

Beyond My Own Shadow: Collaborative Latent Prompting for Personalized Question Answering

Yinghan Shi¹, Kai Fu¹, Xinyu Liu¹, Quanyou Chu¹,
Ming Du¹, Hongya Wang¹, Yanghua Xiao², Bo Xu^{1*}

¹School of Information and Intelligent Science, Donghua University

²Shanghai Key Laboratory of Data Science,

College of Computer Science and Artificial Intelligence, Fudan University

{2242829, 2242825, 2242850, 2242865}@mail.dhu.edu.cn

{duming, hywang}@dhu.edu.cn, shawyh@fudan.edu.cn, xubo@dhu.edu.cn

Abstract

While Large Language Models (LLMs) excel in general question answering, they typically generate uniform responses, ignoring users' diverse intents. Current Personalized Question Answering (PQA) methods face dual limitations: individual-based prompting struggles with data sparsity, while collaborative explicit prompting suffers from inaccurate user matching and attention dispersion caused by text concatenation. To tackle these challenges, we propose Collaborative Latent Prompting (CLP), a PQA framework that fuses collaborative knowledge within a dynamic latent space rather than relying on explicit textual concatenation. First, our Multi-Grained Collaborative Retrieval (MGCR) module ensures precise retrieval of relevant reference users through a three-stage refinement. Next, the Dynamic Q-Former (DQ-Former) utilizes question-modulated dynamic latent queries to extract compact personalized features from noisy texts. Finally, these features are projected into dense soft prompts to implicitly steer a frozen LLM. Experiments across three PQA domains and three backbone LLMs show consistent improvements over representative PQA baselines.

1 Introduction

While Large Language Models (LLMs) perform well in general question answering (Wang et al., 2025; Hu et al., 2025; Nian et al., 2025; Aftabi and Farzi, 2026), they usually follow a uniform paradigm, generating the same standard answers for the same questions. However, different users often have different intents based on their own backgrounds, even when they ask the exact same question (Xu et al., 2025; Kasela et al., 2024). For instance, given the query “How should I prepare for

an interview?”, a fresh graduate may need basic interview skills, while an experienced engineer may look for team management strategies. Therefore, understanding these personal intents to generate customized answers—formally known as Personalized Question Answering (PQA)—has become an important research topic.

Current PQA methods generally fall into two paradigms. The first, Individual-based Prompting (Figure 1(a)), augments the prompt with the target user's historical records. However, this strategy struggles in data-sparse scenarios or new topics where relevant personal records are limited (Thonet et al., 2025; Zhang et al., 2025c). To overcome this, recent work has explored the second paradigm, Collaborative Explicit Prompting (Figure 1(b)) (Shi et al., 2025; Sun et al., 2025), which aggregates the historical records of similar users. Although effective in mitigating data sparsity, explicit textual aggregation introduces two key challenges:

Firstly, identifying truly intent-relevant users remains challenging. Existing methods primarily match users based on topic relevance or interest overlap (Qiu et al., 2025b,a), failing to capture the fine-grained intents of the current query. For instance, a student seeking study strategies and a teacher looking for teaching methods may both have rich records in the “English Education” topic, but their actual intents are fundamentally different. Retrieving the teacher's history for the student may introduce irrelevant or even misleading personalization signals. This suggests that collaborative personalization should move beyond broad topic overlap and perform query-aware, fine-grained intent alignment.

Secondly, explicit context fusion struggles to fully capture personalized signals. Existing methods simply concatenate multi-source texts, but due to context-window constraints, only a limited subset can be included, while personalized signals are sparsely distributed across long histories. More-

*Corresponding author. The work is supported by the Donghua University 2026 Cultivation Project of Discipline Innovation xkcx-202622.

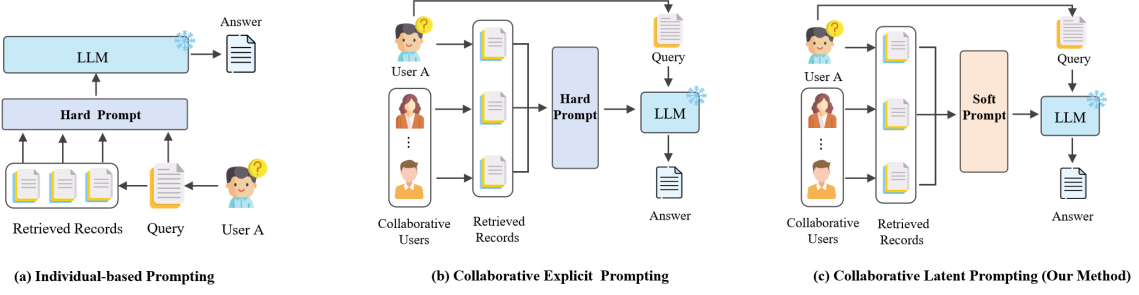


Figure 1: Comparison of PQA paradigms. (a) Individual-based Prompting relies on isolated target user histories. (b) Collaborative Explicit Prompting concatenates extensive group texts, causing severe context dilution and noise. (c) Our method introduces Collaborative Latent Prompting, dynamically compressing multi-source histories into dense soft prompts for robust, adaptive generation.

