

Beyond Isolated Behaviors: Hierarchical User Modeling for LLM Personalization

Liang Wang^{1*,3}, Xinyi Mou^{1*}, Xiaoyou Liu^{1,3}
Tiannan Wang³, Yuqing Wang³, Zhongyu Wei^{1,2†}

¹School of Data Science, Fudan University

²Shanghai Innovation Institute

³OPPO

liangwang25@m.fudan.edu.cn, {xymou20, zywei}@fudan.edu.cn

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse domains, yet personalizing their outputs to individual users remains an open challenge. Existing approaches predominantly adopt a flat behavioral paradigm, aggregating user behaviors without an explicit account of how they are organized into deeper behavioral structures. In this work, we draw on Pierre Bourdieu’s Theory of Practice to propose PHF (Practice-Habitus-Field), a sociologically grounded framework that reconceptualizes LLM personalization through three hierarchical levels: individual behaviors as practices, their temporal accumulation into stable dispositions as habitus, and shared regularities across similar users as fields. We instantiate PHF through PHF_{Compass}, a lightweight and model-agnostic implementation based on a frozen LLM. Experiments on the Language Model Personalization (LaMP) benchmark demonstrate consistent improvements across diverse tasks, while further analyses validate the interpretability and extensibility of the learned behavioral structures.¹

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across diverse tasks and domains, driving growing interest in personalized language generation, role-playing agents, and social simulation (Mou et al., 2026; Chen et al., 2024; Mou et al., 2024). These capabilities support applications including dialogue systems, recommendation, and healthcare interventions (He et al., 2025; Deng et al., 2024; Ning et al., 2025; Hebert et al., 2024; Deng et al., 2025; Li et al., 2024; Bao et al., 2024). In personalized settings, user expectations

¹Code repository: https://github.com/RaidonWong/PHF_code. Work completed during the internship of Liang Wang and Xiaoyou Liu at OPPO.

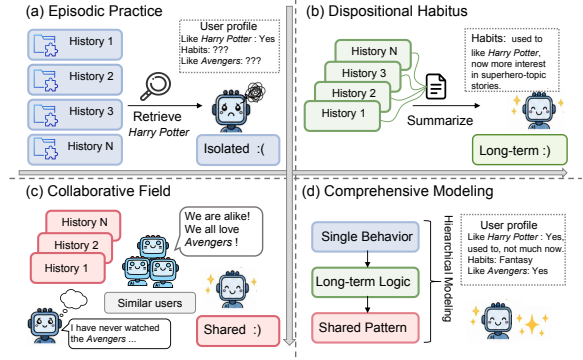


Figure 1: Comparison of personalization paradigms. (a) Existing methods adopt a flat behavioral paradigm, retrieving or encoding user behaviors without explicit structural modeling. (b) Habitus captures the temporal accumulation of repeated behaviors into stable dispositions. (c) Field captures shared behavioral regularities among users with similar dispositions. (d) PHF integrates practice, habitus, and field into a unified hierarchical framework for LLM personalization.

are inherently heterogeneous: responses appropriate for one user may not align with the expectations of another. Consequently, effective personalization requires models to capture the underlying behavioral patterns that shape user preferences and behaviors.

Recent approaches to LLM personalization have explored retrieval-augmented prompting, learned user representations, and parameter-efficient adaptation (Salemi et al., 2024a; Tan et al., 2024b; Liu et al., 2025). While these methods have shown promising results, they predominantly adopt a *flat* behavioral paradigm that treats user behaviors as an unordered collection of observed instances, as illustrated in Figure 1(a). This flat formulation gives rise to two key limitations: (1) **behavior fragmentation**: within each user, behavioral histories lack internal structure, where they are treated as isolated instances rather than being consolidated into stable long-term patterns. (2) **user isolation**: users are modeled independently, without leveraging shared

behavioral regularities across users with similar behavioral profiles, limiting generalization when user history is sparse. Together, these limitations make it difficult to capture the multi-level behavioral structures that govern user behavior, leading to suboptimal personalization.

Addressing these limitations requires a structured view of user behavior that models both how behaviors consolidate within individuals and how regularities are shared across users. We draw conceptual inspiration from Pierre Bourdieu’s Theory of Practice (Bourdieu, 1990, 1977). In particular, the theory motivates two design principles for personalization: recurring actions can be summarized into relatively stable individual dispositions, and supra-individual regularities can provide contextual signals beyond a single user’s history. Based on this perspective, we propose **PHF** (Practice-Habitus-Field), a hierarchical user modeling framework that addresses both limitations through a unified multi-level representation. Individual behaviors are abstracted as *practices* and consolidated into *habitus*, which represents long-term user tendencies (Figure. 1(b)); users with similar habitus are further associated with *fields* that encode shared behavioral regularities (Figure. 1(c)). Collectively, PHF formulates LLM personalization as a hierarchical behavioral modeling problem (Figure. 1(d)) that jointly captures within-user disposition formation and cross-user behavioral structure.

To implement PHF, we develop PHF_{Compass} (C**O**mbined M**O**deling of P**R**actice, A**C**cumulated S**T**ructure, and S**H**ared signals), a lightweight model compatible with diverse backbones. PHF_{Compass} represents practices by mapping individual behaviors into discrete semantic prototypes, filtering surface-level noise while preserving core behavioral semantics. These practices are then temporally aggregated into habitus representations that encode stable long-term behavioral dispositions for each user. To capture collective structure, PHF_{Compass} further organizes users into latent fields derived from their habitus, enabling users with similar dispositions to benefit from shared behavioral regularities. During inference, both habitus and field representations are provided to the LLM. This enables personalized generation that reflects both individualized dispositions and broader collective patterns. Our main contributions are summarized as follows:

(1) We propose **PHF** (Practice-Habitus-Field), a theory-inspired hierarchical framework for LLM

personalization. PHF captures both the temporal organization of behavior within users and shared regularities across behaviorally similar users, while explicitly treating its computational realization as replaceable.

(2) We develop PHF_{Compass}, a lightweight instantiation of PHF using residual behavioral abstraction, temporal aggregation, and collaborative clustering. The underlying LLM remains frozen, and alternative Habitus and Field implementations confirm that the framework is not tied to this particular pipeline.

(3) Experiments on six public LaMP tasks demonstrate strong performance across classification and generation. Further analyses validate the interpretability of the learned behavioral structures and support the utility of the proposed sociological perspective.

2 Related Work

2.1 LLM-based Personalization

Despite strong general capabilities in reasoning, knowledge-intensive tasks, and domain applications (Joshi et al., 2017; Liang et al., 2025; Wu et al., 2025; Jia et al., 2026), LLMs remain fundamentally limited in personalization. Since standard LLMs are trained on broadly aggregated corpora, they capture general linguistic and task patterns but contain limited information about individual behavioral tendencies, preferences, and long-term decision logic (Tao et al., 2025; Xi et al., 2025). Existing LLM personalization methods can be broadly characterized by how they utilize user behavioral histories. Retrieval-based methods personalize generation by selecting relevant behaviors from user history and incorporating them into prompts (Salemi et al., 2024b; Kumar et al., 2024; Shi et al., 2025; Zhang et al., 2024; Wang et al., 2025). While flexible, these approaches operate at the level of individual behaviors without aggregating them into a holistic user representation. Aggregation-based methods instead construct unified user representations through summarized preferences (Zhang, 2024), learned soft prompts or PEFT modules (Li and Liang, 2021; Zhuang et al., 2024; Zhang et al., 2025b), compressed user embeddings (Liu et al., 2025), or retrieval optimization (Salemi et al., 2024a). These methods merge behavioral signals into user-level representations but do so without explicitly modeling how behaviors are internally organized or how regularities

