintenance burdens. Moreover, personalization systems can amplify behavioral biases or reinforce user-specific echo chambers if not properly monitored. Future work should explore strategies for scalable model management, mechanisms for auditing personalization behavior, and safeguards that ensure personalized systems remain fair, transparent, and aligned with user expectations.

References

Aishwarya Gowda AG, Hui-Kai Su, and Wen-Kai Kuo. 2024. Personalized e-commerce: Enhancing customer experience through machine learning-driven personalization. In *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, pages 1–5. IEEE.

Kemal Bicakci and Nazife Baykal. 2004. Server assisted signatures revisited. In *Cryptographers' Track at the RSA Conference*, pages 143–156. Springer.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024a. From persona to

personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

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

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

Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. 2024. A review of modern recommender systems using generative models (gen-recsys). In *Proceedings of the 30th ACM SIGKDD conference on Knowledge Discovery and Data Mining*, pages 6448–6458.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Aman Gupta, Anup Shirgaonkar, Angels de Luis Balaguer, Bruno Silva, Daniel Holstein, Dawei Li, Jennifer Marsman, Leonardo O Nunes, Mahsa Rouzbahman, Morris Sharp, and 1 others. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint arXiv:2401.08406*.

Lin He, Fuchang Li, Haikun Xu, Wenbo Xia, Xuefei Zhang, and Xiaofeng Tao. 2023. Blockchain-based vehicular edge computing networks: the communication perspective. *Science China Information Sciences*, 66(7):172301.

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

Michael Hechter and Karl-Dieter Opp. 2001. Social norms.

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

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 preprint arXiv:2310.11564*.

Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do llms understand user preferences? evaluating llms on user rating prediction. *Preprint*, arXiv:2305.06474.

Jaehyung Kim and Yiming Yang. 2025. Few-shot personalization of llms with mis-aligned responses. In *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)*, pages 11943–11974.

Young-Ho Kim, Eun Kyoung Choe, Bongshin Lee, and Jinwook Seo. 2019. Understanding personal productivity: How knowledge workers define, evaluate, and reflect on their productivity. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

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 preprint arXiv:2308.07968*.

Hongwei Li, Dongxiao Liu, Yuanshun Dai, Tom H Luan, and Shui Yu. 2015. Personalized search over encrypted data with efficient and secure updates in mobile clouds. *IEEE Transactions on Emerging Topics in Computing*, 6(1):97–109.

Li Li, Yuxi Fan, Mike Tse, and Kuo-Yi Lin. 2020. A review of applications in federated learning. *Computers & Industrial Engineering*, 149:106854.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

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

Yehuda Lindell. 2020. Secure multiparty computation. *Communications of the ACM*, 64(1):86–96.

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

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 preprint arXiv:2311.09180*.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. Memgpt: Towards llms as operating systems.

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

Mustafa Ayobami Raji, Hameedat Bukola Olodo, Timothy Tolulope Oke, Wilhelmina Afua Addy, Onyeka Chrisanctus Ofodile, and Adedoyin Tolulope Oyewole. 2024. E-commerce and consumer behavior: A review of ai-powered personalization and market trends. *GSC advanced research and reviews*, 18(3):066–077.

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

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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

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

J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. Collaborative filtering recommender systems. In *The adaptive web: methods and strategies of web personalization*, pages 291–324. Springer.

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

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R Fung, Hou Pong Chan, ChengXiang Zhai, and Heng Ji. 2024. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. *arXiv preprint arXiv:2402.11060*.

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

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 *EMNLP*.

Zhaoxuan Tan, Zinan Zeng, Qingkai Zeng, Zhenyu Wu, Zheyuan Liu, Fengran Mo, and Meng Jiang. 2025a. Can large language models understand preferences in personalized recommendation? *arXiv preprint arXiv:2501.13391*.

Zhaoxuan Tan, Zixuan Zhang, Haoyang Wen, Zheng Li, Rongzhi Zhang, Pei Chen, Fengran Mo, Zheyuan Liu, Qingkai Zeng, Qingyu Yin, and 1 others. 2025b. Instant personalized large language model adaptation via hypernetwork. *arXiv preprint arXiv:2510.16282*.