over, this text-level fusion inevitably mixes personalized and generic information, forcing the model to filter noise during generation. Consequently, due to attention dispersion, the model struggles to fully capture sparse but useful personalized signals (Yang et al., 2025b; Zhang et al., 2025a). To address this, we separate information filtering from answer generation by compressing histories into latent personalized features that guide the LLM more efficiently.

To tackle these dual challenges, we propose Collaborative Latent Prompting (CLP), a novel PQA framework that first retrieves relevant users across multiple granularities and then extracts multi-perspective personalized features within a latent space. First, we design a **Multi-Grained Collaborative Retrieval (MGCR) module**. Through a progressive three-stage refinement (topic-level, macro-level, and micro-level), MGCR aims to retrieve intent-relevant reference users more accurately. To efficiently capture sparse personalized signals from both target and collaborative user histories, we introduce the **Dynamic Q-Former (DQ-Former)**. Inspired by vision-language architectures (Li et al., 2023), DQ-Former uses dynamic learnable latent queries to selectively aggregate these signals through a multi-head cross-attention layer. The resulting latent features are then projected by an MLP into dense soft prompts, which implicitly steer the frozen LLM toward the user’s specific intent. This implicit steering mechanism mitigates context-window limitations and attention dispersion caused by explicit text concatenation.

The main contributions of this paper are summarized as follows:

- We propose CLP, a novel PQA framework that combines multi-grained collaborative retrieval

with latent prompting to alleviate individual data sparsity and attention dispersion.

- We introduce DQ-Former, a dynamic latent-query module that converts target and collaborative user histories into compact soft prompts, reducing reliance on explicit text concatenation.
- Extensive experiments across three PQA domains and three backbone LLMs show that CLP consistently and significantly outperforms strong PQA baselines.

2 Related Work

2.1 Personalized Question Answering

Personalized Question Answering (PQA) aims to tailor Large Language Model (LLM) responses to individualized intents (Xu et al., 2025; Liu et al., 2025b; Zhang et al., 2026). Early methods primarily relied on Parameter-Efficient Fine-Tuning (PEFT) (Thonet et al., 2025; Lester et al., 2021) or individual retrieval paradigms, explicitly augmenting contexts with past interactions (Shu et al., 2026; Rasmussen et al., 2025). While effective for short-form tasks (Qiu et al., 2025b; Zhang et al., 2025b; Salemi et al., 2024), recent research tackles complex, long-form PQA (Kumar et al., 2024), utilizing strategies like PoT (Salemi et al., 2026) for multi-directional prompting. However, these methods remain confined to sparse individual histories and inefficient textual aggregation, causing context dispersion. To address this, our work shifts toward collaborative latent prompting, dynamically integrating multi-source knowledge without overloading the LLM’s context window.

2.2 Collaborative Prompting

Collaborative Prompting mitigates individual data sparsity by drawing insights from similar users.