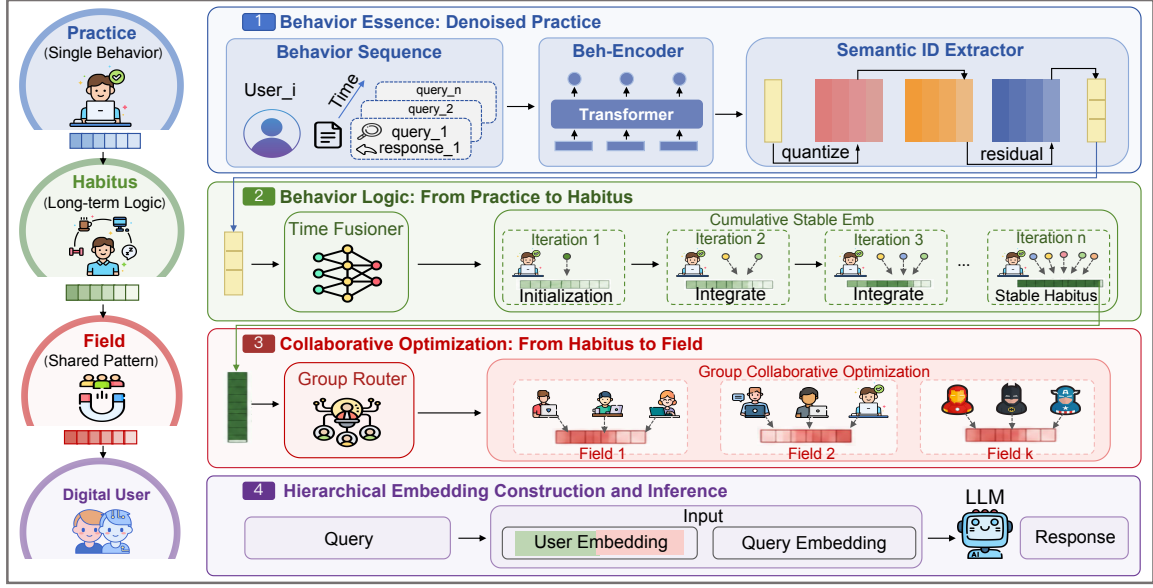


Figure 2: Overview of the PHF framework and the $\text{PHF}_{\text{Compass}}$ implementation. (1) Raw behaviors are abstracted into denoised practice essence via semantic ID extraction. (2) Practices are temporally aggregated into stable habitus representations. (3) Users are organized into latent fields through a group router. (4) Habitus and field embeddings are combined with the query embedding to condition the LLM for personalized generation.

are shared across users. In both cases, user behavior is treated as a flat collection of observed instances, leaving the within-user behavioral structure and cross-user behavioral regularity largely unaddressed.

2.2 User Embedding Modeling

Learning user representations from behavioral data has long been central to personalized systems (Elkahky et al., 2015; Purificato et al., 2024; Ko et al., 2022; Mu, 2018). Recent LLM personalization methods further model users through textual profiles, behavioral summaries, or trainable latent representations derived from interaction and dialogue histories (Hebert et al., 2024; Doddapaneni et al., 2024; Liu et al., 2025; Tang et al., 2024). Temporal weighting and semantic quantization techniques have also been introduced to better capture salient behavioral signals (Hossein-zadeh et al., 2026; Dai et al., 2022; Rajput et al., 2023; Zheng et al., 2024; Deng et al., 2025; Chen et al., 2025). CURP (Wang et al., 2026) models users with both a stable inner representation and a query-aware dynamic representation via PQ quantization, balancing the effectiveness and efficiency. Different from prior work that primarily models behaviors independently, our approach organizes user behaviors into hierarchical behavioral structures, capturing both stable individual dispositions and

collective regularities across users.

3 Method

3.1 PHF Framework Setup

We introduce PHF, a hierarchical user modeling framework for LLM personalization, illustrated in the left part of Figure. 2. PHF organizes personalization through three behavioral levels: *practice*, *habitus*, and *field*. Individual user behaviors are first treated as practices, which are progressively accumulated into habitus, stable behavioral dispositions that characterize each user’s long-term tendencies. Users with similar habitus are further organized into shared fields that capture collective behavioral regularities across behaviorally similar users. Personalized generation is conditioned on both habitus and field representations, yielding the following task formulation:

Let $\mathcal{U} = \{u_1, u_2, \dots, u_N\}$ denote a set of N users. Each user u_i is associated with a chronologically ordered interaction history $\mathcal{H}_i = \{h_{i1}, h_{i2}, \dots, h_{iT_i}\}$, where each $h_{it} = (q_{it}, r_{it})$ is a query-response pair. Given a new query Q_i , the goal is to generate a personalized response conditioned on the user’s behavioral history: $P(\mathcal{R}_i | Q_i, \mathcal{H}_i)$. Following the PHF hierarchy, each interaction is abstracted into a practice representation, temporally aggregated into a habitus embedding $\mathbf{h}_i$, and associated with a field embedding $\mathbf{f}_i$. Per-

sonalized generation is thus conditioned on both individual and collective signals:

$$P(\mathcal{R}_i \mid \mathcal{Q}_i, \mathbf{h}_i, \mathbf{f}_i). \quad (1)$$

3.2 PHF_{Compass} Architecture

To instantiate PHF, we develop PHF_{Compass}, a lightweight and model-agnostic architecture that realizes the three levels with dedicated components. The architecture is shown in the right part of Figure. 2, and training consists of two stages. The first stage learns denoised *practice* representations. The second constructs *habitus* and *field* representations and trains lightweight projections for personalized generation. We also explore alternative pathways for implementing PHF in Appendices B and C.

Enhanced Practice: Denoised Behavior Essence As illustrated in Figure. 2-(1), raw user behaviors often contain substantial surface variation. Directly modeling continuous embeddings risks overfitting to contextual details rather than capturing stable behavioral semantics. To obtain more robust representations, we follow recent semantic quantization approaches (Deng et al., 2025; Chen et al., 2025) and abstract each interaction into a discrete semantic prototype (Wang et al., 2026). This discretization serves as a behavioral abstraction process: behaviors with different surface forms but similar underlying intent converge toward shared prototypes.

Given an interaction h_{it} , we first encode it into a dense representation:

$$\mathbf{x}_{it} = \text{Enc}(h_{it}), \quad (2)$$

where $\text{Enc}(\cdot)$ denotes a lightweight encoder. Detailed implementation choices are deferred to the Appendix K.

We then apply multi-level residual vector quantization to progressively decompose each interaction into coarse-to-fine semantic components:

$$\mathbf{q}_l = \arg \min_{\mathbf{c} \in \mathcal{C}_l} \|\mathbf{r}_{l-1} - \mathbf{c}\|_2, \quad \mathbf{r}_l = \mathbf{r}_{l-1} - \mathbf{q}_l, \quad (3)$$

where $\mathbf{r}_0 = \mathbf{x}_{it}$ and $\mathcal{C}_l$ denotes the codebook at the l -th quantization level. Higher levels capture dominant behavioral intent, while lower levels progressively refine finer-grained distinctions. The final practice representation is reconstructed as:

$$\hat{\mathbf{x}}_{it} = \sum_{l=1}^L \mathbf{q}_l, \text{SID}(p_{it}) = (i_1, i_2, \dots, i_L), \quad (4)$$

where $(i_1, \dots, i_L)$ denotes the hierarchical Semantic ID of the interaction.

The quantization module is optimized through $\mathcal{L}_{\text{RQ}}$, a combination of reconstruction, commitment, and codebook regularization objectives:

$$\begin{aligned} \mathcal{L}_{\text{RQ}} = & \|\mathbf{x}_{it} - \hat{\mathbf{x}}_{it}\|_2^2 + \lambda_v \sum_{l=1}^L \|\mathbf{r}_{l-1} - \text{sg}(\mathbf{q}_l)\|_2^2 \\ & + \lambda_v \sum_{l=1}^L \|\text{sg}(\mathbf{r}_{l-1}) - \mathbf{q}_l\|_2^2 + \lambda_u \sum_{l=1}^L \text{KL}(\bar{p}_l \parallel \mathcal{U}). \end{aligned} \quad (5)$$

The first term minimizes reconstruction error between the original and quantized representations. The second and third terms are symmetric commitment losses that encourage the encoder outputs and codebook entries to converge, where $\text{sg}(\cdot)$ denotes the stop-gradient operator. The final term is a codebook utilization regularizer that penalizes deviation of the codeword usage distribution $\bar{p}_l$ from the uniform distribution $\mathcal{U}$, preventing codebook collapse. Following prior work (Deng et al., 2025; Chen et al., 2025), the codebooks are initialized through balanced K-Means and are maintained as task-specific. Additional implementation details are provided in the Appendix L.

