

# Beyond Retrieval: Learning Compact User Representations for Scalable LLM Personalization

Heng Cao<sup>\*</sup>, Fan Zhang<sup>\*</sup>, Jian Yao<sup>†§</sup>, Yujie Zheng<sup>◇</sup>, Changlin Zhao<sup>\*</sup>, Lu Hao<sup>\*</sup>,  
Yuxuan Wei<sup>\*</sup>, Wangze Ni<sup>‡</sup>, Huaiyu Fu<sup>\*</sup>, Yuqian Sun<sup>\*</sup>, Xuyan Mo<sup>\*</sup>

<sup>\*</sup>Microsoft <sup>◇</sup>Shanghai International Studies University <sup>‡</sup>Zhejiang University  
<sup>†</sup>The Hong Kong Polytechnic University

{hengcao, zhanfa, neilz, luhao, yuxwei,

huaiyufu, yuqiansun, xuyanmo}@microsoft.com

nigel97.yao@connect.polyu.hk zhengyujie@shisu.edu.cn niwangze@zju.edu.cn

## Abstract

Personalizing large language models requires adapting model behavior to individual users while preserving robustness and deployment-scale efficiency. Existing approaches typically personalize LLMs either at the input level, by retrieving user histories or constructing profile prompts, or at the parameter level, by maintaining user-specific parameter-efficient modules. The former makes personalization sensitive to retrieval quality and prompt design, whereas the latter incurs storage and maintenance costs that grow with the user population. To address these limitations, we propose **TAP-PER**<sup>1</sup> (Temporal Attentive Prefix for **PER**sonalization), a prefix-based framework that encodes user preferences as learnable representations, avoiding the serialization of user histories into prompts and replacing heavy per-user adapters with lightweight user-state prefix embeddings. Inspired by personalized recommendation systems, TAP-PER decomposes user modeling into user-state and query-conditioned components, and incorporates temporal signals to capture the evolving nature of user interests. Experiments on six LaMP tasks show that TAP-PER consistently outperforms prompt-based and model-based baselines across classification, rating, and generation settings. Moreover, TAP-PER uses 130× fewer per-user parameters than OPPU and roughly half the total parameter footprint of PER-PCS at the 1,000-user scale, demonstrating scalable personalization without prompt-serialized histories or heavy per-user adapters.

## 1 Introduction

Large language models (LLMs) (Brown et al., 2020; Grattafiori et al., 2024; Yang et al., 2025) have achieved strong performance across a wide range of natural language processing tasks, yet their

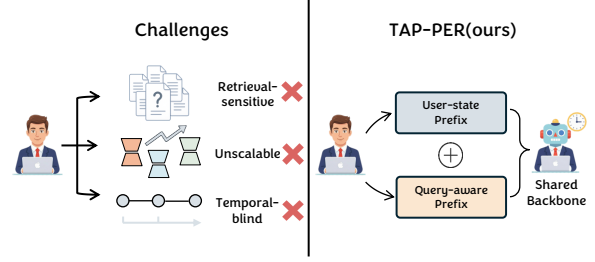


Figure 1: Existing personalized LLMs face three challenges: retrieval sensitivity, poor scalability, and temporal blindness (left). TAP-PER addresses them via a user-state prefix and a temporal-aware query-conditioned prefix injected into a shared backbone (right).

outputs often remain generic: the same input is typically answered in the same way regardless of the user’s long-term interests, stylistic preferences, or task-specific goals. This limitation has motivated growing interest in personalized LLMs (Liu et al., 2025a; Chen et al., 2024; Kirk et al., 2024), which aim to adapt model behavior to individual users while retaining the general capabilities of pretrained language models.

Existing approaches to LLM personalization largely fall into two categories. The first category personalizes LLMs at the input level. Retrieval-augmented generation (RAG) methods retrieve relevant user histories and prepend them to the query, while profile-augmented generation (PAG) methods summarize user histories into natural-language profiles (Chen et al., 2024; Li et al., 2023; Mysore et al., 2024; Richardson et al., 2023). These methods are flexible and require no parameter updates, but personalization is delegated to how user context is retrieved, summarized, and serialized into the prompt. The second category personalizes LLMs at the parameter level, typically through parameter-efficient fine-tuning (PEFT). For example, One PEFT Per User (OPPU) trains a dedicated adapter for each user (Tan et al., 2024b), while PER-PCS reduces redundancy by decom-

<sup>§</sup>Corresponding author.

<sup>1</sup>Code: <https://github.com/heng380/TAP-PER>.

posing and recombining personalized adapter components (Tan et al., 2024a). More recent methods use progressive LoRA-based mixture-of-experts or hypernetworks to generate personalized adapters from user profiles (Zhang et al., 2025; Tan et al., 2026). These methods provide a parameterized route to personalization by encoding user preferences directly in model parameters. Despite recent progress, existing personalized LLM methods still face three practical limitations.

**Retrieval sensitivity.** Prompt-based methods such as RAG and PAG are sensitive to retrieval and prompt quality. Table 1 shows that the optimal RAG/PAG setting varies across tasks, suggesting limited robustness and possible privacy risks from exposing user histories.

**Limited scalability.** PEFT-based methods often allocate dedicated parameters (e.g., LoRA modules) for individual users, resulting in a parameter footprint that grows with the user population. Representative approaches such as OPPU (Tan et al., 2024b) and Profile-to-PEFT (P2P) (Tan et al., 2026) either maintain user-specific parameters or generate them on-the-fly via a hypernetwork. PER-PCS (Tan et al., 2024a) reduces redundancy through collaborative decomposition, but still requires a sizable pool of personalized parameters. OPPU and PER-PCS are also commonly combined with optional RAG/PAG variants, while P2P consumes a textual user profile to generate an adapter, adding another user-context pipeline.

**Temporal unawareness.** Most methods treat histories as static context and overlook preference evolution, although recent interactions are often more indicative of current intent.

These limitations lead us to ask: *Can we achieve LLM personalization without serializing user histories into prompts or maintaining heavy per-user adapters, while remaining scalable to large populations?*

Inspired by personalized recommendation systems, which model user preferences as learned embeddings, we propose **TAP-PER** (Temporal Attentive Prefix for **PER**sonalization), a dual-track prefix framework for scalable LLM personalization. Recommendation systems have long shown that user preferences are multi-timescale, with persistent interests and transient, context-dependent intents (Zhou et al., 2018; Sun et al., 2020), and that temporal dynamics are crucial for sequential

user modeling (Li et al., 2020). As illustrated in Figure 1, TAP-PER follows this principle by decomposing personalization into a *user-state prefix* for long-term preferences and a *query-aware record prefix* for dynamically activating relevant historical records. It further incorporates relative time gaps and sequence order into record selection, and uses a shared LoRA-based bridge to integrate the prefix signals end-to-end, eliminating explicit prompt construction while keeping per-user parameters lightweight. In summary, our contributions are as follows:

- To address the retrieval sensitivity of prompt-based personalization and the scalability limitations of per-user model adaptation, we propose TAP-PER, a prefix-based framework that represents user preferences as compact learnable states. Instead of serializing user histories into prompts or maintaining heavy per-user adapters, TAP-PER shifts personalization toward end-to-end optimized preference representations.
- We design a personalized prefix mechanism that decomposes user preferences into two complementary signals: a user-state prefix for persistent preferences and a query-aware record prefix for instance-specific historical evidence. The record prefix uses temporal attention with learned time-gap and order-gap biases to capture evolving user interests, while a shared bridge LoRA integrates the personalized prefix signals into the task-adapted backbone.
- Across six LaMP tasks, TAP-PER consistently outperforms both prompt-based and model-based baselines while shrinking per-user storage by  $130\times$  relative to OPPU and cutting the total parameter footprint to roughly half of PER-PCS at the 1,000-user scale; its compact per-user prefix state additionally supports lightweight online adaptation from streaming feedback without retraining the shared backbone.

## 2 Method

**Task formulation** Following prior work (Tan et al., 2024b), we consider a personalized text generation setting over a set of users  $\mathcal{U}$ . Each user  $u \in \mathcal{U}$  is associated with an interaction history  $\mathcal{H}_u$ . Given an input query  $q$ , the goal is to generate an output  $r$  that reflects the user’s preferences.

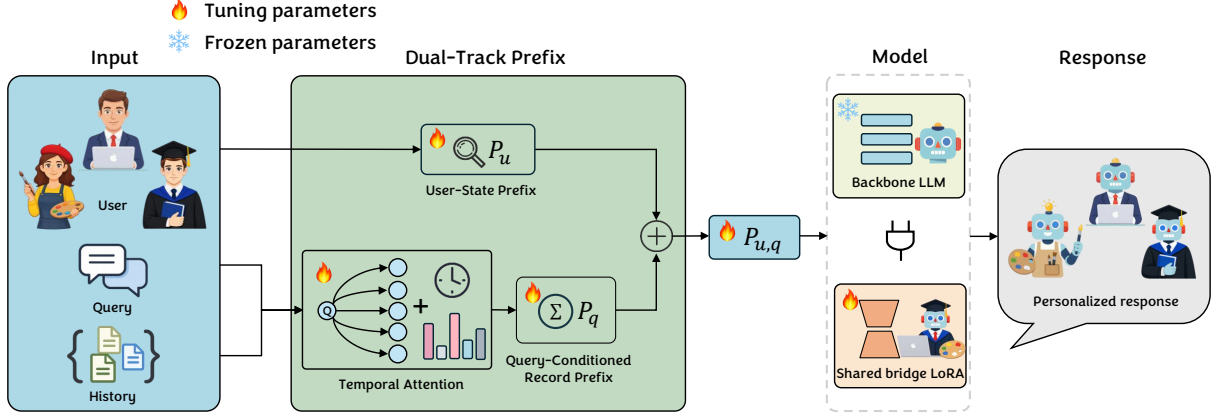


Figure 2: Overview of TAP-PER’s personalized prefix learning. With the task-adapted backbone  $\theta_{\text{task}}$  frozen, a user-state prefix  $\mathbf{P}_u$  is fetched from a per-user embedding table and a query-aware record prefix  $\mathbf{P}_q$  is produced by temporal-decayed attention over the user’s full interaction history. The two are summed and prepended to the query, while a shared bridge LoRA  $\Delta\theta_{\text{bridge}}$  is jointly trained to deeply condition the backbone on the prefix signals. Stage 1 (task-level LoRA adaptation) precedes this stage and is omitted from the diagram.