Leonard Tetzlaff, Florian Schmiedek, and Garvin Brod. 2021. Developing personalized education: A dynamic framework. *Educational Psychology Review*, 33(3):863–882.

Valerie Trifts and Hamed Aghakhani. 2019. Enhancing digital entertainment through personalization: The evolving role of product placements. *Journal of Marketing Communications*, 25(6):607–625.

Bing Wang, Xinnian Liang, Jian Yang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. 2023a. Enhancing large language model with self-controlled memory framework. *arXiv preprint arXiv:2304.13343*.

Danqing Wang, Kevin Yang, Hanlin Zhu, Xiaomeng Yang, Andrew Cohen, Lei Li, and Yuandong Tian. 2023b. Learning personalized story evaluation. *arXiv preprint arXiv:2310.03304*.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H Chi, and 1 others. 2025. Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory. *arXiv preprint arXiv:2511.20857*.

Le Wu, Yong Ge, Qi Liu, Enhong Chen, Richang Hong, Junping Du, and Meng Wang. 2017. Modeling the evolution of users' preferences and social links in social networking services. *IEEE Transactions on Knowledge and Data Engineering*, 29(6):1240–1253.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the web conference 2021*, pages 2980–2991.

Aakas Zhiyuli, Yanfang Chen, Xuan Zhang, and Xun Liang. 2023. Bookgpt: A general framework for book recommendation empowered by large language model. *arXiv preprint arXiv:2305.15673*.

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. *Advances in Neural Information Processing Systems*, 37:100783–100815.

A Cold-Start and Debiasing Analysis

A.1 Warming Up Cold-Start Users.

Figure 2 presents the evaluation results for inactive (cold-start) users, those in the bottom 25% based on the number of historical interactions, across both the LaMP-2M (classification) and LaMP-4 (text generation) tasks. We evaluate two variants of LoGo: the base version and the RAG-enhanced version. The results show that incorporating global knowledge $\mathcal{M}_G$ consistently enhances model performance, with improvements of up to 18.05%. These findings highlight the effectiveness of leveraging collective knowledge to improve cold-start performance, enabling LoGo to overcome sparse user-specific signals by drawing on shared behavioral patterns.

A.2 Mitigating Bias in High-Activity Users.

Figure 2 reports the experimental results on highly active users, defined as the top 25% of users with the richest historical interactions. This subgroup is particularly prone to prediction bias, often caused by overfitting to individualized behavior patterns. We evaluate both prediction performance and diversity ratio to demonstrate the debiasing capability of our LoGo model. In particular, we use the normalized entropy of each user's prediction distribution as a surrogate measure of diversity, with a detailed definition provided in later. The results demonstrate that using only local knowledge can lead to excessive personalization, limiting generalization and reducing output diversity. In contrast, augmenting the model with global knowledge $\mathcal{M}_G$ consistently improves both predictive performance and diversity. This suggests that global knowledge acts as a form of population-level regularization, helping the model balance between personalization and general trends, ultimately mitigating user-specific biases.

The Diversity Measurement To evaluate the biasing ratio of user predictions, we define a diversity measurement where a large diversity indicates low biasing. To this end, we define an entropy-based method to measure the diversity of the prediction results. Take classification tasks as an example, each user u is associated with a set of predicted labels from a classifier:

$$L_u = \{\ell_1, \ell_2, \ldots, \ell_m\}, \quad \ell_i \in \{1, 2, \ldots, n\}$$

From this, we define a discrete empirical distribution $P_u = \{p_{u,1}, p_{u,2}, \ldots, p_{u,n}\}$, where:

$$p_{u,i} = \frac{\text{count of label } i \text{ in } L_u}{m}$$

That is, $p_{u,i}$ is the relative frequency of label i among the m predictions for user u, and $\sum_{i=1}^{n} p_{u,i} = 1$. Given a distribution P_u, we define the diversity using normalized entropy as:

$$Div(P_u) = \frac{H(P_u)}{H_{\text{max}}} = \frac{-\sum_{i=1}^{n} p_{u,i} \log_b p_{u,i}}{\log_b n},$$

where $H(P_u) = -\sum_{i=1}^{n} p_{u,i} \log_b p_{u,i}$ is the shannon entropy of the distribution and $H_{\text{max}} = \log_b n$ is the maximum possible entropy occurs when the distribution is uniform.