Habitus: Modeling Long-Term Behavioral Dispositions As illustrated in Figure. 2-(2), to model the stable behavioral dispositions formed through repeated behaviors, we aggregate the sequence of denoised practice representations $\{\hat{\mathbf{x}}_{it}\}_{t=1}^{T_i}$ for user u_i over time with a time fusioner. Since more recent behaviors tend to better reflect a user’s current preferences, we simplify this as a temporally weighted aggregation:

$$\mathbf{h}_i = \sum_{t=1}^{T_i} w_t \hat{\mathbf{x}}_{it}, \quad w_t = \frac{\exp(\sqrt{t})}{\sum_{\tau=1}^{T_i} \exp(\sqrt{\tau})}. \quad (6)$$

The square-root scaling ensures that the recency effect grows sublinearly, preventing recent behaviors from dominating the representation while still preserving temporal sensitivity. The resulting habitus vector $\mathbf{h}_i$ encodes a coherent representation of the user’s long-term behavioral disposition, which subsequently serves as the basis for modeling collective behavioral structure in the field. Appendix B further compares alternative temporal weighting schemes.

Field: Collective Behavioral Structure As illustrated in Figure. 2-(3), we use a group router over habitus representations to capture shared behavioral regularities across users. The intuition is related to collaborative filtering (Shi et al., 2025): users with similar long-term representations may provide a useful group-level prior. However, unlike classical collaborative filtering, personalization datasets typically lack sufficient overlapping observations across identical items or contexts. We therefore approximate this shared structure by inducing latent fields through clustering over user-level habitus representations.

Specifically, we induce latent fields by partitioning users with their habitus via K-Means:

$$\{\mu_k\}_{k=1}^K = \arg \min_{\{\mu_k\}} \sum_{i=1}^N \min_k \|\mathbf{h}_i - \mu_k\|_2^2, \quad (7)$$

where each centroid μ_k corresponds to a field embedding that encodes the shared behavioral disposition of users within that cluster. Each user is then assigned to the nearest field:

$$\mathbf{f}_i = \mu_{k^*}, \quad k^* = \arg \min_k \|\mathbf{h}_i - \mu_k\|_2. \quad (8)$$

The resulting field representations capture collective behavioral regularities across users with similar long-term dispositions, providing complementary signals that are particularly informative under sparse-history and cold-start settings.

3.3 Generation with Hierarchical User Representation

Finally, as shown in Figure. 2-(4), we condition the downstream LLM on both habitus and field representations. Specifically, we project them into the LLM embedding space:

$$\phi_h : \mathbb{R}^d \rightarrow \mathbb{R}^{d_{\text{llm}}}, \quad \phi_f : \mathbb{R}^d \rightarrow \mathbb{R}^{d_{\text{llm}}}. \quad (9)$$

The final input sequence is constructed as:

$$\mathbf{X}_{\text{in}} = [\phi_h(\mathbf{h}_i), \phi_f(\mathbf{f}_i), \text{Emb}(\mathcal{Q}_i)]. \quad (10)$$

The model is optimized using the standard language modeling objective:

$$\mathcal{L}_{\text{LM}} = - \sum_{t=r}^M \log P(x_t | \mathbf{X}_{\text{in}}, x_{<t}). \quad (11)$$

Only lightweight projection layers are trained, while the underlying LLM remains frozen. This

design enables scalable personalization grounded in both individual behavioral dispositions and collective social structure. During inference, each user is assigned to the nearest field based on cluster centroids learned during training, requiring no access to other users' data.

4 Experiments

4.1 Experiment Setup

4.1.1 Dataset

We conduct experiments on the Language Model Personalization (LaMP) benchmark (Salemi et al., 2024b). Following prior work (Liu et al., 2025; Shi et al., 2025; Tan et al., 2024b,a; Zhang et al., 2025a), we evaluate our framework on all six public tasks². The selected tasks span both 3 classification tasks and 3 generation tasks. The classification tasks contain: (1) **LaMP-1**: Personalized Citation Identification; (2) **LaMP-2**: Personalized Movie Tagging; (3) **LaMP-3**: Personalized Product Rating; The generation tasks contain: (4) **LaMP-4**: Personalized News Headline Generation; (5) **LaMP-5**: Personalized Scholarly Title Generation; and (6) **LaMP-7**: Personalized Tweet Paraphrasing. Since the test sets are held out by the benchmark organizers, we report results on the official validation sets using the provided time-based splits. More details are provided in Appendix H.

4.1.2 Evaluation Metrics

Following the default evaluation protocol of the LaMP benchmark, we adopt task-specific metrics as follows: **Accuracy** for LaMP-1; **Accuracy** and **F1-score** for LaMP-2; **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** for LaMP-3; and **ROUGE-1 (R-1)**, **ROUGE-L (R-L)** for LaMP-4, LaMP-5, and LaMP-7.

4.2 Implementation Details

We use **Contriever** (Izacard et al., 2021) as the default encoder and **Qwen2.5-7B-Instruct** (Yang et al., 2024) as the decoder. For residual quantization codebooks, we use $L = 3$ layers with codebook sizes of 32, 32, and 16. The default number of retrieved histories is set to $k = 8$, and the number of field clusters is set to $K = 10$. We optimize the model using AdamW with a learning rate of 1×10^{-4} . All experiments are conducted on 8

²We didn't include LaMP-6 Personalized Email Subject Generation as it is not publicly available.

Task	Metric	Zero Shot	Flat: Retrieval-Based				Flat: Aggregation-Based			PHF Implements					
			Recency	BM25	Dense	ROPG	OPPU	DiffMean	PPlug	T-H	T-F	L-F	T-H+T-F	T-H+L-F	PHF _{Compass}
LaMP-1	Acc $\uparrow$	0.464	0.504	0.504	0.529	0.526	-	-	0.520	0.488	0.487	0.535	0.525	<u>0.569</u>	0.631
LaMP-2	Acc $\uparrow$	0.366	0.449	0.453	0.490	0.489	0.444	0.374	0.503	0.457	0.467	<u>0.528</u>	0.518	0.506	0.531
	F1 $\uparrow$	0.337	0.432	0.437	0.468	0.455	0.400	0.343	0.448	0.444	0.433	<u>0.490</u>	0.478	0.474	0.461
LaMP-3	MAE $\downarrow$	0.498	0.456	0.495	0.414	0.483	0.426	0.504	<u>0.314</u>	0.355	0.367	0.419	0.324	0.334	0.260
	RMSE $\downarrow$	0.825	0.811	0.859	0.750	0.840	0.740	0.834	0.805	0.691	0.725	0.762	0.660	<u>0.655</u>	0.579
LaMP-4	R-1 $\uparrow$	0.145	0.167	0.170	0.171	0.165	<u>0.192</u>	0.146	0.182	0.168	0.152	0.168	0.158	0.174	0.193
	R-L $\uparrow$	0.123	0.150	0.151	0.153	0.148	<u>0.170</u>	0.126	0.162	0.151	0.136	0.151	0.140	0.155	0.172
LaMP-5	R-1 $\uparrow$	0.441	0.461	0.463	0.465	0.461	0.475	0.442	<u>0.512</u>	0.471	0.476	0.476	0.485	0.489	0.516
	R-L $\uparrow$	0.360	0.387	0.389	0.391	0.385	0.398	0.358	<u>0.443</u>	0.392	0.398	0.403	0.411	0.412	0.453
LaMP-7	R-1 $\uparrow$	0.448	0.418	0.464	0.462	0.447	-	0.444	<u>0.524</u>	0.456	0.414	0.463	0.418	0.459	0.535
	R-L $\uparrow$	0.392	0.365	0.408	0.406	0.393	-	0.388	<u>0.465</u>	0.400	0.365	0.406	0.369	0.403	0.475

Table 1: Performance of all methods on the LaMP benchmark. Existing personalization methods are organized under the flat paradigm, while PHF Implements validate each level of the proposed hierarchy. The best results are shown in **bold**, and the second-best are underlined. $\uparrow$ indicates that higher values are better. “-” denotes that the method is not applicable. All metrics are the average of 3 independent reruns with different random seeds.

NVIDIA A100 (80GB) GPUs. Additional details are provided in the Appendix N.

4.3 Baselines

We compare PHF_{Compass} against representative personalization methods organized in 3 categories, along with alternative implements of PHF.

Zero-Shot: The base **Qwen2.5-7B-Instruct** model generates responses solely from the input query without incorporating any user information.

Flat: Retrieval-Based: These methods personalize by selecting individual behaviors from user history without aggregating them into a unified representation. (1) **Recency** retrieves the most recent k behaviors according to temporal order. (2) **BM25** (Robertson and Zaragoza, 2009) retrieves query-relevant histories using sparse lexical matching. (3) **Dense** (Izacard et al., 2021) uses dense semantic embeddings for retrieval. (4) **ROPG** (Salemi et al., 2024a) further improves dense retrieval through reinforcement learning optimization of the retriever.