Formally, we learn a conditional generator:

$$r = \text{LLM}(q, u, \mathcal{H}_u; \theta), \quad (1)$$

where personalization is achieved by incorporating both user identity and historical interactions.

**Overall Framework** TAP-PER adopts a two-stage training paradigm: **Stage 1 (Task Adaptation)** fine-tunes the backbone on data aggregated across users to acquire shared task capabilities, and **Stage 2 (Personalized Prefix Learning)** introduces user-specific prefixes jointly optimized with a shared LoRA-based bridge. Figure 2 illustrates the Stage 2 pipeline.

### 2.1 Task-level Adaptation

In the first stage, we adapt the backbone model to the target task using data aggregated from all users. Each task is trained independently with its own Stage 1 pipeline. Specifically, we fine-tune the base LLM with Low-Rank Adaptation (LoRA) (Hu et al., 2022) on query–response pairs  $\{(q_i, r_i)\}$ , optimizing the cross-entropy loss

$$\mathcal{L}_{\text{task}} = \sum_i \text{CE}(\text{LLM}(q_i; \theta_{\text{task}}), r_i), \quad (2)$$

where  $\theta_{\text{task}}$  denotes the task-adapted parameters (i.e., the base model with LoRA updates), and  $\text{CE}(\cdot)$  is the cross-entropy loss. Following LoRA, the weight update for each target matrix is parameterized as  $\Delta W = B^{(t)} A^{(t)}$ , where  $A^{(t)} \in \mathbb{R}^{r \times k}$  and  $B^{(t)} \in \mathbb{R}^{d \times r}$  with rank  $r \ll \min(d, k)$ , while the backbone parameters remain frozen. After optimization, the learned LoRA parameters are merged

into the backbone:

$$W_{\text{task}} = W + B^{(t)} A^{(t)}, \quad (3)$$

which serves as the initialization for the subsequent personalization stage. Detailed hyperparameter settings are provided in Appendix A.1.

### 2.2 Personalized Prefix Learning

In Stage 2, we freeze the task-adapted backbone  $\theta_{\text{task}}$  and learn user-specific adaptations through a lightweight prefix-based parameterization, together with a shared bridge LoRA. By merging the Stage 1 LoRA into the backbone before this stage, we decouple task adaptation from personalization: the backbone retains general task capabilities, and the Stage 2 parameters focus exclusively on modeling user preferences without risking catastrophic forgetting of the learned task knowledge.

**User-State Prefix.** Let  $L$  denote the prefix length and  $d$  the hidden dimension of the backbone. We model persistent user characteristics via a learnable embedding table  $\mathbf{E} \in \mathbb{R}^{|\mathcal{U}| \times (Ld)}$ . For each user  $u$ , the embedding  $\mathbf{e}_u$  is reshaped into a prefix matrix  $\mathbf{P}_u \in \mathbb{R}^{L \times d}$ , which captures long-term preferences shared across queries.

**Query-Conditioned Record Prefix.** Given a query  $q$  and the user’s full interaction history  $\mathcal{H}_u = \{h_1, \dots, h_j, \dots, h_n\}$ , we first obtain the representation  $\mathbf{z}_{h_j}$  for each record  $h_j$  by mean-pooling its token embeddings from the frozen backbone’s embedding layer. Since the embedding layer is frozen, these record representations are pre-computed once

and cached, so that no per-step encoding of history records is required during training or inference. The query representation  $\mathbf{z}_q$  is obtained in the same manner. We then compute relevance scores over *all* historical records using a DIN-style attention mechanism (Zhou et al., 2018), dispensing with the separate retriever (e.g., BM25) used by RAG-style baselines:

$$s_j = \text{MLP}(\mathbf{z}_q \parallel \mathbf{z}_{h_j} \parallel (\mathbf{z}_q - \mathbf{z}_{h_j}) \parallel (\mathbf{z}_q \odot \mathbf{z}_{h_j})), \quad (4)$$

where  $\parallel$  denotes concatenation,  $\odot$  denotes element-wise product, and the MLP consists of two linear layers with GELU activation. The scores are then augmented with temporal bias:

$$\tilde{s}_j = s_j - \lambda_t \log(1 + \Delta t_j) - \lambda_o \log(1 + \Delta \pi_j), \quad (5)$$

where  $\Delta t_j$  denotes the time gap between the query and record  $h_j$ , and  $\Delta \pi_j$  denotes the order index gap between the query and record  $h_j$  in the user history. The decay coefficients  $\lambda_t$  and  $\lambda_o$  are learnable scalar parameters.

We normalize scores and aggregate historical signals:

$$\alpha_j = \text{softmax}(\tilde{s}_j), \quad (6)$$

and construct the query-conditioned prefix as

$$\mathbf{P}_q = \text{MLP}\left(\sum_{j=1}^n \alpha_j \mathbf{z}_{h_j}\right), \quad (7)$$

where the MLP projects the aggregated representation into the prefix space  $\mathbb{R}^{L \times d}$ . Letting the attention range over the full history rather than a pre-selected top- $k$  subset removes the dependence on an external retriever and lets the model learn end-to-end which records are relevant to the query; we empirically validate this design choice in §4.5.

**Prefix Composition.** We combine persistent and query-conditioned signals via element-wise addition:

$$\mathbf{P}_{u,q} = \mathbf{P}_u + \mathbf{P}_q. \quad (8)$$

This parameter-free fusion follows the multi-field intuition used in recommender systems and keeps the framework independent of task-specific fusion modules. Appendix A.6 compares it with concatenation, gated-sum, and MLP fusion; the more expressive alternatives do not yield consistent gains.

**Bridge LoRA.** Since prefix tokens interact with the backbone only through the attention mechanism, their influence on intermediate representations is limited. To enable deeper integration of personalized signals, we introduce a shared bridge LoRA  $\Delta\theta_{\text{bridge}}$ , applied to the query, key, value, and output projection layers of each transformer block. Unlike the Stage 1 LoRA, which is merged into the backbone to encode task knowledge, the bridge LoRA is initialized from scratch and trained jointly with the prefix parameters in Stage 2, conditioning the backbone’s internal representations on the prefix signals while keeping the number of additional shared parameters small. We empirically validate this design in Section 4.2, where removing the bridge causes the largest performance drop. Detailed hyperparameter settings including LoRA ranks are provided in Appendix A.1.

**Training Objective.** We optimize the prefix parameters (user embeddings  $\mathbf{E}$ , the attention MLP, decay coefficients  $\lambda_t, \lambda_o$ , and the projection MLP) together with the bridge LoRA  $\Delta\theta_{\text{bridge}}$ :

$$\mathcal{L}_{\text{pers}} = \sum_{(u,q,r)} \text{CE}(\text{LLM}(\mathbf{P}_{u,q}, q; \theta'), r), \quad (9)$$

where  $\theta' = \theta_{\text{task}} + \Delta\theta_{\text{bridge}}$ .  $\mathbf{P}_u$  and  $\mathbf{P}_q$  can be viewed as parameterized counterparts of PAG and RAG, respectively, encoded as end-to-end-optimized prefixes rather than natural-language prompts.

### 3 Experimental Setup

**Datasets.** Following prior work (Tan et al., 2024b; Zhang et al., 2025), we conduct experiments on the Large Language Model Personalization (LaMP) benchmark (Salemi et al., 2024b). LaMP covers a diverse set of personalization tasks spanning classification, regression, and text generation. We adopt the same data split as in OPPU (Tan et al., 2024b). Detailed task descriptions and dataset statistics are provided in Appendix A.2 and A.3.

**Baselines.** We compare TAP-PER with two families of personalization baselines: prompt-based methods, including RAG (Salemi et al., 2024b; Lewis et al., 2020) and PAG (Richardson et al., 2023); and model-based methods, including OPPU (Tan et al., 2024b), PER-PCS (Tan et al., 2024a), P2P (Tan et al., 2026), and PROPER (Zhang et al., 2025). P2P encodes a

| Task                                         | Metric | Prompt-based |       |       |       |       | Model-based |       |       |         |       |       |        |                   |                              |                              |
|----------------------------------------------|--------|--------------|-------|-------|-------|-------|-------------|-------|-------|---------|-------|-------|--------|-------------------|------------------------------|------------------------------|
|                                              |        | RAG          |       |       | PAG   |       | OPPU        |       |       | PER-PCS |       |       | P2P    | TAP-PER           |                              |                              |
|                                              |        | k=1          | k=2   | k=4   | k=0   | k=1   | base        | +rag  | +pag  | base    | +rag  | +pag  | single | base <sup>†</sup> | +P <sub>q</sub> <sup>‡</sup> | +P <sub>u</sub> <sup>§</sup> |
| LaMP-1: PERSONALIZED CITATION IDENTIFICATION | Acc ↑  | 0.659        | 0.642 | 0.659 | 0.683 | 0.675 | 0.654       | 0.675 | 0.715 | 0.626   | 0.626 | 0.691 | 0.724  | 0.642             | 0.699                        | 0.732                        |
|                                              | F1 ↑   | 0.659        | 0.645 | 0.664 | 0.683 | 0.675 | 0.656       | 0.674 | 0.711 | 0.619   | 0.620 | 0.685 | 0.720  | 0.642             | 0.696                        | 0.728                        |
| LaMP-2N: PERSONALIZED NEWS CATEGORIZE        | Acc ↑  | 0.790        | 0.796 | 0.803 | 0.787 | 0.790 | 0.794       | 0.809 | 0.817 | 0.748   | 0.798 | 0.809 | 0.803  | 0.789             | 0.808                        | 0.838                        |
|                                              | F1 ↑   | 0.542        | 0.560 | 0.571 | 0.532 | 0.546 | 0.544       | 0.596 | 0.600 | 0.505   | 0.589 | 0.608 | 0.598  | 0.528             | 0.575                        | 0.622                        |
| LaMP-2M: PERSONALIZED MOVIE TAGGING          | Acc ↑  | 0.586        | 0.592 | 0.598 | 0.592 | 0.596 | 0.679       | 0.542 | 0.521 | 0.440   | 0.520 | 0.551 | 0.645  | 0.595             | 0.657                        | 0.751                        |
|                                              | F1 ↑   | 0.545        | 0.549 | 0.553 | 0.550 | 0.552 | 0.584       | 0.438 | 0.422 | 0.400   | 0.450 | 0.471 | 0.601  | 0.551             | 0.596                        | 0.682                        |
| LaMP-3: PERSONALIZED PRODUCT RATING          | MAE ↓  | 0.420        | 0.375 | 0.393 | 0.304 | 0.357 | 0.218       | 0.223 | 0.250 | 0.393   | 0.205 | 0.188 | 0.201  | 0.286             | 0.188                        | 0.170                        |
|                                              | RMSE ↓ | 0.796        | 0.744 | 0.866 | 0.598 | 0.694 | 0.453       | 0.430 | 0.540 | 0.581   | 0.491 | 0.472 | 0.507  | 0.567             | 0.453                        | 0.430                        |
| LaMP-4: PERSONALIZED NEWS HEADLINE GEN.      | R-1 ↑  | 0.198        | 0.202 | 0.201 | 0.193 | 0.200 | 0.202       | 0.207 | 0.195 | 0.188   | 0.197 | 0.201 | 0.209  | 0.202             | 0.212                        | 0.219                        |
|                                              | R-L ↑  | 0.178        | 0.182 | 0.182 | 0.173 | 0.179 | 0.182       | 0.188 | 0.175 | 0.173   | 0.181 | 0.184 | 0.192  | 0.182             | 0.191                        | 0.198                        |
| LaMP-5: PERSONALIZED SCHOLARLY TITLE GEN.    | R-1 ↑  | 0.503        | 0.502 | 0.516 | 0.491 | 0.512 | 0.502       | 0.510 | 0.491 | 0.457   | 0.472 | 0.470 | 0.495  | 0.512             | 0.507                        | 0.538                        |
|                                              | R-L ↑  | 0.451        | 0.454 | 0.464 | 0.441 | 0.463 | 0.450       | 0.460 | 0.439 | 0.405   | 0.422 | 0.421 | 0.450  | 0.462             | 0.459                        | 0.485                        |