The range of $Div(P)$ is in $[0, 1]$, where a low values means the predictions are concentrated in a few classes, i.e., less diversity, and a high value means the predictions are spread across many classes, i.e., high diversity.

For text generation tasks, we cannot directly apply this metric since the output space is continuous. To adapt it, we first convert the generated texts into embeddings and then apply a clustering algorithm (e.g., k-means) to discretize the continuous space. This yields a discrete distribution over clusters for each user. Consequently, we obtain cluster-based distributions for the generated texts, analogous to class distributions in classification tasks, which can then be used to compute diversity via normalized entropy.

B Evaluation Setting

B.1 Dataset

We summarize the tasks below to clarify their input–output formats and personalization requirements. Dataset statistics are provided in Table 3.

Task in LaMP	Base LLM Training			Personal PEFT Training			
	#Q	L_{in}	L_{out}	#Q	#History	L_{in}	L_{out}
LamP-2N: News Classification	3,662	68.2	1.3	6,033	219.9	63.5	1.1
LamP-2M: Movie Tag Prediction	3,181	92.1	1.4	3,302	55.6	92.6	2.0
LamP-3: Product Rating	22,388	128.7	1.0	112	959.8	211.9	1.0
LamP-4: News Headline Generation	7,275	33.9	9.2	6,275	270.1	25.2	11.1
LamP-5: Scholarly Title Generation	16,075	162.1	9.7	107	442.9	171.6	10.3

Table 3: Dataset statistics extracted from the original table. #Q is the number of queries, L_{in} and L_{out} denote average input and output lengths, and #History is the number of historical interactions. The table is from Tan et al. (2024b).

- **LaMP-2N: Personalized News Classification.** This task evaluates whether a model can categorize news articles according to an individual user's reading and writing patterns. Given an article written by user u, together with that user's prior article–label history, the model must assign the article to one of 15 possible news categories.

- **LaMP-2M: Personalized Movie Tag Prediction.** Here, the goal is to infer which descriptive tag a user is likely to apply to a movie. The model receives a movie synopsis along with the user's past movie–tag interactions and must select the appropriate tag from a 15-class label set, reflecting the user's tagging preferences.

- **LaMP-3: Personalized Product Rating.** This task examines user-conditioned rating prediction. Given a review and the historical review–rating pairs written by user u, the model predicts the rating the user would assign, choosing from five discrete levels (1–5). The task can be viewed as either ordinal classification or personalized regression.

- **LaMP-4: Personalized News Headline Generation.** This task assesses style-sensitive generation. The model is given an article and a profile summarizing the user's prior article–headline pairs. Using these personal writing patterns, the model must generate a headline tailored to the target user's stylistic tendencies.

- **LaMP-5: Personalized Scholarly Title Generation.** Similar in structure to LaMP-4 but grounded in the academic domain, this task asks the model to produce a title for a given scholarly abstract or article. The model relies on the author's past article–title history to generate a title consistent with the user's preferred phrasing and stylistic conventions.

B.2 Training Details

We follow the training protocol of Tan et al. (2024b). The dataset is partitioned into two subsets: the top 100 most active users, which we reserve for personalization experiments, and the remaining users, which we use to train the general instruction-following model. We first fine-tune the base model on the non-top-100 portion to obtain robust instruction-following behavior. This fine-tuned model then serves as the initialization for constructing both the global knowledge and the user-specific local memories for the top 100 active users. We also adopt the prompt templates introduced by Tan et al. (2024b) for all personalization tasks; details are provided in Appendix H of their work.