Flat Aggregation-Based: These methods construct a unified user representation by aggregating behaviors, but without explicitly modeling how behaviors are structured. (1) **OPPU** (Tan et al., 2024b) trains dedicated LoRA (Hu et al., 2022) adapters for each user. (2) **DiffMean** (Zhang et al., 2025a) models user style through residual activation differences between personalized and non-personalized responses. (3) **PPlug** (Liu et al., 2025) aggregates historical interaction embeddings into a unified query-aware user representation.

PHF Variants: To validate each level of the PHF hierarchy, we implement variants that operationalize habitus and field through different mechanisms. (1) **T-H** (Zhang, 2024) constructs habitus by summarizing user histories into textual persona descriptions using **Qwen2.5-14B-Instruct**. (2) **T-F** con-

structs field similar with T-H to summarize collective preferences and behavioral tendencies within each user group. (3) **L-F** (Shi et al., 2025) constructs field through cross-user retrieval over behavior pools from similar users. (4) **T-H+T-F** combines textual habitus summaries with text-based field descriptions. (5) **T-H+L-F** combines textual habitus summaries with cross-user retrieval-based field modeling.

4.4 Main Results

The experimental results on the validation set are shown in Table 1, and we can observe that:

(1) PHF_{Compass} **consistently achieves the best performance across all tasks, yielding substantial improvements over the strongest flat-paradigm baselines on classification tasks and consistent gains on generation tasks.** Among other methods, retrieval-based approaches benefit from semantic matching (e.g., BM25 and Dense consistently outperform Recency and Zero-Shot), but remain limited by their reliance on individual behaviors. Aggregation-based methods such as PPlug and OPPU achieve stronger results by constructing holistic user representations, yet still underperform PHF variants. This suggests that while behavioral aggregation is beneficial, explicitly modeling the structure of user behavior is crucial for capturing the multi-level regularities that drive effective personalization.

(2) **Habitus and field are complementary, and jointly modeling both yields stronger performance than modeling either level alone.** The results show that either level alone improves over flat baselines (e.g., T-H and T-F perform better than some flat paradigms on most tasks), and combining both yields further gains. Comparing different PHF implements, we can further find that PHF_{Compass} achieves strongest results overall, confirming our

Method	LaMP-1		LaMP-2		LaMP-3		LaMP-4		LaMP-5		LaMP-7	
	Acc $\uparrow$		Acc $\uparrow$	F1 $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$	R-1 $\uparrow$	R-L $\uparrow$	R-1 $\uparrow$	R-L $\uparrow$	R-1 $\uparrow$	R-L $\uparrow$
PHF _{Compass}	0.631		0.531	0.461	0.260	0.579	0.193	0.172	0.516	0.453	0.525	0.465
w/o Habitus	0.501		0.401	0.349	0.321	0.649	0.178	0.157	0.490	0.425	0.491	0.431
w/o Field	0.598		0.527	0.435	0.264	0.588	0.199	0.176	0.517	0.453	0.529	0.467
w/o Temporal	0.503		0.538	0.471	0.265	0.584	0.201	0.180	0.519	0.454	0.518	0.460
w/o Codebook	0.496		0.508	0.428	0.293	0.594	0.156	0.138	0.458	0.381	0.500	0.439

Table 2: Ablation study of PHF_{Compass} on the LaMP benchmark. “w/o” denotes removing the corresponding component from PHF_{Compass}. “w/o Temporal” replaces temporally weighted aggregation with simple averaging, and “w/o Codebook” removes residual quantization and use original embedding.

Encoder + Decoder	LaMP-2		LaMP-4	
	Acc	F1	R-1	R-L
Contriever+Qwen	0.531	0.461	0.199	0.176
Contriever+LLaMA	0.510	0.423	0.215	0.193
RoBERTa+Qwen	0.523	0.451	0.198	0.177
RoBERTa+LLaMA	0.533	0.437	0.213	0.191

Table 3: Overall performance of models with different encoders and decoders on LaMP-2 and LaMP-4.

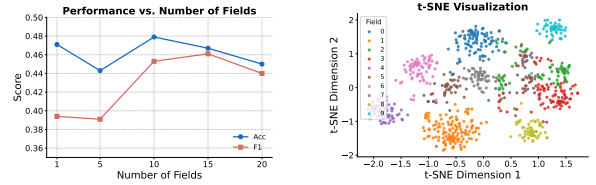
dedicated behavioral modeling outperforms text-based alternatives like T-H+T-F and T-H+L-F.

(3) **Different tasks benefit from different levels of the PHF hierarchy, highlighting the distinct roles of habitus and field in personalization.** Classification tasks (LaMP-1 to 3) primarily depend on user behavioral logic, so field modeling is particularly beneficial by compensating for sparse individual histories. As a result, field-enhanced methods like T-H, T-H+T-F and PHF_{Compass} obviously lead the classification tasks. In contrast, generation tasks rely on content stylistic consistency, so training-based aggregation methods like OPU and PPlug remain competitive. PHF_{Compass}, benefiting from the robust modeling of practice and habitus (see Sec. 4.5), also show prominent performance on these tasks. Overall, PHF_{Compass} achieves the strongest performance across both task types, demonstrating the advantage of modeling multiple levels of user behavior.

4.5 Ablation Study

We conduct an ablation study to evaluate the contribution of each component in PHF_{Compass}. As shown in Table 2, we analyze the model from two perspectives:

The effect of PHF constructs. Removing habitus (*w/o Habitus*) causes the largest performance degradation across all tasks, confirming that fine-grained individual behavioral representations are fundamental to effective personalization. Removing field (*w/o Field*) leads to more moderate declines, particularly on classification tasks, indicating that collaborative signals are especially benefi-



(a) Performance of different field number on LaMP-2

(b) The t-SNE visualization of user cluster fields.

Figure 3: The influence of field granularity.

cial when behavioral reasoning rather than stylistic variation drives personalization. These results show that **both habitus and field contribute meaningfully to personalization, with habitus playing the dominant role.** We further find that Field benefits users with long histories but slightly harms users with short histories. As shown in Appendix D, a confidence-based dynamic gate disables Field when its assignment is unreliable, thereby improving overall performance.

The effect of implementation choices of PHF_{Compass}. Replacing temporal weighting (*w/o Temporal*) with uniform averaging mainly affects tasks involving evolving user preferences, suggesting that temporal aggregation provides complementary gains beyond habitus construction itself. Appendix B further reports alternative implementation pathways for Habitus aggregation. In addition, removing residual quantization (*w/o Codebook*) consistently degrades performance across all tasks, with especially large drops on generation benchmarks. This highlights the importance of behavioral abstraction: without discrete practice representations, surface-level noise propagates through the entire hierarchy. Therefore, **the effectiveness of PHF also depends on our implementation choices.**

4.6 Further Analysis

We further conduct additional analysis on LaMP-2 (classification) and LaMP-4 (generation).

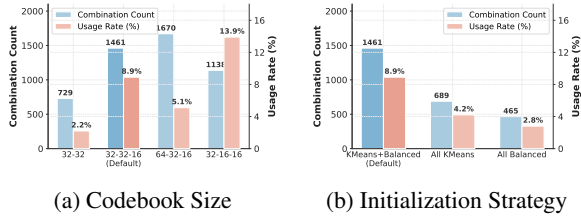


Figure 4: Impact of codebook configuration. The notation *a-b-c* indicates the codebook size for each layer. “All KMeans” applies standard K-Means initialization for all three layers.

4.6.1 Impact of Field Granularity

A key design choice in PHF is the granularity of the field, which determines how finely the user population is partitioned into behavioral communities. This granularity is controlled by the number of clusters K : a smaller K yields coarser fields that group diverse users together, potentially diluting shared behavioral signals with outlier patterns; a larger K produces finer fields with higher internal homogeneity but fewer users per cluster, weakening the collaborative signal due to data sparsity. This trade-off is evident in Figure. 3a: performance improves as K increases from small values, but begins to decline once the fields become overly fragmented. Appendix C further evaluates alternative Field construction and assignment methods. The best balance is achieved at $K = 10$, where both metrics are jointly optimized. Figure. 3b further illustrates the corresponding t-SNE visualization of user habitus, showing that the induced fields are well separated and internally cohesive.