Table 1: Comparison results on LaMP benchmark across prompt-based and model-based personalization. **Bold** and underline denote the best and second-best result per metric, respectively.  $k$  denotes the number of retrieved items, with  $k=0$  indicating no retrieval.  $\uparrow$  indicates higher is better;  $\downarrow$  indicates lower is better. For OPPU and PER-PCS, the columns are cumulative:  $+rag$  adds RAG on top of *base*, and  $+pag$  further adds PAG on top of  $+rag$ . For P2P, *single* denotes the single hypernetwork-generated LoRA variant without retrieval or profile stacking. For TAP-PER:  $^{\dagger}base$  = task-adapted model (Stage 1 only);  $^{\ddagger}+P_q$  = adding query-aware record prefix;  $^{\S}+P_u$  = further adding user-state prefix (full model). The  $P_q$  and  $P_u$  columns parallel the  $+rag$  and  $+pag$  columns of OPPU/PER-PCS, as they serve analogous roles through learned prefix representations.

profile built from user history and uses a hypernetwork to generate user-specific LoRA parameters. PROPER progressively combines population-, group-, and user-level adaptation through routed LoRA experts; because adding its results would make the already dense main table unreadable, we report the six-task comparison in Appendix A.5. We use Llama-3.1-8B (Grattafiori et al., 2024) as the backbone for all methods. BM25 (Trotman et al., 2014) is used for retrieval-based baselines, whereas TAP-PER directly attends over each user’s full history without an external retriever. More details are provided in Appendix A.4 and A.1.

**Evaluation Metrics.** Following LaMP (Salemi et al., 2024b), we use accuracy and F1-score for classification tasks (LaMP-1, LaMP-2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for LaMP-3, and ROUGE-1 and ROUGE-L (Lin, 2004) for text generation tasks (LaMP-4, LaMP-5). Higher values indicate better performance for all metrics, except for MAE and RMSE where lower values are better.

## 4 Experimental Results

We evaluate TAP-PER on the LaMP benchmark through six questions:

**RQ1:** Does TAP-PER outperform prompt-based and model-based personalization baselines?

**RQ2:** What does each component contribute?

**RQ3:** How important are temporal signals?

**RQ4:** How scalable is TAP-PER?

**RQ5:** How sensitive is TAP-PER to retrieval pre-filtering and prefix length?

**RQ6:** Can TAP-PER support online adaptation?

### 4.1 Main Results (RQ1)

Table 1 compares TAP-PER with prompt-based and model-based personalization baselines. TAP-PER achieves the best performance on all task-metric pairs across the six LaMP tasks. The gains are especially clear on tasks requiring stronger personalization: compared with the best OPPU variant, TAP-PER improves accuracy by 7.2 points on LaMP-2M and ROUGE-1 by 2.8 points on LaMP-5. We further verify the same trend with an alternative Qwen3-4B backbone in Appendix C.2.

The results also expose the instability of prompt-based personalization: the best RAG size, the usefulness of PAG, and whether the two should be combined vary substantially across tasks, and this sensitivity persists when RAG/PAG is stacked on OPPU or PER-PCS. By replacing prompt-time context with learned prefixes, TAP-PER avoids this brittleness; the incremental gains from *base* to  $+P_q$  to  $+P_u$  further indicate that query-aware and user-state modeling are complementary (Section 4.2).

### 4.2 Component Ablation (RQ2)

To answer RQ2, we ablate the bridge LoRA, the query-aware record prefix  $P_q$ , and the user-state prefix  $P_u$  from the full TAP-PER model. As shown in Table 2, removing any component degrades or

| Task    | Metric            | Base  | Full         | −bridge | − $\mathbf{P}_q$ | − $\mathbf{P}_u$ |
|---------|-------------------|-------|--------------|---------|------------------|------------------|
| LaMP-1  | Acc $\uparrow$    | 0.642 | <b>0.732</b> | 0.659   | 0.699            | 0.699            |
|         | F1 $\uparrow$     | 0.642 | <b>0.728</b> | 0.659   | 0.695            | 0.696            |
| LaMP-2N | Acc $\uparrow$    | 0.789 | <b>0.838</b> | 0.814   | 0.837            | 0.808            |
|         | F1 $\uparrow$     | 0.528 | <b>0.622</b> | 0.595   | 0.611            | 0.575            |
| LaMP-2M | Acc $\uparrow$    | 0.595 | <b>0.751</b> | 0.702   | 0.733            | 0.657            |
|         | F1 $\uparrow$     | 0.551 | <b>0.682</b> | 0.624   | 0.659            | 0.596            |
| LaMP-3  | MAE $\downarrow$  | 0.286 | <b>0.170</b> | 0.241   | 0.241            | 0.188            |
|         | RMSE $\downarrow$ | 0.567 | <b>0.430</b> | 0.526   | 0.543            | 0.453            |
| LaMP-4  | R-1 $\uparrow$    | 0.202 | <b>0.219</b> | 0.211   | 0.218            | 0.212            |
|         | R-L $\uparrow$    | 0.182 | <b>0.198</b> | 0.190   | <b>0.198</b>     | 0.191            |
| LaMP-5  | R-1 $\uparrow$    | 0.512 | <b>0.538</b> | 0.508   | 0.523            | 0.507            |
|         | R-L $\uparrow$    | 0.462 | <b>0.485</b> | 0.449   | 0.473            | 0.459            |

Table 2: Component ablation of TAP-PER. *Base* is the Stage-1 task-adapted model. *Full* is the complete TAP-PER with all components. *−bridge*, *− $\mathbf{P}_q$* , and *− $\mathbf{P}_u$*  denote removing the bridge LoRA, query-aware record prefix, and user-state prefix, respectively. **Bold** denotes the best result per row.

does not improve performance, confirming that they provide complementary gains. The bridge LoRA is important for integrating prefix signals into the task-adapted backbone (e.g., removing it causes large drops on LaMP-1 and LaMP-5), while  $\mathbf{P}_u$  and  $\mathbf{P}_q$  capture different personalization cues:  $\mathbf{P}_u$  is more helpful on tasks requiring stable user preferences (e.g., LaMP-2M and LaMP-5), whereas  $\mathbf{P}_q$  is especially important when query-specific historical evidence is needed (e.g., LaMP-3).

### 4.3 Temporal Analysis (RQ3)

To answer RQ3, we examine the effect of the temporal biases in the query-aware record prefix through coefficient ablation and attention visualization.

**Temporal-bias ablation.** We ablate the time-gap and order-gap decay coefficients,  $\lambda_t$  and  $\lambda_o$ , on LaMP-2M and LaMP-5. As shown in Table 3, the full model performs best across all metrics. Removing both biases causes the largest drop on LaMP-2M, while disabling either one gives intermediate results, indicating that both temporal recency and sequence order contribute to personalization. The effect is smaller on LaMP-5, suggesting that the usefulness of temporal cues is task-dependent.

**Attention case study.** Figure 3 visualizes a representative LaMP-2M example. Without temporal biases, TAP-PER fails to emphasize recent query-relevant records and predicts an incorrect tag, similar to the non-personalized baseline. In contrast, the full model assigns higher attention to recent

| Task    | Metric         | Base  | Full         | $\lambda_t=0$ | $\lambda_o=0$ | $\lambda_t=\lambda_o=0$ |
|---------|----------------|-------|--------------|---------------|---------------|-------------------------|
| LaMP-2M | Acc $\uparrow$ | 0.595 | <b>0.751</b> | 0.723         | 0.733         | 0.714                   |
|         | F1 $\uparrow$  | 0.551 | <b>0.682</b> | 0.641         | 0.660         | 0.635                   |
| LaMP-5  | R-1 $\uparrow$ | 0.512 | <b>0.538</b> | 0.516         | 0.511         | 0.513                   |
|         | R-L $\uparrow$ | 0.462 | <b>0.485</b> | 0.464         | 0.452         | 0.464                   |

Table 3: Ablation on temporal biases in the query-aware record prefix. *Base* is the Stage-1 task-adapted model. *Full* is the complete TAP-PER.  $\lambda_t=0$  disables the time-gap decay,  $\lambda_o=0$  disables the order-gap decay, and  $\lambda_t=\lambda_o=0$  removes both temporal biases while retaining content-based attention. **Bold** denotes the best result per row.

relevant histories and recovers the ground-truth tag. This qualitative pattern is consistent with the ablation results in Table 3.