C Black-Box LoGo Variant

Beyond its parametric formulation, LoGo can also be instantiated in settings where the underlying LLM is accessible only through an inference API and its parameters cannot be modified. In this black-box scenario, both local and global knowledge are incorporated through carefully designed prompts rather than learned adapters, allowing LoGo to convey personalized and population-level knowledge without requiring any finetuning of the base model.

C.1 Modeling Local Knowledge

In the non-parametric setting, local knowledge is injected into the black-box LLM through prompt engineering rather than trainable adapters. The goal is to supply the model with user-specific behavioral evidence that can guide its predictions without modifying model parameters. We begin by constructing a basic prompt that summarizes or exposes the relevant portion of the user's history. This typically takes the form of a short textual snippet describing the user's past behavior or a set of exemplar interactions. For illustration, the movie-tagging task may

include a prompt such as:

Modeling Local Knowledge (User Profile)

You will receive a personalized knowledge between the movies and their taggings along with a list of new movies and their descriptive tags. Update the knowledge by identifying the correlations between the movies and their tags, while preserving key insights from the original.

Personalized knowledge: `[updated personalized knowledge]`.

New movies with tags: `[A list of movies with tags]`

C.2 Modeling Global Knowledge

In addition to user-specific information, personalization benefits from recognizing population-level regularities that evolve over time. In the non-parametric setting, where model parameters cannot be modified, we represent global knowledge through external textual artifacts—such as summaries, exemplars, or aggregated user profiles—that encapsulate shared behavioral patterns across the user base. This global knowledge serves as a stable, population-wide prior that complements the personalized signals injected through local knowledge.

To construct this global knowledge, we adopt a time-structured procedure that mirrors the temporal dynamics present in real-world user behavior. We partition all interactions into T chronological segments,

$$\{\mathcal{H}^{(1)}, \mathcal{H}^{(2)}, \ldots, \mathcal{H}^{(T)}\},$$

where each segment corresponds to a distinct period of activity. At each phase t, we collect the user profiles or personalized memories generated during that period and prompt the black-box LLM to synthesize a population-level summary. This summary captures the dominant themes, preferences, and behavioral patterns emerging in phase t and is used to update the evolving global knowledge. For example:

Modeling Global Knowledge

You will receive a global knowledge and a set of personalized memories. Update the global knowledge by identifying patterns between paper titles and their most relevant references, while preserving key insights from the original. Output a bulleted list with 20 items.

Global knowledge: `[updated global knowledge]`.

Personalized memories: `[A list of personalized memories]`.

By iteratively aggregating these time-conditioned summaries, the resulting global knowledge $\mathcal{M}_G$ becomes an external, temporally aware representation of collective behavior. This knowledge can then be inserted into prompts during inference, enabling black-box LLMs to incorporate population-level context even without parametric finetuning.

C.3 Mitigating Local–Global Conflicts

Local knowledge $\mathcal{M}_L$ provides user-specific evidence, whereas global knowledge $\mathcal{M}_G$ reflects population-level tendencies. When these two sources disagree—such as when a user's behavior deviates from dominant trends—a black-box LLM must still produce a stable and personalized output. Because API-only models expose no trainable parameters, conflict mitigation must be achieved through prompt design rather than adaptive parameter updates.

To reconcile these signals, we introduce a *prompt-based mediator* that integrates both knowledge components into a unified context. Instead of modifying the underlying model, the mediator is implemented as a structured instruction guiding the LLM on how to balance personalization against global priors. The prompt explicitly presents the user's local evidence alongside the global summary and instructs the model to weigh them appropriately. For example:

Mitigating Local-Global Conflict

Here is the current user's knowledge: `[local knowledge]` with the additional user information: `[Retrieval Items]` Here is the global knowledge: `[global knowledge]`. You need to balance their contributions.