4.6.2 Model-Agnostic Extensibility

To validate the generalizability of PHF_{Compass} across different model architectures, we replace the default **Contriever** encoder and **Qwen2.5-7B-Instruct** decoder with **RoBERTa-large** (Liu et al., 2019) and **LLaMA-3-8B-Instruct** (Llama Team, 2024), respectively. As shown in Table 3, PHF_{Compass} maintains strong performance across all four encoder-decoder combinations, with only minor variation across configurations. This robustness demonstrates that the PHF framework is not tied to any specific backbone architecture. We further observe that LLaMA performs better on generation tasks, while Qwen achieves comparable or stronger results on classification tasks, suggesting that although backbone choice influences task preference, **the effectiveness of PHF remains consistent across architectures**. Appendix G further examines cross-task Practice-codebook transfer, pro-

viding additional evidence for the model-agnostic extensibility of PHF.

4.6.3 Impact of Codebook Configuration

The codebook size directly determines the number of discrete practice categories and therefore the expressive capacity of user behavior modeling. To analyze this impact, we conduct experiments using different **codebook size**. As shown in Figure. 4(a), a small codebook fails to capture sufficient behavioral diversity, limiting personalization, while an excessively large codebook leads to sparse codeword usage and degraded performance due to under-trained entries. Furthermore, we analyze different **initialization strategy** (Figure. 4(b)). We find that the first two quantization layers primarily capture dominant behavioral patterns, for which standard K-Means provides a better fit than enforcing uniform code usage through balanced K-Means. In contrast, the third residual layer models small-magnitude residual signals and is more prone to code collapse under standard K-Means. This results in the final 32-32-16 configuration with a hybrid initialization strategy: standard K-Means for the first two layers and balanced K-Means for the third. This achieves both high behavioral coverage and stable code utilization.

4.6.4 Interpretability of PHF_{Compass}

PHF_{Compass} uses RQ-VAE to extract both shared patterns and personalized signals in a highly interpretable manner: (1) movies in user behavior histories with same semantic ID exhibit strong thematic coherence: e.g., 0-14-2 groups romance narratives set in fantastical worlds (*Wonder Woman*, *Howl’s Moving Castle*). (2) even under the same coarse tag like *romance*, different level-1 codes reveal fine-grained distinctions: 0- emphasizes otherworldly settings, while 16- focuses on identity and emotional realism (*Juno*, *Safe Haven*). (3) identical movies yield distinct IDs when paired with different tags, e.g., *Inception* as *action* gives 8-21-12, but as *psychology* it becomes 15-21-12, with only the level-1 code changing while deeper layers (-21-12) remain stable. These results suggest that the hierarchical quantization disentangles user-specific intent at the top level from shared semantic structure in deeper residual layers. More details are provided in Appendix J.

5 Conclusion

In this work, we propose PHF (Practice–Habit–Field), a sociologically inspired framework that organizes LLM personalization into a unified behavioral hierarchy, jointly modeling within-user behavioral formation and cross-user behavioral regularity. Under this framework, we develop PHF_{Compass}, a lightweight and model-agnostic implementation that abstracts user behaviors into denoised practices, aggregates them into stable habitus, and situates users within shared fields that capture collective behavioral regularities. Experiments on the LaMP benchmark demonstrate consistent improvements across both classification and generation tasks, while further analyses confirm the interpretability of the learned behavioral structures and the extensibility of the framework across different model architectures.

Limitations

This work has several limitations that suggest directions for future research.

Our experiments are conducted exclusively on the LaMP benchmark, which primarily consists of English-language text-based tasks with static user histories. The effectiveness of PHF on multilingual, multimodal, or domain-specific personalization scenarios (e.g., e-commerce, healthcare) remains to be validated. Additionally, since LaMP does not provide continuously evolving user interactions, we are unable to evaluate revolving habitus updates or dynamic field re-clustering in a streaming setting, which we consider an important direction for future work.

We evaluate personalization using standard classification and generation metrics such as Accuracy, F1, and ROUGE. These metrics may not fully capture the nuanced aspects of personalization quality, such as user satisfaction, perceived relevance, or behavioral consistency over time.

While PHF_{Compass} demonstrates one effective way to operationalize the PHF hierarchy, the specific implementations of each level are by no means the only options. For instance, habitus could be modeled through recurrent or attention-based temporal architectures rather than weighted aggregation, and fields could be induced through graph-based community detection or dynamic clustering rather than static K-Means. Exploring alternative implementations under the PHF framework is a promising avenue for future work.

Ethical Considerations

Personalization inherently involves modeling individual user behavior, which raises potential concerns regarding privacy and data usage. In this work, all experiments are conducted on the publicly available LaMP benchmark, which consists of anonymized user interaction data. Our framework operates without accessing other users’ raw data during inference — field embeddings are derived from cluster centroids learned during training, and no cross-user data sharing occurs at test time. Nevertheless, deploying personalization systems in real-world settings requires careful attention to user consent and data protection regulations. Additionally, clustering users into shared fields based on behavioral similarity may group users in ways that reinforce existing biases in the training data. We encourage future work to investigate fairness-aware extensions of the PHF framework and to evaluate personalization systems under broader ethical criteria beyond task performance.

Acknowledgments

The research is supported by National Key R&D Program of China (Grant No. 2023YFF1204800) and the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (Grant No. 2025-GZL-RGZN-BTBX-02028). The project’s computational resources are partially supported by CFFF platform of Fudan University.

We only use AI assistants as tools for paper refinement and code implementation assistance, and we do not engage in any practice of using AI to directly generate paper content.

References

- Zhijie Bao, Qingyun Liu, Ying Guo, Zhengqiang Ye, Jun Shen, Shirong Xie, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2024. Piors: Personalized intelligent outpatient reception based on large language model with multi-agents medical scenario simulation. *arXiv preprint arXiv:2411.13902*.
- Pierre Bourdieu. 1977. *Outline of a Theory of Practice*. Cambridge University Press.
- Pierre Bourdieu. 1990. *The logic of practice*. Stanford university press.
- Ben Chen, Xian Guo, Siyuan Wang, Zihan Liang, Yue Lv, Yufei Ma, Xinlong Xiao, Bowen Xue, Xuxin Zhang, Ying Yang, Huangyu Dai, Xing Xu, Tong Zhao, Mingcan Peng, Xiaoyang Zheng, Chao Wang, Qihang Zhao, Zhixin Zhai, Yang Zhao, Bocho Liu,