### 4.4 Parameter Efficiency (RQ4)

We evaluate scalability from two aspects: storage cost and per-user training time, comparing TAP-PER with OPPU and PER-PCS in Figure 4.

**Storage efficiency.** OPPU maintains a dedicated LoRA adapter for each user ( $\sim 4.2\text{M}$  parameters), whereas TAP-PER stores only a 32K user-state prefix per user, yielding a  $130\times$  reduction in per-user storage. TAP-PER’s remaining footprint is *shared* across all users: a bridge LoRA ( $\sim 4.2\text{M}$ ) plus the DIN-style attention and projection MLPs ( $\sim 201\text{M}$ ), totaling  $\sim 205\text{M}$  shared parameters. As shown in Figure 4 (top), with 1,000 users, TAP-PER uses only 5.7% of OPPU’s total parameters and about 53% of PER-PCS’s, showing that it scales mainly through lightweight user embeddings rather than large personalized adapter modules.

This trainable state is separate from the non-trainable history cache. With FP16 record vectors ( $d=4096$ ),  $H$  cached records require  $2Hd$  bytes per user, or approximately 0.8/4.1/8.2 GB (decimal) in total for 1,000 users with  $H=100/500/1000$ . The cache therefore grows linearly with retained history and can be pruned or quantized independently of the trainable prefix.

**Training time.** For each user, TAP-PER only updates the user-state prefix  $\mathbf{P}_u$ , while the query-aware prefix  $\mathbf{P}_q$  is produced by shared parameters. As shown in Figure 4 (bottom), this reduces per-user training time by roughly half compared with OPPU. PER-PCS is faster because it does not perform per-user gradient updates, but TAP-PER offers a stronger accuracy-efficiency trade-off. In addition, the  $\mathbf{P}_q$ -only variant requires no per-user

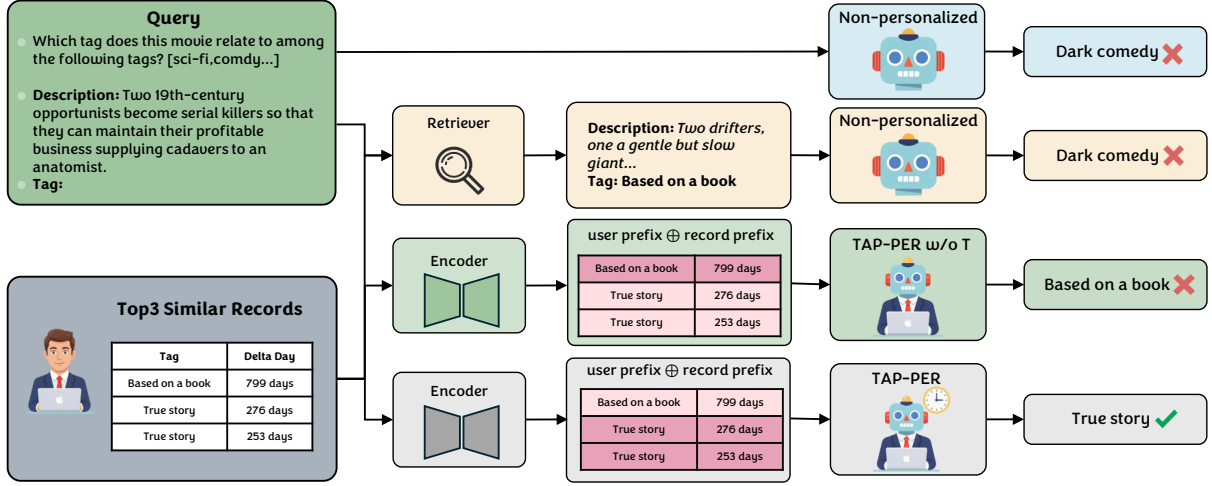


Figure 3: Case study on a representative query from LaMP-2M. The non-personalized baseline and the temporal-free TAP-PER variant predict incorrect tags, whereas the full TAP-PER correctly identifies the ground-truth tag by attending to recent query-relevant records. Darker red indicates higher attention weight.

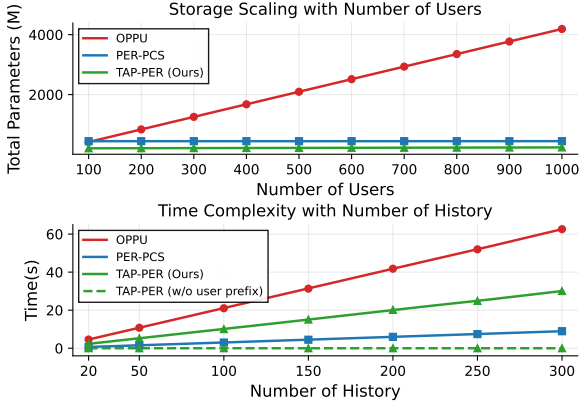


Figure 4: Storage and time scaling comparison. Top: total parameter count as a function of user population. Bottom: per-user training time as a function of history length on LaMP-2M; the dashed green line denotes the TAP-PER variant without user-state prefix  $\mathbf{P}_u$ , which requires no per-user gradient updates.

training, providing a lightweight option when on-line user optimization is impractical.

**Inference efficiency.** On one A800 GPU, full-history attention plus prefix projection takes 0.73/0.74/1.04 ms for 100/500/1000 cached records. For comparison, Llama-3.1-8B has a 108.14 ms time to first token for a 1,024-token prompt and 28.9 ms/token decoding latency; with a 32-token output, the 1,000-record prefix computation is about 0.1% of total generation time. TAP-PER also avoids the extra prompt tokens introduced by RAG/PAG. Appendix C.1 reports per-user update time on the two generation tasks.

#### 4.5 Sensitivity Analysis (RQ5)

To answer RQ5, we study two design choices: history scope and prefix length. For history scope, we insert a BM25 pre-filter before temporal attention and vary the retained records  $k \in \{10, 20, 30, 50, \text{All}\}$ , where “All” denotes our default setting over the full user history. For prefix length, we vary  $L \in \{2, 4, 8, 16, 32\}$ . We evaluate on a representative classification task (LaMP-2M) and a representative generation task (LaMP-5), with results shown in Figure 5.

**History scope.** The “All” setting performs best on both tasks. Adding a BM25 pre-filter consistently reduces performance, and varying  $k$  within  $\{10, 20, 30, 50\}$  does not recover the full-history result. This suggests that explicit pre-retrieval is unnecessary: the temporal attention in  $\mathbf{P}_q$  can learn to down-weight less relevant records while preserving useful signals that a sparse retriever may discard.

**Prefix length.** Performance peaks at a moderate prefix length. Short prefixes underfit user-specific signals, while longer prefixes bring little additional gain. We therefore use the full-history setting with  $L = 8$  as the default configuration.

#### 4.6 Online Adaptation to Streaming Feedback (RQ6)

A deployed personalization system should absorb new feedback without full retraining. Since TAP-PER stores user-specific state in a compact  $L \times d$  prefix  $\mathbf{P}_u$  ( $\sim 32\text{K}$  scalars), it can be updated online by optimizing only  $\mathbf{P}_u$ , leaving all shared param-

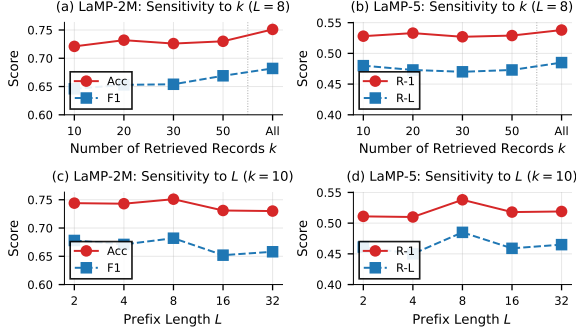


Figure 5: Sensitivity of TAP-PER to the number of retrieved records  $k$  (top, with “All” denoting our default setting that attends over the full user history) and the prefix length  $L$  (bottom). Each subplot shows two metrics: Acc (solid) and F1 (dashed) for LaMP-2M, and R-1 (solid) and R-L (dashed) for LaMP-5.

eters unchanged. To answer RQ6, we simulate streaming feedback on LaMP-2M and LaMP-5.

**Protocol.** For each evaluation user, we split the training history chronologically into two equal halves: the earlier 50% is used for bootstrap (Stage 2) training, and the later 50% is partitioned into  $T=5$  equally sized chunks  $\{C_1, \dots, C_T\}$  that simulate streaming feedback. The original test set is kept unchanged and used for evaluation throughout. After bootstrap training, chunks arrive sequentially; at step  $t$ , each method updates on  $C_t$  with the same SGD budget and is then evaluated on the fixed test set, yielding a trajectory  $R_0, \dots, R_T$ .

**Variants.** We compare four variants initialized from the same bootstrap checkpoint: **Static**, which performs no online update; **Online-Emb**, which updates only the per-user prefix  $P_u$ ; **Online-Full**, which updates  $P_u$  together with the shared  $P_q$  network and bridge LoRA; and **Full-Retrain**, which retrains TAP-PER from scratch on all data observed so far and serves as an upper bound.

**Results.** As shown in Figure 6, Static remains flat, confirming that the model cannot benefit from new feedback without online updates. Online-Full nearly matches Full-Retrain at the final step, with less than  $\sim 1$  point difference on both tasks. This shows that incremental updates can recover most of the benefit of full retraining without replaying past data.

Online-Emb provides a more lightweight alternative. By updating only the  $\sim 32K$ -parameter user prefix, it steadily improves on LaMP-5, where each user has dense histories, but gains less on LaMP-

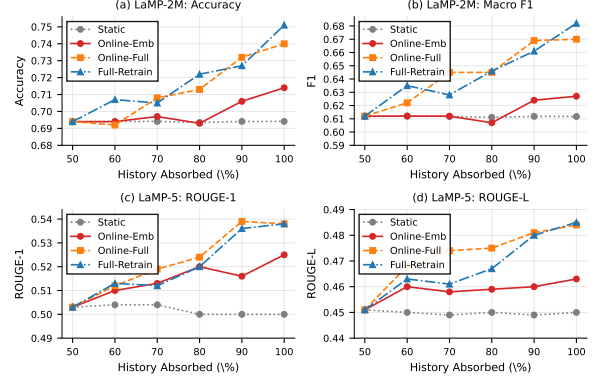


Figure 6: Online adaptation to streaming feedback on LaMP-2M (top) and LaMP-5 (bottom).  $x$ -axis: fraction of training history absorbed (50% = bootstrap, 60–100% = after each of the five streaming chunks);  $y$ -axis: test-set metric.