Which tag does this movie relate to among the following tags? Just answer with the tag name without further explanation. tags: [...]

description: `[movie descriptions]`

tag: []

To further strengthen the mediator, we optionally augment the prompt with additional retrieved examples. These elements help the model distinguish persistent user preferences from outliers and make the conflict resolution more robust. The mediator thus serves a similar conceptual purpose to the parametric mediator in the main LoGo framework, but operates entirely through textual conditioning.

Inference. Given a user query q_u, inference proceeds by instantiating the mediator prompt with

Method	LaMP-2N		LaMP-2M		LaMP-3		LaMP-4		LaMP-5	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	MAE ↓	RMSE ↓	R-1 ↑	R-L ↑	R-1 ↑	R-L ↑
Base Model	0.744	0.520	0.457	0.379	0.304	0.582	0.159	0.142	0.508	0.417
Base Model + Rand.	0.755	0.536	0.513	0.398	0.318	0.563	0.159	0.156	0.423	0.355
Base Model + RAG	0.762	0.553	0.513	0.420	0.259	0.543	0.177	0.159	0.507	0.427
Base Model + Profile	0.798	0.608	0.556	0.454	0.241	0.538	0.162	0.145	0.529	0.435
Base Model + Profile w. RAG	0.809	0.621	0.581	0.472	0.250	0.535	0.193	0.173	0.512	0.432
LoGo	0.757	0.537	0.504	0.409	0.290	0.562	0.186	0.167	0.524	0.421
LoGo + RAG	0.774	0.579	0.524	0.432	0.250	0.535	0.191	0.171	0.515	0.434
LoGo + Profile	0.822	0.652	0.589	0.484	0.286	0.567	0.182	0.163	**0.534**	**0.439**
LoGo + Profile w. RAG	**0.828**	**0.658**	**0.601**	**0.498**	**0.239**	**0.523**	**0.203**	**0.182**	0.532	0.427

Table 4: Experimental results of the white-box implementation of LoGo using Claude 3.7. R-1 and R-L denote ROUGE-1 and ROUGE-L, respectively. ↑ indicates that higher values are better, while ↓ indicates that lower values are preferred. **Bold** indicates the best results.

the constructed local knowledge $\mathcal{M}_L$ and global knowledge $\mathcal{M}_G$, then forwarding the resulting text to the black-box model:

$$\hat{r} = \text{LLM}\big(\mathcal{F}_{\text{prompt}}(q_u, \mathcal{M}_L, \mathcal{M}_G)\big) .$$

In this way, the final prediction reflects population-level regularities while remaining sensitive to the user's individual behavior, achieving conflict mitigation without any parametric adaptation.

C.4 Experimental Results

We evaluate the effectiveness of LoGo in a black-box setting, where both global and local knowledge are modeled externally via prompts without access to model parameters. The results are presented in Table 4, and several key observations emerge:

Personalization Improves Performance. Similar to the white-box case, the use of local knowledge, either via RAG or profile summaries, consistently outperforms the non-personalized base model and random baselines. This confirms that even without parameter tuning, black-box LLMs benefit from personalization through prompt augmentation.

Global Knowledge Offers Complementary Gains. Incorporating global knowledge via LoGo yields additional improvements. Specifically, LoGo variants that include global summaries outperform their base model counterparts across all metrics, indicating the value of shared population-level knowledge in supporting generalization.

Mediator Prompt Effectively Resolves Conflicts. Fusing local and global knowledge through a structured mediator prompt leads to the best results. The *LoGo + Profile w. RAG* variant achieves the strongest performance across nearly all benchmarks, outperforming other methods in both accuracy and generation quality (e.g., F1, ROUGE).

White-Box vs. Black-Box Trade-Off. While the black-box setting yields slightly lower overall performance than the white-box setting, the gap remains modest. This demonstrates that prompt-based knowledge integration is a practical and effective strategy when direct model access is unavailable.

Overall, these findings demonstrate the robustness of LoGo across different access regimes. Even under strict black-box constraints, LoGo delivers strong performance by leveraging personalized local signals and population-level trends, integrated via a prompt-based mediator.