- Jingshan Lv, Xiao Liang, Yuqing Ding, Jing Chen, Chenyi Lei, Wenwu Ou, Han Li, and Kun Gai. 2025. Onesearch: A preliminary exploration of the unified end-to-end generative framework for e-commerce search. *arXiv preprint arXiv:2509.03236*.
- Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. [From persona to personalization: A survey on role-playing language agents](#). *Transactions on Machine Learning Research*.
- Shaojie Dai, Yanwei Yu, Hao Fan, and Junyu Dong. 2022. [Spatio-temporal representation learning with social tie for personalized poi recommendation](#). *Data Science and Engineering*, 7(1):44–56.
- Jiabin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. 2025. Onerec: Unifying retrieve and rank with generative recommender and iterative preference alignment. *arXiv preprint arXiv:2502.18965*.
- Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2024. [Plug-and-play policy planner for large language model powered dialogue agents](#). In *The Twelfth International Conference on Learning Representations*.
- Sumanth Doddapaneni, Krishna Sayana, Ambarish Jash, Sukhdeep Sodhi, and Dima Kuzmin. 2024. [User embedding model for personalized language prompting](#). In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 124–131. Association for Computational Linguistics.
- Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. [A multi-view deep learning approach for cross domain user modeling in recommendation systems](#). In *Proceedings of the 24th international conference on world wide web*, pages 278–288.
- Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Yiheng Sun, Zerui Chen, Ming Liu, and Bing Qin. 2025. [Simulation-free hierarchical latent policy planning for proactive dialogues](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Liam Hebert, Krishna Sayana, Ambarish Jash, Alexandros Karatzoglou, Sukhdeep Sodhi, Sumanth Doddapaneni, Yanli Cai, and Dima Kuzmin. 2024. Persoma: Personalized soft prompt adapter architecture for personalized language prompting. *arXiv preprint arXiv:2408.00960*.
- Mehdi Hosseinzadeh, Fa Zhu, Amir Masoud Rahmani, Tofan Agung Eka Prasetya, Omar Almomani, Gholamreza Zare, Parisa Khoshvaght, Aso Darwesh, Thantrira Porntaveetus, and Jan Lansky. 2026. [TiLLM-Rec: Temporal-interval-aware large language model for sequential recommendation under irregular user interactions](#). *ICT Express*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations*.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.
- Zheng Jia, Shengbin Yue, Wei Chen, Siyuan Wang, Yidong Liu, Zejun Li, Yun Song, and Zhongyu Wei. 2026. [Ready jurist one: Benchmarking language agents for legal intelligence in dynamic environments](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10351–10376. Association for Computational Linguistics.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611. Association for Computational Linguistics.
- Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. 2022. [A survey of recommendation systems: recommendation models, techniques, and application fields](#). *Electronics*, 11(1):141.
- Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.
- Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.
- 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. Association for Computational Linguistics.
- Jingcong Liang, Shijun Wan, Xuehai Wu, Yitong Li, Qianglong Chen, Duyu Tang, Siyuan Wang, and Zhongyu Wei. 2025. Hardcorelogic: Challenging large reasoning models with long-tail logic puzzle games. *arXiv preprint arXiv:2510.12563*.
- 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 (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Llama Team. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. 2026. [From individual to society: A survey on social simulation driven by large language model-based agents](#). *ACM Computing Surveys*, 58(11):280:1–280:41.
- Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. [Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4789–4809. Association for Computational Linguistics.
- Ruihui Mu. 2018. [A survey of recommender systems based on deep learning](#). *IEEE Access*, 6:69009–69022.
- Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O’Banion, and Jun Xie. 2025. [User-llm: Efficient llm contextualization with user embeddings](#). In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1219–1223.
- Erasmus Purificato, Ludovico Boratto, and Ernesto William De Luca. 2024. User modeling and user profiling: A comprehensive survey. *arXiv preprint arXiv:2402.09660*.
- Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, Maciej Kula, Ed Chi, and Maheswaran Sathiamoorthy. 2023. [Recommender systems with generative retrieval](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 10299–10315. Curran Associates, Inc.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: Bm25 and beyond](#). *Found. Trends Inf. Retr.*, 3(4):333–389.
- Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024a. [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. 2024b. [LaMP: When large language models meet personalization](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392. Association for Computational Linguistics.
- 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](#). In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1294–1304.
- Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024a. [Personalized pieces: Efficient personalized large language models through collaborative efforts](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- 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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.
- Yihong Tang, Bo Wang, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. [MORPHEUS: Modeling role from personalized dialogue history by exploring and utilizing latent space](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7664–7676. Association for Computational Linguistics.
- Meiling Tao, Chenghao Zhu, Dongyi Ding, Tiannan Wang, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2025. [Personafeedback: A large-scale human-annotated benchmark for personalization](#). *arXiv preprint arXiv:2506.12915*.
- Liang Wang, Xinyi Mou, Xiaoyou Liu, Xuanjing Huang, and Zhongyu Wei. 2026. [Curp: Codebook-based continuous user representation for personalized generation with llms](#). *arXiv preprint arXiv:2602.00742*.
- Piaohong Wang, Motong Tian, Jiaxian Li, Yuan Liang, Yuqing Wang, Qianben Chen, Tiannan Wang, Zhicong Lu, Jiawei Ma, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2025. [O-mem: Omni memory system for personalized, long horizon, self-evolving agents](#). *arXiv preprint arXiv:2511.13593*.
- Yingqian Wu, Qiushi Wang, Zefei Long, Rong Ye, Zhongtian Lu, Xianyin Zhang, Bingxuan Li, Wei Chen, Liwen Zhang, and Zhongyu Wei. 2025. [Fin-team: A multi-agent collaborative intelligence system for comprehensive financial scenarios](#). In *CCF*

International Conference on Natural Language Processing and Chinese Computing, pages 443–455. Springer.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, Qi Zhang, and Tao Gui. 2025. [The rise and potential of large language model based agents: A survey](#). *Science China Information Sciences*, 68(2):121101.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.

Jiarui Zhang. 2024. Guided profile generation improves personalization with llms. *arXiv preprint arXiv:2409.13093*.

Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025a. [Personalized text generation with contrastive activation steering](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7128–7141. Association for Computational Linguistics.

Kai Zhang, Yejin Kim, and Xiaozhong Liu. 2024. Personalized llm response generation with parameterized memory injection. *arXiv preprint 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](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16399–16411. Association for Computational Linguistics.

Bowen Zheng, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, Ming Chen, and Ji-Rong Wen. 2024. [Adapting large language models by integrating collaborative semantics for recommendation](#). In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 1435–1448. IEEE.

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 Scope of the PHF Framework

Bourdieu’s Theory of Practice motivates the organization of personalization around recurring actions, persistent individual dispositions, and shared supra-individual regularities. PHF uses this organization as a problem-formulation framework rather than claiming a computational reproduction of the complete theory. Concepts such as capital, embodiment, social reproduction, power relations, and social-space stratification require demographic, relational, and institutional information that LaMP does not contain. Within the available behavioral setting, PHF contributes a hierarchical expression that combines Practice-level denoising, Habitus-level temporal representation, and Field-level shared regularities. Table 4 contrasts this perspective with several traditional paradigms.

Framework property	User emb.	VQ	CF	PHFCompass
Behavior-level denoising	×	✓	×	✓
Temporal user representation	×	×	×	✓
Cross-user shared regularities	×	✓	✓	✓
Hierarchical expression	×	×	×	✓

Table 4: Comparison with conventional user embeddings, vector quantization (VQ), and collaborative filtering (CF).

B Alternative Habitus Aggregation

We compare the default square-root recency weighting with other fixed functions and a learned single-query attention-pooling module containing approximately 0.6M parameters. All variants use the same diagnostic training budget and differ only in Habitus aggregation. Table 5 shows that no fixed or learned function dominates both tasks: exponential weighting is strongest on LaMP-2, whereas square-root weighting is strongest on LaMP-4. Learned attention is competitive but not uniformly superior. We therefore retain normalized square-root weighting as a lightweight default rather than a theoretically privileged implementation.

Habitus aggregation	LaMP-2 Acc	LaMP-4 R-1
Uniform	0.4184	0.1904
Linear	0.4610	0.1918
Logarithmic	0.4695	0.1844
Exponential	0.4738	0.1876
Square root (default)	0.4369	0.1945
Learned attention	0.4688	0.1936

Table 5: Alternative Habitus aggregation under a shared diagnostic budget.

C Alternative Field Constructions

We first assess clustering stability across repeated runs. The induced Field partition is relatively stable on LaMP-4 (pairwise ARI = 0.86, NMI = 0.90) and moderately stable on LaMP-2 (ARI = 0.59, NMI = 0.71), indicating that Field reliability depends on the task and user distribution. We then vary the construction or assignment mechanism while keeping the downstream pipeline fixed. As shown in Table 6, GMM component means perform best on LaMP-2, while K-Means with $K = 10$ performs best on LaMP-4. Soft cosine assignment is viable but not universally superior. These results support PHF’s shared-regularity level while showing that static hard K-Means is only one possible realization.

Field construction/assignment	LaMP-2 Acc	LaMP-4 R-1
K-Means, $K = 5$	0.4270	0.1888
K-Means, $K = 10$	0.4369	0.1945
K-Means, $K = 20$	0.4277	0.1854
GMM component means, $K = 10$	0.4496	0.1875
Cosine-softmax mixture ($\tau = 0.1$)	0.4369	0.1888

Table 6: Alternative Field constructions under a shared diagnostic budget.

D Conditional Field Contribution and Confidence Gating

The Field effect is concentrated among users with sufficiently informative histories. Table 7 shows that Field slightly hurts short-history users but improves long-history users by 5.0% relatively. A plausible explanation is that short histories yield less stable Habitus estimates and therefore less reliable Field assignments. Field is consequently better viewed as a conditional group-level regularizer than as a universally beneficial signal.

History group	w/o Field R-1	w/ Field R-1	Effect (abs./rel.)
Short	0.1873	0.1847	−0.0026 / −1.4%
Long	0.1838	0.1930	+0.0091 / +5.0%

Table 7: Field contribution by user-history length on LaMP-4.