2M, where each feedback chunk contains only a few samples. Overall, TAP-PER provides a feasible online-learning pathway: dense user histories can be handled by updating only  $P_u$ , while sparse histories may benefit from updating the shared personalization modules, without modifying the backbone or retraining from scratch. This experiment is a controlled probe with chronologically ordered feedback and a fixed test set rather than a full production simulation; robustness to noisy feedback is analyzed in Appendix B.3.

## 5 Related Work

**LLM Personalization.** Existing approaches fall into two families (Liu et al., 2025a; Chen et al., 2024; Kirk et al., 2024). Prompt-based methods inject user context via retrieval (Salemi et al., 2024b; Mysore et al., 2024) or natural-language profiles (Richardson et al., 2023; Li et al., 2023), but are sensitive to retrieval and prompt construction. Model-based methods encode preferences into PEFT parameters: OPPU (Tan et al., 2024b) trains one adapter per user; PROPER (Zhang et al., 2025) progressively combines population-, group-, and user-level LoRA experts; PER-PCS (Tan et al., 2024a) reassembles shared adapter pieces; and P2P (Tan et al., 2026) generates adapters from profiles via a hypernetwork. HYDRA (Zhuang et al., 2024) factorizes reranking and adaptation into shared bases and personal heads for a black-box target LLM. These methods reduce different deployment costs, but still maintain, generate, or assemble personalized parameters and may retain a retrieval/profile pipeline. Complementary methods

maintain context outside the backbone: PersonaDB (Sun et al., 2025) refines hierarchical personas collaboratively, and LoGo (Wang et al., 2025) combines local personal memory and collective global memory. Other work optimizes retrieval (Salemi et al., 2024a), elicits preferences through dialogue (Wu et al., 2025), or merges preference-specific checkpoints post-hoc (Jang et al., 2023). Earlier UserAdapter (Zhong et al., 2021) attached small per-user modules to pretrained encoders. Closer to our representation-based view, PPlug (Liu et al., 2025b) encodes a user’s history into one input-level embedding, while CURP (Wang et al., 2026) extracts multi-dimensional traits using a user encoder and prototype codebook. TAP-PER instead combines a persistent user-state prefix with a query-aware temporal record prefix, integrated by a shared bridge LoRA.

**Prefix and Prompt Tuning.** Prefix and prompt tuning (Li and Liang, 2021; Lester et al., 2021; Liu et al., 2022) steer frozen transformers via learnable continuous vectors, alongside PEFT variants such as LoRA (Hu et al., 2022), adapters (Houlsby et al., 2019), and BitFit (Ben Zaken et al., 2022). HyperPrompt (He et al., 2022) generates task-conditioned prompts with a hypernetwork. In personalization, SPT (Huang et al., 2024) learns a bank of soft prompts and a context-conditioned retriever for dialogue. TAP-PER instead treats each user as a learnable prefix and pairs it with a query-aware temporal record prefix.

**Sequential Modeling in Recommendation.** Personalized recommendation has long modeled temporal dynamics and query-aware user interests. Early sequential recommenders encode interaction sequences with recurrent or convolutional networks (Hidasi et al., 2016; Tang and Wang, 2018), and self-attentive successors such as SASRec (Kang and McAuley, 2018) and BERT4Rec (Sun et al., 2019) capture long-range dependencies within a user’s history. DIN (Zhou et al., 2018) introduces query-conditioned attention over user histories, DIEN (Zhou et al., 2019) models the evolution of user interests over time, and TiSASRec (Li et al., 2020) incorporates time intervals into self-attention to distinguish recent and distant interactions. More recently, LLMs have been applied directly to recommendation, either as rating predictors or preference reasoners (Dai et al., 2023; Kang et al., 2023; Petrov and MacDonald, 2023; Wu et al., 2024). TAP-PER ports

these inductive biases to LLM personalization via DIN-style attention with learnable time-gap and order-gap decays, complementing the user-state prefix with context-dependent evidence.

## 6 Conclusion

We presented TAP-PER, a prefix-based framework that reformulates LLM personalization as learned user representation modeling. TAP-PER replaces prompt-serialized user histories and heavy per-user adapters with lightweight prefix states, combining a user-state prefix for persistent preferences, a query-aware temporal record prefix over cached histories, and a shared bridge LoRA for effective integration with the task-adapted backbone. Our results suggest that effective LLM personalization need not rely on retrieval-time prompt engineering or heavy user-specific adapter modules: a compact, end-to-end-optimized user representation can match or surpass both families while reducing per-user trainable storage by  $130\times$  over OPPU and the total parameter footprint to about half of PER-PCS at the 1,000-user scale. For an unseen user, the shared query-conditioned prefix provides an immediate fallback, and a user-state prefix can be trained asynchronously once sufficient interactions accumulate.

## Limitations

TAP-PER has several limitations. First, although  $P_q$  attends over the user’s full cached history without an external retriever, cache storage still grows linearly with history length, and attention quality may degrade for extremely long or noisy histories. We have not studied cache pruning, quantization, or hybrid learned retrieval in such regimes. Second, our evaluation focuses on six LaMP tasks. We exclude LaMP-7 because its unpaired history tweets and rewrite queries do not provide the task-consistent history input–output pairs assumed by our supervised Stage 2; LongLaMP would likewise require adapting its protocol and history construction. Evaluation on these and other benchmarks remains future work. Third, our experiments cover Llama-3.1-8B and Qwen3-4B only, and we have not verified whether the gains hold at substantially larger backbone sizes. The online study is also a controlled simulation with chronological feedback and a fixed clean test set; despite the added label-noise probe, production feedback can be delayed, biased, non-stationary, or adversarial. Finally, un-

seen users initially lack  $\mathbf{P}_u$ . Our shared  $\mathbf{P}_q$  fallback performs well for held-out users with histories, but users with no interaction history receive only the task-level model until evidence becomes available, and we do not inductively generate  $\mathbf{P}_u$  from a profile.

## Ethical Impact

TAP-PER is designed to reduce direct exposure of raw histories: user preferences are encoded into continuous prefix embeddings and a shared bridge LoRA rather than injected verbatim into model inputs. This may reduce one surface for prompt-level leakage, but it does not establish a formal privacy guarantee (Carlini et al., 2021). The per-user embeddings and history cache still encode behavioral patterns derived from private histories and should be treated as sensitive artifacts: they require encryption, access control, retention limits, and deletion mechanisms, and users should retain the ability to inspect, update, or delete their personalized representations.

Beyond privacy, personalized LLMs can amplify existing biases or reinforce filter-bubble effects by over-fitting to past behavior, and the temporal decay used in TAP-PER—by design—up-weights recent interactions, which may further narrow the range of content surfaced to a user. We therefore recommend periodic auditing of model outputs across demographic and interest groups, together with transparent controls over the strength of personalization, before deploying frameworks like TAP-PER in user-facing applications. All datasets used in this work are publicly released research benchmarks (LaMP), and no additional user data was collected for this study.

## Acknowledgments

This work is supported by the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (No. 2026C02A1236).

## References

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. [BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie

Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. [Extracting training data from large language models](#). In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

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

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. [The entropy mechanism of reinforcement learning for reasoning language models](#). *arXiv preprint arXiv:2505.22617*.

Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering ChatGPT’s capabilities in recommender systems. *arXiv preprint arXiv:2305.02182*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.

Yun He, Steven Zheng, Yi Tay, Jai Gupta, Yu Du, Vamsi Aribandi, Zhe Zhao, YaGuang Li, Zhao Chen, Donald Metzler, Heng-Tze Cheng, and Ed H Chi. 2022. [HyperPrompt: Prompt-based task-conditioning of transformers](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8678–8690. PMLR.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea

- Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799.
- 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 *International Conference on Learning Representations*.
- Qiushi Huang, Xubo Liu, Tom Ko, Bo Wu, Wenwu Wang, Yu Zhang, and Lilian Tang. 2024. [Selective prompting tuning for personalized conversations with LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16212–16226. Association for Computational Linguistics.
- Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. [Personalized soups: Personalized large language model alignment via post-hoc parameter merging](#). *arXiv preprint arXiv:2310.11564*.
- Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE.
- 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. *arXiv preprint arXiv:2305.06474*.
- Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024. [The benefits, risks and bounds of personalizing the alignment of large language models to individuals](#). *Nature Machine Intelligence*, 6(4):383–392.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- 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*.
- Jiacheng Li, Yujie Wang, and Julian McAuley. 2020. Time interval aware self-attention for sequential recommendation. In *Proceedings of the 13th international conference on web search and data mining*, pages 322–330.
- 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.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
- Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025a. [A survey of personalized large language models: Progress and future directions](#). *arXiv preprint arXiv:2502.11528*.
- Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025b. LLMs + persona-plugin = personalized LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9373–9385.
- Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68.
- Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. [Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers](#). In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 198–219, Miami, Florida, USA. Association for Computational Linguistics.
- Aleksandr V Petrov and Craig Macdonald. 2023. Generative sequential recommendation with GPTRec. In *Gen-IR Workshop at SIGIR*.
- 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*.
- Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024a. Optimization methods for personalizing large language models through retrieval augmentation. In *Proceedings of SIGIR*.

- 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.
- Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi Ren 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.
- Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1441–1450.
- Ke Sun, Tiejun Qian, Tong Chen, Yile Liang, Quoc Viet Hung Nguyen, and Hongzhi Yin. 2020. Where to go next: Modeling long-and short-term user preferences for point-of-interest recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 214–221.
- 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*, pages 6459–6475. 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*, pages 6476–6491. Association for Computational Linguistics.
- Zhaoxuan Tan, Zixuan Zhang, Haoyang Wen, Zheng Li, Rongzhi Zhang, Pei Chen, Fengran Mo, Zheyuan Liu, Qingkai Zeng, Qingyu Yin, and Meng Jiang. 2026. [Instant personalized large language model adaptation via hypernetwork](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23557–23580, San Diego, California, United States. Association for Computational Linguistics.
- Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of WSDM*.
- Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, pages 58–65.
- 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*.
- Zehong Wang, Junlin Wu, Zhaoxuan Tan, Bolian Li, Xianrui Zhong, Zheli Liu, and Qingkai Zeng. 2025. From personal to collective: On the role of local and global memory in llm personalization. *arXiv preprint arXiv:2509.23767*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. 2024. [A survey on large language models for recommendation](#). *World Wide Web*, 27(5):60.
- Shujin Wu, Yi R. Fung, Cheng Qian, Jeonghwan Kim, Dilek Hakkani-Tur, and Heng Ji. 2025. Aligning LLMs with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. 2025. [Diversity-aware policy optimization for large language model reasoning](#). *arXiv preprint arXiv:2505.23433*.
- Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. 2025. [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.
- Wanjuan Zhong, Duyu Tang, Jiahai Wang, Jian Yin, and Nan Duan. 2021. [Useradapter: Few-shot user learning in sentiment analysis](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1484–1488. Association for Computational Linguistics.
- Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019.

Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5941–5948.

Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1059–1068.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. [HY-DRA: Model factorization framework for black-box LLM personalization](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 100783–100815.

| Task    | Rank | Epochs | learning rate | batch size |
|---------|------|--------|---------------|------------|
| LaMP-1  | 8    | 3      | $3e^{-4}$     | 8          |
| LaMP-2N | 8    | 3      | $3e^{-4}$     | 8          |
| LaMP-2M | 8    | 3      | $3e^{-4}$     | 8          |
| LaMP-3  | 8    | 3      | $5e^{-4}$     | 32         |
| LaMP-4  | 8    | 3      | $3e^{-4}$     | 8          |
| LaMP-5  | 8    | 3      | $3e^{-4}$     | 8          |

Table 4: Stage 1 (Task Adaptation) hyperparameters.

| Task    | Rank | Epochs | learning rate | batch size |
|---------|------|--------|---------------|------------|
| LaMP-1  | 4    | 3      | $2e^{-4}$     | 32         |
| LaMP-2N | 4    | 3      | $1e^{-4}$     | 16         |
| LaMP-2M | 4    | 3      | $1e^{-4}$     | 8          |
| LaMP-3  | 4    | 3      | $2e^{-4}$     | 32         |
| LaMP-4  | 4    | 3      | $2e^{-4}$     | 32         |
| LaMP-5  | 4    | 3      | $1e^{-4}$     | 16         |

Table 5: Stage 2 (Personalized Prefix Learning) hyperparameters. The prefix length  $L=8$  is fixed across all tasks, and  $\mathbf{P}_q$  attends over the user’s full history (no retriever).

## A Experimental Details

### A.1 Hyperparameter Settings

Tables 4 and 5 summarize the hyperparameter settings for Stage 1 (task adaptation) and Stage 2 (personalized prefix learning), respectively. Both stages apply LoRA to the query, key, value, and output projection layers. The Stage 1 task LoRA is merged into the backbone after training, while the Stage 2 bridge LoRA remains as a separate shared module.

### A.2 LaMP Task Descriptions

We employ the LaMP benchmark (Salemi et al., 2024b), which encompasses a diverse set of personalization tasks covering both text classification and generation. We evaluate on the following tasks:

- **LaMP-1 (Personalized Citation Identification):** A binary text classification task that involves predicting citation choices. Given a new paper and two candidate references, the model determines which paper will be cited based on the user’s historical publications. This task evaluates a model’s ability to capture user-specific citation preferences.
- **LaMP-2N (Personalized News Categorization):** A categorical text classification task that involves classifying news articles into one of 15 categories. Given an article written by a user, the model predicts its category using the user’s history of articles and their corresponding categories.
- **LaMP-2M (Personalized Movie Tagging):** A categorical text classification task that involves predicting one of 15 tags for a movie based on a user’s tagging history. This task evaluates the model’s ability to assign tags to a movie description using the user’s historical movie-tag pairs.
- **LaMP-3 (Personalized Product Rating):** An ordinal rating prediction task. The model predicts an integer rating from one to five for a given review based on the user’s historical review–rating pairs, and is evaluated with MAE and RMSE.
- **LaMP-4 (Personalized News Headline Generation):** A text generation task that involves generating personalized news headlines for given articles based on the author’s historical article-title pairs. This task evaluates the model’s ability to replicate the author’s stylistic nuances in headline generation.
- **LaMP-5 (Personalized Scholarly Title Generation):** A text generation task that involves generating titles for research articles based on the author’s historical article-title pairs, extending personalized text generation into scholarly domains.

We exclude the following two tasks from our evaluation:

- **LaMP-6 (Personalized Email Subject Generation):** A text generation task that involves generating subject lines for emails based on the user’s historical email-subject pairs. This task is excluded because the corresponding dataset is not publicly available.
- **LaMP-7 (Personalized Tweet Paraphrasing):** A text generation task that rewrites an input tweet in a user’s style. Its histories comprise unpaired user tweets, whereas the query requests a rewrite of a separate input. This format does not provide the task-consistent history input–output pairs assumed by our supervised Stage 2 setup, so a faithful extension requires a different training protocol. HYDRA (Zhuang et al., 2024) similarly excludes this task. We leave adaptation to LaMP-7 to future work.

LongLaMP would offer a useful test of substantially longer histories. However, its data protocol

| Task    | Stage 1 |        | Stage 2 |          | Eval. |       |
|---------|---------|--------|---------|----------|-------|-------|
|         | #User   | #Q     | #User   | #History | #User | #Q    |
| LaMP-1  | 6,442   | 7,919  | 100     | 318      | 100   | 123   |
| LaMP-2N | 274     | 3,662  | 49      | 220      | 49    | 6,033 |
| LaMP-2M | 829     | 3,181  | 100     | 56       | 100   | 3,302 |
| LaMP-3  | 19,899  | 22,388 | 101     | 960      | 101   | 112   |
| LaMP-4  | 1,543   | 7,275  | 100     | 270      | 100   | 6,275 |
| LaMP-5  | 14,581  | 16,075 | 101     | 443      | 101   | 107   |

Table 6: Dataset statistics. #User is the number of users, #Q is the number of queries, and #History is the average number of historical records per user for training.

and history construction must be aligned with our supervised Stage 2 setup to avoid an inequivalent comparison. We therefore report this as a limitation rather than retrofit an incompatible protocol, and plan to release the preprocessing and evaluation scripts to facilitate such extensions.

### A.3 Dataset Statistics

Table 6 summarizes the dataset statistics for each stage. Stage 1 uses all users except the most active ones for task adaptation, while Stage 2 trains on the selected most active users with their full interaction histories. The evaluation set shares the same users as Stage 2.

### A.4 Baseline Details

- **RAG (Retrieval-Augmented Generation) (Salemi et al., 2024b):** Following the retrieval-augmented personalization method presented in LaMP, the user’s query is augmented with the top- $k$  retrieved items from the user’s history corpus using BM25. We take  $k \in \{1, 2, 4\}$  to study the effect of retrieval size. No model parameters are modified.
- **PAG (Profile-Augmented Generation) (Richardson et al., 2023):** The user’s input sequence is concatenated with a natural language profile summarizing the user’s preferences and behavior patterns. Following the implementation of Tan et al. (2024b), user profiles are generated using the Vicuna-7B model. We evaluate PAG without retrieval ( $k=0$ ) and combined with retrieval augmentation ( $k=1$ ), following the setting of Richardson et al. (2023). No model parameters are modified.
- **OPPU (One PEFT Per User) (Tan et al., 2024b):** A dedicated LoRA adapter is trained for each user using their individual data. We

evaluate three variants: *base* (LoRA only, no prompt augmentation), *+rag* (LoRA combined with  $k=1$  retrieved records), and *+pag* (LoRA combined with user profile). LoRA is applied to the same target modules as our method with rank 8.

- **PER-PCS (Personalized Pieces) (Tan et al., 2024a):** A collaborative PEFT framework that selects representative “sharer” users via clustering, trains individual LoRA adapters and post-hoc gating vectors for each sharer, and assembles personalized PEFT parameters for target users by selecting and weighting sharer pieces in an auto-regressive, training-free manner. We evaluate three variants following OPPU: *base*, *+rag*, and *+pag*.
- **P2P (Profile-to-PEFT) (Tan et al., 2026):** A hypernetwork-based personalization framework that maps an encoded user profile directly to a full set of LoRA adapter parameters, eliminating per-user training at deployment time. The user profile is first generated by an upstream LLM summarizing the user’s history (following the PAG protocol), then encoded and fed to a hypernetwork that produces a user-specific LoRA. The generated LoRA is loaded into the backbone to produce personalized outputs. We use a single variant per task, since P2P already incorporates PAG (its input is a user profile) and produces a personalized LoRA in one step.
- **PROPER (Zhang et al., 2025):** A progressive PEFT framework that bridges population- and user-level adaptation through group-level LoRA experts, a user-aware router, and a LoRA-aware router. The released code depends on a different TRL/Transformers stack, so we implement PROPER in the same Transformers/PEFT pipeline used for all other methods while retaining its progressive adaptation design.

### A.5 Comparison with PROPER

Table 7 compares TAP-PER with PROPER under the same backbone, split, training environment, and evaluation protocol. TAP-PER performs better on all six tasks and all reported metrics. This comparison complements Table 1: PROPER reduces sparsity through group-level adaptation, whereas TAP-PER stores personalization in compact prefixes and uses a shared temporal history encoder.

| Task    | Metric            | PROPER | TAP-PER      |
|---------|-------------------|--------|--------------|
| LaMP-1  | Acc $\uparrow$    | 0.691  | <b>0.732</b> |
|         | F1 $\uparrow$     | 0.695  | <b>0.728</b> |
| LaMP-2N | Acc $\uparrow$    | 0.820  | <b>0.838</b> |
|         | F1 $\uparrow$     | 0.610  | <b>0.622</b> |
| LaMP-2M | Acc $\uparrow$    | 0.694  | <b>0.751</b> |
|         | F1 $\uparrow$     | 0.625  | <b>0.682</b> |
| LaMP-3  | MAE $\downarrow$  | 0.206  | <b>0.170</b> |
|         | RMSE $\downarrow$ | 0.496  | <b>0.430</b> |
| LaMP-4  | R-1 $\uparrow$    | 0.203  | <b>0.219</b> |
|         | R-L $\uparrow$    | 0.188  | <b>0.198</b> |
| LaMP-5  | R-1 $\uparrow$    | 0.489  | <b>0.538</b> |
|         | R-L $\uparrow$    | 0.440  | <b>0.485</b> |

Table 7: Comparison with PROPER on the six LaMP tasks.

| Task    | Metric         | Add          | Concat | Gated        | MLP          |
|---------|----------------|--------------|--------|--------------|--------------|
| LaMP-2M | Acc $\uparrow$ | 0.744        | 0.742  | 0.741        | <b>0.747</b> |
|         | F1 $\uparrow$  | 0.672        | 0.670  | 0.668        | <b>0.674</b> |
| LaMP-5  | R-1 $\uparrow$ | <b>0.529</b> | 0.522  | 0.525        | 0.518        |
|         | R-L $\uparrow$ | 0.472        | 0.464  | <b>0.476</b> | 0.469        |

Table 8: Comparison of prefix fusion strategies. These controlled runs use the same checkpoint and protocol for all four variants; their absolute values should be compared within this table.