To control unreliable assignments, we compute the margin between the distances to the second-nearest and nearest Field centroids. Let $d_{i,(1)}$ and $d_{i,(2)}$ denote these distances. The gated Field rep-

resentation is

$$\begin{aligned} m_i &= d_{i,(2)} - d_{i,(1)}, \\ g_i &= \mathbb{I}[m_i > Q_{0.25}(m)], \\ \tilde{\mathbf{f}}_i &= g_i \mathbf{f}_i. \end{aligned} \quad (12)$$

where $Q_{0.25}(m)$ is the lowest-margin quartile threshold estimated on the validation population. This inference-time diagnostic disables Field for uncertain assignments and retains it otherwise. The gated mixture reaches 0.1940 ROUGE-1, compared with 0.1923 without Field. The gate does not establish universal Field gains, but demonstrates that reliability-aware activation can preserve useful cross-user information while suppressing noisy assignments.

ROUGE may underreport group-level personalization effects, so we additionally compare structural properties of generated outputs over all 1,500 LaMP-4 examples. Table 8 shows a modest but consistent reduction in style distance and improvement in binary form-feature matching.

Structural metric	Effect	95% CI
Style-distance reduction	−4.0% relative	[2.1%, 5.8%]
Form-feature match	+0.6 p.p.	[+0.24, +0.94]

Table 8: Structural personalization effects of Field on LaMP-4.

E Representation Diagnostics across PHF Levels

We test whether the learned representations exhibit the properties intended by the PHF hierarchy. Practice recurrence is assessed by comparing full Semantic IDs, shared prefixes, and random pairs. Habitus persistence is assessed through within-user similarity against temporally permuted and cross-user controls. Field regularity is assessed through query-tag agreement within the same Field against random users. Table 9 shows the expected ordering at all three levels, providing direct representation-level evidence beyond downstream performance.

Level	Intended property	Evidence
Practice	Recurrence	full-ID cos. 0.617 > prefix 0.555 > random 0.507
Habitus	Persistence	same-user 0.944 > permuted 0.822; 97.1% satisfy same > cross
Field	Cross-user regularity	same-Field agreement 16.8% > random 9.1%

Table 9: Diagnostics of the intended property at each PHF level.

F Variance and Statistical Significance

Because generation scores are close, we report standard deviation over three runs and paired-bootstrap 95% confidence intervals for the primary-metric difference $\text{PHF}_{\text{Compass}} - \text{PPlug}$. Table 10 shows low run-to-run variance. The intervals for LaMP-4 and LaMP-5 exclude zero, while the LaMP-7 interval reaches zero at its lower bound; we therefore describe the evidence as strongest on LaMP-4/5 and borderline on LaMP-7 rather than claiming uniform significance.

Task	Std.	Paired difference 95% CI
LaMP-4	0.0004	[0.0043, 0.0175]
LaMP-5	0.0046	[0.0101, 0.0261]
LaMP-7	0.0030	[0.0000, 0.0114]

Table 10: Variance and paired-bootstrap confidence intervals on generation tasks.

G Cross-Task Practice-Codebook Portability

We evaluate whether Practice abstractions can be reused across tasks. A source-task codebook is frozen and used to quantize histories in the target task; target-side Field centers are then re-estimated, and only target projection modules are trained. As shown in Table 11, the transferred codebook retains 92% of target-specific performance from LaMP-4 to LaMP-2 and 98% in the reverse direction. Thus, the Practice vocabulary is not confined to one task, although the remaining gap and target-side adaptation requirement preclude a claim of zero-shot cross-task deployment.

Source → Target	Target-specific	Transferred	Retained
LaMP-4 → LaMP-2	0.4619	0.4270	92%
LaMP-2 → LaMP-4	0.1933	0.1897	98%

Table 11: Cross-task portability of the frozen Practice codebook.

H Dataset Details

Detailed statistics and example input output pair for all six tasks are provided in Table 12. We use the time-split of the benchmark.

I Integration with RAG

Our preliminary experiments show that retrieval-augmented generation (RAG) can provide useful

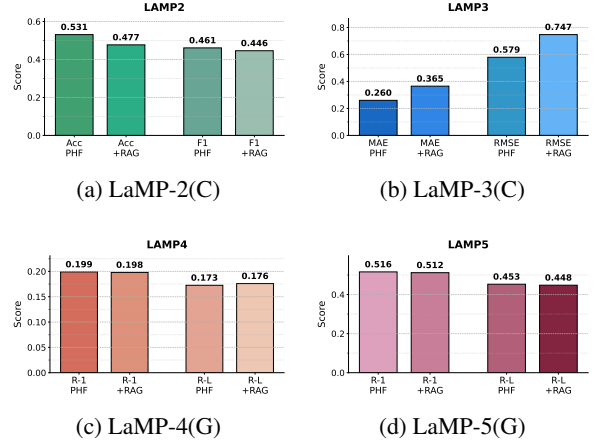


Figure 5: Performance comparison across LaMP-2 to 5. Each subfigure shows scores for PHF and PHF+RAG variants.

personalization signals. We therefore investigate whether integrating RAG with our habitus and field embeddings further improves performance. As shown in Figure 5, we concatenate retrieved user history with the PHF embeddings as additional context. Surprisingly, this integration consistently degrades performance across tasks. We attribute this to the noise and excessive length introduced by raw retrieved text, which dilutes the focused signal from our compact latent embeddings. The effect is especially pronounced in classification tasks (LaMP-2 and LaMP-3), where performance drops significantly. In contrast, generation tasks (LaMP-4 and LaMP-5) show only minor fluctuations, suggesting greater robustness to input noise. These results indicate that our latent habitus and field embeddings already capture essential user characteristics effectively, eliminating the need for redundant or noisy textual context from RAG.

J Interpretability

To further exhibit the interpretability of the Semantic ID based practice expression, we further exhibit some examples as shown in Table 13 and 15.

K Detailed Procedure for Encoder Refinement

Although the main framework only requires a lightweight interaction encoder, in practice we further refine the encoder to better align contextual semantics with user-specific behavioral expression.

Task	#Train	#Dev	In Len	#Cls or Out Lens	Hist Len	Input Format	Output Format	History Format
LaMP-1	6,542	1,500	51.43	2	84.15	For an author who has written the paper with title "{title}", which reference is related? Just answer with <1> or <2>. <1>: "{ref1}" <2>: "{ref2}"	<1>	title: {title} abstract: {abs}
LaMP-2	5,073	1,410	92.39	15	86.76	Which tag does this movie relate to among the following tags? Just answer with the tag name. tags: [sci-fi, action, ...] description: {movie}	sci-fi	description: {des} tag: {tag}
LaMP-3	20,000	2,500	128.18	5	185.40	What is the score of the following review on a scale of 1 to 5? Just answer with 1-5. review: {review}	3	text: {review} score: {score}
LaMP-4	12,500	1,500	29.97	10.07	204.59	Generate a headline for the following article: {article}	How I Got 'Rich'	title: {title} text: {article}
LaMP-5	14,682	1,500	162.34	9.71	87.88	Generate a title for the following abstract of a paper: {abstract}	Distributed Partial Clustering	title: {title} text: {abstract}
LaMP-7	13,437	1,498	29.72	16.96	15.71	Paraphrase the following tweet without any explanation: {tweet}	gotta make the most of my last full day in ktown	text: {tweet}

Table 12: Dataset statistics and input–output formats for different tasks in the LaMP benchmark. “#” denotes the number of instances, “Cls” indicates classification tasks, and placeholders in braces { } are replaced with dataset-specific values.

Table 13: Representative movies and their corresponding semantic IDs on LaMP-2

Semantic ID	Genre	Movie (Title + Summary)
7-10-1	fantasy	<i>Avengers: Infinity War</i> – Heroes unite to stop Thanos from collecting the six Infinity Stones and imposing his will on reality.
7-10-1	fantasy	<i>The Lord of the Rings</i> – Aragorn defends Gondor while Frodo and Sam journey toward Mount Doom to destroy the One Ring.
8-10-1	action	<i>Avengers: Endgame</i> – The remaining Avengers assemble one last time to reverse Thanos’ snap and restore the universe.
8-10-1	action	<i>Iron Man 2</i> – Tony Stark battles government pressure, rival inventors, and his own failing health as Iron Man.
7-18-3	fantasy	<i>Harry Potter</i> series – A young wizard uncovers secrets about his parents and battles dark forces throughout his school years.
7-18-3	fantasy	<i>Star Wars: The Last Jedi</i> – Rey seeks training from a reluctant Luke Skywalker as the Resistance prepares for its final stand.
0-14-2	romance	<i>Wonder Woman</i> – An Amazon princess leaves her island to fight in World War I and discover the power of love and truth.
0-14-2	romance	<i>Howl’s Moving Castle</i> – A cursed girl finds refuge and romance in a wizard’s enchanted, walking castle.
16-14-2	romance	<i>Juno</i> – A sharp-witted teenager navigates an unexpected pregnancy and makes an unconventional choice about her future.
16-14-2	romance	<i>Safe Haven</i> – A mysterious woman hiding from her past finds healing and love in a small coastal town.

Table 14: Semantic ID example for some combinations.

Movie (Title + Summary)	Tag	Semantic ID
<i>Inception</i> – A thief who steals secrets from dreams is hired to implant an idea into a target’s subconscious.	action	8-21-12
	sci-fi	5-21-12
	twist ending	12-21-12
<i>Interstellar</i> – Explorers use a wormhole to travel across galaxies and save humanity from extinction.	action	26-26-7
	dystopia	14-26-7
	twist ending	12-26-7
<i>The Dark Knight</i> – Batman battles the Joker, a criminal who tests the moral limits of Gotham’s heroes.	action	21-25-13
	dystopia	14-25-13
	violence	20-25-13

Table 15: Same movie paired with diverse user tags on LaMP-2

Specifically, given an interaction

$$h_{it} = (q_{it}, r_{it}),$$

we separately encode the query and response:

$$\mathbf{q}_{it} = \text{Enc}(q_{it}), \quad \mathbf{r}_{it} = \text{Enc}(r_{it}). \quad (13)$$

The final interaction embedding is constructed through weighted fusion:

$$\mathbf{x}_{it} = \text{norm}(\alpha \cdot \text{norm}(\mathbf{q}_{it}) + \beta \cdot \text{norm}(\mathbf{r}_{it})), \quad (14)$$

where α and β balance contextual information and user-specific expression.

To further improve semantic consistency, we introduce a contrastive alignment objective between two complementary semantic views. The first view directly encodes the concatenated interaction:

$$\mathbf{x}_{it}^{(1)} = \text{norm}(\text{Enc}([q_{it}, r_{it}])), \quad (15)$$

while the second view corresponds to the weighted fusion representation:

$$\mathbf{x}_{it}^{(2)} = \text{norm}(\alpha \cdot \text{norm}(\mathbf{q}_{it}) + \beta \cdot \text{norm}(\mathbf{r}_{it})). \quad (16)$$

Algorithm 1 Encoder Refinement with Contrastive Alignment

Require: Interaction set $\mathcal{H}$

- 1: **for** each interaction $h_{it} = (q_{it}, r_{it})$ **do**
- 2: Encode query embedding $\mathbf{q}_{it}$
- 3: Encode response embedding $\mathbf{r}_{it}$
- 4: Construct fused representation $\mathbf{x}_{it}^{(2)}$
- 5: Encode concatenated interaction $\mathbf{x}_{it}^{(1)}$
- 6: Compute contrastive alignment loss
- 7: Update encoder parameters
- 8: **end for**
- 9: **return** Refined encoder $\text{Enc}(\cdot)$

We treat the two representations as a positive pair and optimize a symmetric InfoNCE objective:

$$\mathcal{L}_{\text{cont}} = \frac{1}{2} \left[\mathcal{H} \left(\frac{\text{sim}(\mathbf{x}_{it}^{(1)}, \mathbf{x}_{it}^{(2)})}{\tau} \right) + \mathcal{H} \left(\frac{\text{sim}(\mathbf{x}_{it}^{(2)}, \mathbf{x}_{it}^{(1)})}{\tau} \right) \right], \quad (17)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, τ is a temperature parameter, and $\mathcal{H}$ denotes cross-entropy loss.

This refinement objective encourages the fused representation to preserve the joint semantics captured by direct interaction encoding while maintaining a disentangled representation of contextual grounding and user-specific behavioral expression.

L Detailed Procedure for Balanced K-Means Initialization

To improve codebook stability during residual quantization, especially in deeper quantization layers where residual magnitudes become progressively smaller, we initialize the final codebook using balanced K-Means.

Given residual representations

$$\mathcal{R} = \{\mathbf{r}_1, \dots, \mathbf{r}_N\},$$

balanced K-Means aims to partition residuals into K approximately balanced clusters:

$$\min_{\{\mathcal{C}_k\}} \sum_{k=1}^K \sum_{\mathbf{r} \in \mathcal{C}_k} \|\mathbf{r} - \boldsymbol{\mu}_k\|_2^2, \quad (18)$$

subject to:

$$|\mathcal{C}_k| \approx \frac{N}{K}, \quad \forall k. \quad (19)$$

Algorithm 2 Balanced K-Means Initialization

Require: Residual set $\mathcal{R}$, number of clusters K

- 1: Initialize centroids $\{\boldsymbol{\mu}_k\}_{k=1}^K$
- 2: **repeat**
- 3: Assign residuals to nearest centroids under balance constraints
- 4: Update centroids using assigned residuals
- 5: **until** cluster assignments converge
- 6: **return** Balanced codebook centroids $\{\boldsymbol{\mu}_k\}_{k=1}^K$

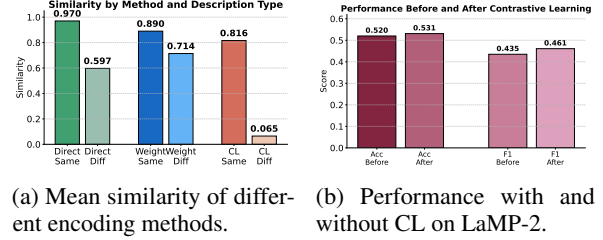


Figure 6: The impact of contrastive learning on the encoder.

Here, $\boldsymbol{\mu}_k$ denotes the centroid of cluster $\mathcal{C}_k$. Compared with standard K-Means, balanced clustering prevents excessive concentration of residuals into a small number of centroids, thereby improving codebook utilization and stabilizing subsequent quantization training.

In practice, we only apply balanced K-Means to initialize the deepest residual codebook, while shallower codebooks are initialized using standard K-Means.

M Impact of Contrastive Learning on the Encoder

Contrastive learning effectively reduces the spurious similarity caused by length imbalance between query and response, and avoids the artificial similarity introduced by simple weighted summation. Taking LaMP-2 as an example shown in Fig 6, where the query is a long movie description and the response is only 1–2 words, we observe that directly encoding the same query with different responses yields nearly identical embeddings (similarity = 0.970), making them hard to distinguish. While weighted fusion improves separability, it introduces extra similarity: the same response paired with different queries shows increased similarity (from 0.597 to 0.714), which is still suboptimal. In contrast, our contrastive learning approach shown in Fig 6(a) achieves the best separation by aligning the fused representation with its reconstructed counterpart which greatly helped the downstream RQ-VAE. Fig 6(b) further validates its effective-

ness: removing contrastive learning leads to clear drop in both Accuracy and F1.

N Implementation Details

We use Contriever (Izacard et al., 2021) as the default interaction encoder and Qwen2.5-7B-Instruct (Yang et al., 2024) as the decoder. In the extensibility experiments (Section 3), we additionally evaluate RoBERTa-large (Liu et al., 2019) as the encoder and LLaMA-3-8B-Instruct (Llama Team, 2024) as the decoder.

The codebook consists of $L = 3$ layers with codebook sizes of 32, 32, and 16 for layers 1, 2, and 3, respectively. Codebooks in layers 1 and 2 are initialized using standard K-Means on the training data, while layer 3 employs balanced K-Means initialization to promote uniform codeword usage. The commitment loss coefficient λ_v and the codebook update coefficient are both set to 0.25. The usage regularization weight λ_u in Equation (8) is set to 0.05. The temperature is set to 0.05. The random seed we use are 0, 1, 42. We use rouge_score package to calculate ROUGE.

The hyperparameters α and β for contrastive learning are set to 1.0 and 1.2, respectively. The default number of retrieved histories is $k = 8$. For field-aware user clustering, users are partitioned into $K = 10$ groups by default. We conduct contrastive learning for all tasks except LaMP-1 and LaMP-7 due to data format constraints.

The MLP consists of two layers with dimensions $\text{encoder_dim} \rightarrow \text{decoder_dim} \rightarrow \text{decoder_dim}$, using GELU activations. We optimize the model using AdamW with a batch size of 4 per GPU and a learning rate of 1×10^{-4} . Gradient accumulation with step size 8 is applied to simulate the target effective batch size. All experiments are conducted on 8 NVIDIA A100 (80GB) GPUs.

O Licenses

We use publicly available datasets and models that are already licensed. Our code will be released if accepted with MIT License.