## A.6 Prefix Fusion Strategies

We compare element-wise addition with three alternatives: concatenation followed by projection, a learned gated sum, and an MLP over the concatenated prefixes. Table 8 shows no consistently superior expressive fusion: the MLP is slightly better on LaMP-2M, while addition gives the best LaMP-5 ROUGE-1 and gated-sum gives the best ROUGE-L. We retain addition because it is parameter-free, stable, and competitive, keeping the contribution focused on prefix construction, temporal attention, and bridge integration.

## B Cold Start and Robustness

### B.1 Unseen-User Generalization

We conduct a strict user-level split on LaMP-2M and LaMP-5: Stage 2 is trained on 50% of users and evaluated on the remaining users, who are never seen during personalized training. Since an unseen user has no learned user-state prefix, we set  $\mathbf{P}_u = \mathbf{0}$  and use only the shared query-conditioned history prefix  $\mathbf{P}_q$ . Table 9 shows that this fallback outperforms Stage 1 and prompt-based baselines on both tasks. Thus, TAP-PER can serve a new user immediately when history is available, without first optimizing user-specific parameters.

| Task    | Metric         | Stage 1 | RAG   | PAG   | RAG+PAG | $\mathbf{P}_q$ only |
|---------|----------------|---------|-------|-------|---------|---------------------|
| LaMP-2M | Acc $\uparrow$ | 0.607   | 0.605 | 0.610 | 0.612   | <b>0.627</b>        |
|         | F1 $\uparrow$  | 0.581   | 0.580 | 0.585 | 0.590   | <b>0.595</b>        |
| LaMP-5  | R-1 $\uparrow$ | 0.528   | 0.522 | 0.514 | 0.518   | <b>0.575</b>        |
|         | R-L $\uparrow$ | 0.480   | 0.484 | 0.464 | 0.485   | <b>0.535</b>        |

Table 9: Strict user-level cold-start results. Stage 2 never observes the evaluation users;  $\mathbf{P}_u$  is zero and the last column uses only shared  $\mathbf{P}_q$ .

## B.2 Early Onboarding with Limited Histories

We further restrict the number of visible histories for Stage 2-unseen users, keep  $\mathbf{P}_u = \mathbf{0}$ , and perform no per-user update. On LaMP-2M, accuracy remains stable from 0.626 to 0.629 across the tested history budgets. On LaMP-5, ROUGE-L rises from 0.504 with the smallest visible-history budget to 0.540 as more records become available. Updating  $\mathbf{P}_u$  from only one or two examples can be unstable; these results support a staged deployment strategy: use shared  $\mathbf{P}_q$  during onboarding, then train or update  $\mathbf{P}_u$  asynchronously after sufficient reliable interactions accumulate. With no history at all, TAP-PER falls back to the Stage 1 task model.

## B.3 Noisy Streaming Feedback

To stress-test the controlled online protocol, we randomly corrupt 10–30% of streaming feedback targets on LaMP-2M and LaMP-5 while evaluating on the original clean test queries. We use the deployment-safe Online-Emb variant, updating only  $\mathbf{P}_u$  and freezing all shared components. At 30% corruption, LaMP-2M F1 remains approximately 0.63; LaMP-5 ROUGE-L changes from 0.476 with clean feedback to 0.468. TAP-PER is therefore reasonably stable under moderate synthetic label noise, but this test does not capture delayed, selection-biased, adversarial, or continuously drifting production feedback.

## C Additional Efficiency Analysis

### C.1 Generation-Task Update Time

Table 10 extends the per-user update-time analysis to LaMP-4 and LaMP-5. All measurements use the same A800 environment and report seconds per user. TAP-PER is about  $2.5\text{--}4\times$  faster than OPUU as history length grows. PER-PCS remains faster because it avoids per-user gradient updates, whereas full TAP-PER is more accurate in Table 1 and supports direct updates of a user’s compact state.

| Task   | Method  | $H=10$      | $H=50$      | $H=100$     | $H=150$     | $H=200$      |
|--------|---------|-------------|-------------|-------------|-------------|--------------|
| LaMP-4 | OPPU    | 4.19        | 19.00       | 37.65       | 54.87       | 76.15        |
|        | TAP-PER | 1.32        | 5.07        | 9.80        | 14.58       | 19.28        |
|        | PER-PCS | <b>0.57</b> | <b>1.41</b> | <b>2.55</b> | <b>3.57</b> | <b>4.62</b>  |
| LaMP-5 | OPPU    | 5.12        | 21.80       | 42.87       | 64.20       | 85.41        |
|        | TAP-PER | 2.08        | 7.68        | 14.36       | 20.98       | 27.85        |
|        | PER-PCS | <b>0.78</b> | <b>2.80</b> | <b>5.49</b> | <b>8.02</b> | <b>10.59</b> |

Table 10: Per-user update time (seconds) on generation tasks as the number of historical records  $H$  increases.

## C.2 Generalization to Alternative Backbone

To evaluate whether TAP-PER generalizes beyond the primary backbone used in our main experiments (Llama-3.1-8B), we conduct additional experiments using Qwen3-4B (Yang et al., 2025) as an alternative backbone. We compare TAP-PER against the same set of baselines on all six evaluated LaMP tasks. Results are shown in Table 11.

The trends largely mirror those on Llama-3.1-8B: TAP-PER (+ $P_u$ ) is the best on 10 out of 12 task-metric pairs. The exception is LaMP-5, where RAG with  $k=2$  achieves the best R-1 (0.512) and R-L (0.449), and OPPU also exceeds TAP-PER on both metrics. This suggests that the smaller backbone offers less headroom for prefix-based personalization on short-form scholarly title generation, where direct lexical cues from retrieved abstracts or per-user adaptation remain strong.

## C.3 Prompt Templates

Table 12 lists the prompt templates used for each task. For prompt-based methods (RAG, PAG), the user profile and/or retrieved records are prepended before the instruction. For TAP-PER, the prefix representations replace these textual augmentations.

## C.4 Case Study: Output Comparison

To complement the quantitative comparison, we provide qualitative output examples. Each table shows the query, the gold answer, and the outputs from a representative non-personalized baseline, prompt-based baselines (RAG and PAG), and two TAP-PER variants (without and with temporal biases). ✓ marks a correct answer and ✗ an incorrect one.

## C.5 Computation Details

All experiments are conducted on a single node equipped with 8 NVIDIA A800 80GB GPUs. Training is distributed across GPUs using HuggingFace Trainer with the accelerate backend, launched via torchrun for multi-GPU data parallelism. Stage 1 task adaptation takes approxi-

mately 5–30 minutes per task depending on dataset size, and Stage 2 personalized prefix learning takes approximately 10–30 minutes per task. All models are implemented using PyTorch, HuggingFace Transformers (Wolf et al., 2020), and PEFT.

## C.6 Use of Existing Artifacts

Our implementation, processed-data instructions, and reproduction commands are publicly available at <https://github.com/heng380/TAP-PER>. All third-party artifacts used in this work are publicly released for research use under their original terms: the LaMP benchmark (Salemi et al., 2024b) (MIT), Llama-3.1-8B (Grattafiori et al., 2024) (Llama 3.1 Community License), and Qwen3-4B (Yang et al., 2025) (Apache 2.0). We also consult the software/data supplements released with OPPU (Tan et al., 2024b) and PER-PCS (Tan et al., 2024a); P2P and PROPER are reproduced in our unified pipeline. Standard libraries (PyTorch, HuggingFace Transformers (Wolf et al., 2020), PEFT) are used under their respective licenses. We use each artifact strictly for its intended research purpose and do not redistribute third-party data or model weights.

## C.7 Potential Future Direction

Beyond improving personalization accuracy, future work can explore output diversity: a personalized model should not collapse to one narrow pattern when a user may have multiple plausible preferences, styles, or decision criteria. Recent work studies entropy preservation and diversity-aware objectives for language-model reasoning (Cui et al., 2025; Yao et al., 2025). Adapting such objectives to personalized generation may help preserve multiple high-quality responses while reducing over-personalization.

## C.8 Use of AI Assistants

We used AI assistants in a limited and clearly scoped manner during the preparation of this work. Specifically, AI assistants were used for: (i) code-

| Task                                         | Metric                                | Prompt-based   |                              |                              |                |                | Model-based    |                              |                              |                |                |                |                              |                   |                              |                              |
|----------------------------------------------|---------------------------------------|----------------|------------------------------|------------------------------|----------------|----------------|----------------|------------------------------|------------------------------|----------------|----------------|----------------|------------------------------|-------------------|------------------------------|------------------------------|
|                                              |                                       | RAG            |                              |                              | PAG            |                | OPPU           |                              |                              | PER-PCS        |                |                | P2P                          | TAP-PER           |                              |                              |
|                                              |                                       | k=1            | k=2                          | k=4                          | k=0            | k=1            | base           | +rag                         | +pag                         | base           | +rag           | +pag           | single                       | base <sup>†</sup> | +P <sub>q</sub> <sup>‡</sup> | +P <sub>u</sub> <sup>§</sup> |
| LaMP-1: PERSONALIZED CITATION IDENTIFICATION | Acc $\uparrow$<br>F1 $\uparrow$       | 0.618<br>0.613 | 0.618<br>0.616               | 0.618<br>0.605               | 0.642<br>0.641 | 0.585<br>0.559 | 0.642<br>0.634 | 0.626<br>0.625               | 0.683<br>0.682               | 0.585<br>0.576 | 0.602<br>0.598 | 0.675<br>0.674 | 0.690<br>0.685               | 0.585<br>0.566    | <u>0.724</u><br><u>0.720</u> | <b>0.813</b><br><b>0.813</b> |
| LaMP-2N: PERSONALIZED NEWS CATEGORIZE        | Acc $\uparrow$<br>F1 $\uparrow$       | 0.770<br>0.520 | 0.777<br>0.530               | 0.787<br>0.550               | 0.783<br>0.537 | 0.785<br>0.555 | 0.785<br>0.557 | 0.793<br>0.600               | <u>0.805</u><br><u>0.604</u> | 0.662<br>0.430 | 0.749<br>0.521 | 0.768<br>0.541 | 0.787<br>0.547               | 0.763<br>0.488    | 0.785<br>0.550               | <b>0.826</b><br><b>0.609</b> |
| LaMP-2M: PERSONALIZED MOVIE TAGGING          | Acc $\uparrow$<br>F1 $\uparrow$       | 0.520<br>0.475 | 0.536<br>0.486               | 0.539<br>0.482               | 0.565<br>0.520 | 0.550<br>0.499 | 0.427<br>0.358 | 0.468<br>0.362               | 0.464<br>0.368               | 0.491<br>0.402 | 0.534<br>0.432 | 0.538<br>0.434 | <u>0.683</u><br>0.525        | 0.545<br>0.505    | 0.580<br><u>0.527</u>        | <b>0.690</b><br><b>0.620</b> |
| LaMP-3: PERSONALIZED PRODUCT RATING          | MAE $\downarrow$<br>RMSE $\downarrow$ | 0.223<br>0.543 | 0.223<br>0.543               | <u>0.196</u><br><u>0.518</u> | 0.241<br>0.559 | 0.250<br>0.567 | 0.223<br>0.559 | <u>0.196</u><br><u>0.518</u> | 0.241<br>0.559               | 0.277<br>0.605 | 0.277<br>0.620 | 0.259<br>0.590 | 0.275<br>0.600               | 0.295<br>0.605    | 0.268<br>0.582               | <b>0.195</b><br><b>0.516</b> |
| LaMP-4: PERSONALIZED NEWS HEADLINE GEN.      | R-1 $\uparrow$<br>R-L $\uparrow$      | 0.176<br>0.157 | 0.175<br>0.156               | <u>0.184</u><br>0.164        | 0.174<br>0.154 | 0.179<br>0.159 | 0.163<br>0.146 | 0.168<br>0.152               | 0.165<br>0.148               | 0.148<br>0.134 | 0.157<br>0.140 | 0.156<br>0.140 | <u>0.184</u><br><u>0.166</u> | 0.174<br>0.154    | 0.175<br>0.154               | <b>0.189</b><br><b>0.168</b> |
| LaMP-5: PERSONALIZED SCHOLARLY TITLE GEN.    | R-1 $\uparrow$<br>R-L $\uparrow$      | 0.489<br>0.422 | <b>0.512</b><br><b>0.449</b> | <u>0.506</u><br><u>0.449</u> | 0.459<br>0.401 | 0.481<br>0.419 | 0.504<br>0.439 | 0.501<br>0.447               | 0.499<br>0.446               | 0.477<br>0.407 | 0.490<br>0.421 | 0.493<br>0.429 | 0.480<br>0.426               | 0.495<br>0.431    | 0.499<br>0.428               | 0.495<br>0.430               |

Table 11: Comparison results on LaMP benchmark with Qwen3-4B as backbone. Notation follows Table 1.

level autocompletion and refactoring suggestions during the implementation of training and evaluation scripts; (ii) editorial drafting and polishing of author-specified technical content; and (iii) help in formatting LaTeX tables and figures. All research ideas, experimental design, analysis, numerical results, claims, and references were supplied, executed, or independently verified by the authors. All AI-assisted text was reviewed and edited by the authors before inclusion.

| Task                                            | Prompt                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
|-------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| LaMP-1: Personalized Citation Identification    | {USER PREFIX + RECORD PREFIX} / ##USER PROFILE: {USER_PROFILE} /<br>##USER HISTORY: {USER_HISTORY}<br>##INSTRUCTION:<br>Identify the most relevant reference for the listed publication by the researcher. Select the reference paper that is most closely related to the researcher’s work.<br>paper title: {TITLE} reference: [1]: “{OPT1}” [2]: “{OPT2}” Answer:                                                                                                                              |
| LaMP-2M: Personalized Movie Tagging             | {USER PREFIX + RECORD PREFIX} / ##USER PROFILE: {USER_PROFILE} /<br>##USER HISTORY: {USER_HISTORY}<br>##INSTRUCTION:<br>Which tag does this movie relate to among the following tags? Just answer with the tag name without further explanation. tags: [sci-fi, based on a book, comedy, action, twist ending, dystopia, dark comedy, classic, psychology, fantasy, romance, thought-provoking, social commentary, violence, true story]<br>description: {QUERY} tag:                            |
| LaMP-2N: Personalized News Categorization       | {USER PREFIX + RECORD PREFIX} / ##USER PROFILE: {USER_PROFILE} /<br>##USER HISTORY: {USER_HISTORY}<br>##INSTRUCTION:<br>Which category does this article relate to among the following categories? Just answer with the category name without further explanation. categories: [travel, education, parents, style & beauty, entertainment, food & drink, science & technology, business, sports, healthy living, women, politics, crime, culture & arts, religion]<br>article: {QUERY} category: |
| LaMP-3: Personalized Product Rating             | {USER PREFIX + RECORD PREFIX} / ##USER PROFILE: {USER_PROFILE} /<br>##USER HISTORY: {USER_HISTORY}<br>##INSTRUCTION:<br>What is the score of the following review on a scale of 1 to 5? just answer with 1, 2, 3, 4, or 5 without further explanation. review: {QUERY} score:                                                                                                                                                                                                                    |
| LaMP-4: Personalized News Headline Generation   | {USER PREFIX + RECORD PREFIX} / ##USER PROFILE: {USER_PROFILE} /<br>##USER HISTORY: {USER_HISTORY}<br>##INSTRUCTION:<br>Generate a headline for the following article.<br>article: {QUERY} headline:                                                                                                                                                                                                                                                                                             |
| LaMP-5: Personalized Scholarly Title Generation | {USER PREFIX + RECORD PREFIX} / ##USER PROFILE: {USER_PROFILE} /<br>##USER HISTORY: {USER_HISTORY}<br>##INSTRUCTION:<br>Generate a title for the following abstract of a paper.<br>abstract: {QUERY} title:                                                                                                                                                                                                                                                                                      |

Table 12: Prompt templates for each personalization task. {USER PREFIX + RECORD PREFIX} denotes the combined prefix  $\mathbf{P}_{u,q} = \mathbf{P}_u + \mathbf{P}_q$ , prepended as soft tokens (present only in TAP-PER; absent for all other methods). {USER\_PROFILE} is a natural language summary of user preferences (present only in PAG variants). {USER\_HISTORY} contains the top- $k$  retrieved records (present only in RAG variants). Each placeholder and its surrounding section header are included only when the corresponding component is active; otherwise the entire block is omitted.

|                            |                                                                                                                                                                                                                                                                                                                                                                                                               |
|----------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Query (qid 81348)</b>   | As America recovers from the Civil War, one man tries to put the pieces of his life back together but finds himself fighting a new battle on the frontier. Cable is an embittered Confederate soldier who returns from the war to reclaim his Arizona homestead from rebel pioneers who sympathize with the Union war effort.                                                                                 |
| <b>Gold tag</b>            | based on a book                                                                                                                                                                                                                                                                                                                                                                                               |
| <b>RAG top-3 retrieved</b> | <i>based on a book</i> — Photographer Robert Kincaid wanders into the life of housewife Francesca Johnson for four days in the 1960s.<br><i>based on a book</i> — Eight people embark on an expedition into the Congo ... stumble across a race of killer apes.<br><i>based on a book</i> — A vampire relates his epic life story of love, betrayal, loneliness, and dark hunger to an over-curious reporter. |
| Non-Personalized           | violence ✗                                                                                                                                                                                                                                                                                                                                                                                                    |
| RAG                        | classic ✗                                                                                                                                                                                                                                                                                                                                                                                                     |
| PAG                        | classic ✗                                                                                                                                                                                                                                                                                                                                                                                                     |
| TAP-PER (no bias)          | based on a book ✓                                                                                                                                                                                                                                                                                                                                                                                             |
| TAP-PER                    | based on a book ✓                                                                                                                                                                                                                                                                                                                                                                                             |

(a) Case study from LaMP-2M.

|                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Abstract</b>            | Implementing security by design in practice often involves the application of threat modeling to elicit security threats and to aid designers in focusing efforts on the most stringent problems first. Existing threat modeling methodologies are capable of generating lots of threats, yet they lack even basic support to triage these threats, except for relying on the expertise and manual assessment by the threat modeler. . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| <b>Gold title</b>          | Risk-based design security analysis.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| <b>RAG top-3 retrieved</b> | <i>Knowledge-enriched security and privacy threat modeling.</i> — Creating secure and privacy-protecting systems entails the simultaneous coordination of development activities along three different yet mutually influencing dimensions: translating security and privacy goals to design choices, analyzing the design for threats, and performing a risk analysis of these threats.<br><i>SPARTA: Security &amp; Privacy Architecture Through Risk-Driven Threat Assessment.</i> — The development of secure and privacy-preserving software systems entails the continuous consideration of the security and privacy aspects of the system under development; the SPARTA prototype facilitates capturing security and privacy design decisions, continuous threat elicitation, and risk analysis.<br><i>Interaction-Based Privacy Threat Elicitation.</i> — Threat modeling involves the systematic identification, elicitation, and analysis of privacy- and/or security-related threats in the context of a specific system; this approach leads to better process guidance and more concrete interpretation of privacy threat types. |
| Non-Personalized           | Risk-Based Threat Modeling: A Case Study on SecureDrop                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| RAG                        | Knowledge-enriched security and privacy threat modeling.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| PAG                        | Risk-Based Threat Modeling: A Case Study on SecureDrop                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| TAP-PER (no bias)          | Risk-Based Threat Modeling: A Case Study on SecureDrop                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| TAP-PER                    | Risk-Based Security Analysis for Threat Modeling                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |

(b) Case study from LaMP-5.

Table 13: Qualitative output comparison. ✓ marks a correct answer and ✗ an incorrect one.