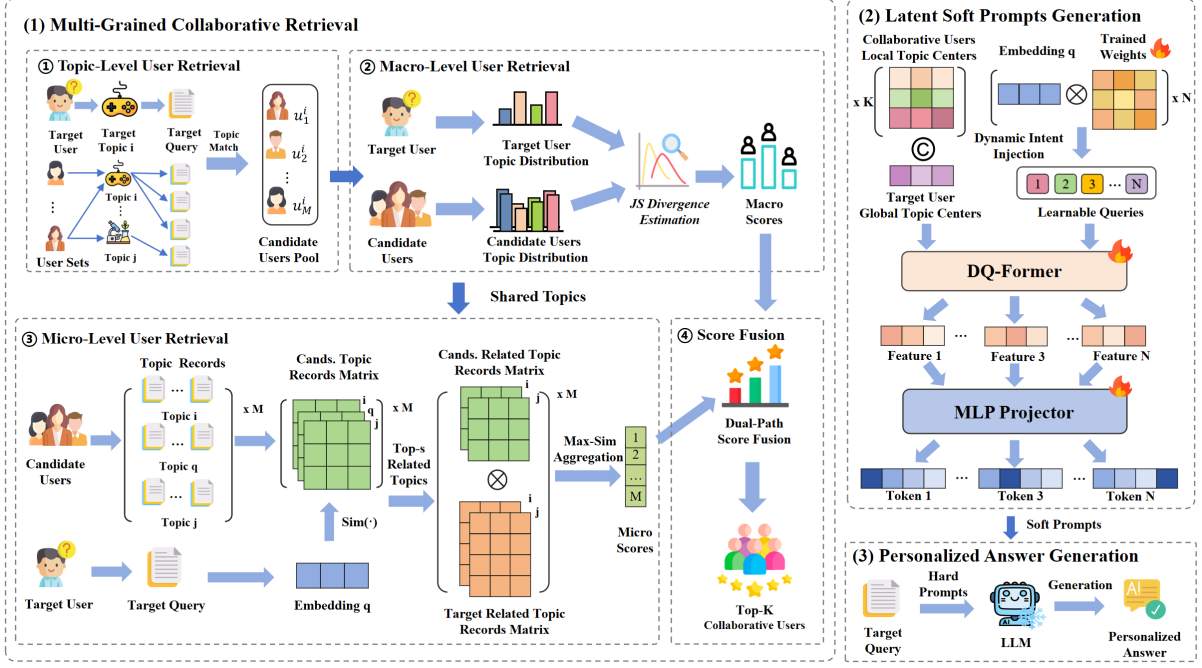


Figure 2: The overall architecture of our proposed CLP framework. It comprises three core modules: (1) the Multi-Grained Collaborative Retrieval; (2) the DQ-Former for dynamically compressing multi-source histories into dense personalized soft prompts; and (3) the generative stage utilizing a frozen LLM conditioned on the injected latent features.

While foundational in recommender systems (He et al., 2017), this paradigm increasingly enhances LLMs for personalized generation (Wu et al., 2024; Qiu et al., 2025b; Shi et al., 2025; Tan et al., 2024; Zhuang et al., 2024). However, explicitly concatenating retrieved collaborative histories restricts context capacity.

To overcome this bottleneck, latent space modeling offers an efficient alternative. Recent latent compression methods—such as ICAE (Ge et al., 2023) and Q-Former (Li et al., 2023)—effectively distill compact representations from redundant inputs. This latent approach has inspired recent explorations into user-specific plugins and latent difference modeling (Liu et al., 2025a; Qiu et al., 2025a; Zhang et al., 2025a). Building on these advancements, our work combines collaborative prompting with latent modeling, utilizing dynamic soft prompts to inject distilled group knowledge and steer the LLM’s generation.

3 Overview

3.1 Problem Formulation

We consider a personalized question-answering setting where each user u is associated with a historical profile $\mathcal{P}_u = \{p_i\}_{i=1}^{|\mathcal{P}_u|}$. Each record is defined

as $p_i = (q_i, c_i, t_i, a_i)$, where q_i , c_i , t_i , and a_i denote the previous question title, detailed description, topic, and key aspects, respectively. The aspects a_i are used only for evaluation and are never exposed during generation. We encode each record with the aspect-free text $\tilde{p}_i = [q_i; c_i]$, while using t_i only for retrieval-stage topic filtering and grouping.

Given a new query q_u from the target user u , the objective is to generate a personalized response $\hat{y}_u$ by leveraging the target user’s profile $\mathcal{P}_u$ and relevant collaborative profiles $\mathcal{P}_v$:

$$\hat{y}_u = \pi_\theta(q_u, \mathcal{P}_u, \mathcal{P}_v),$$

where π_θ denotes the backbone LLM.

3.2 Framework

The overall architecture of our proposed framework is illustrated in Figure 2, consisting of three core components: (1) Multi-Grained Collaborative Retrieval (MGCR); (2) Latent Soft Prompts Generation; and (3) Personalized Answer Generation.

The overall process is as follows: First, given a target user and a current query, MGCR retrieves Top- K relevant collaborative users from a candidate pool. Subsequently, the target user’s global topic records and the Top- K collaborative users’ local topic records are aggregated and fed into the

Soft Prompts Generation module. Within this module, DQ-Former conducts multi-perspective feature extraction via cross-attention, modulated by the current query intent. These features are then projected into soft prompts via an MLP projector. The trainable modules are optimized jointly through supervised fine-tuning. Finally, the extracted soft prompts are combined with the target query and fed into a frozen LLM to generate the final personalized response.

4 Method

4.1 Multi-Grained Collaborative Retrieval

The Multi-Grained Collaborative Retrieval is designed to efficiently locate the most relevant collaborative users from a large-scale candidate pool while ensuring semantic and interest relevance.

4.1.1 Topic-Level User Retrieval

The topic-level user retrieval module is used to perform a strict pre-filtering to avoid topic-irrelevant noise and reduce computational overhead. Prior to fine-grained retrieval, it ensures that a candidate user v has historical records matching the topic t_q of the target query q_u . This hard constraint enforces topic relevance while efficiently pruning the search space. Formally, given a global candidate pool $\mathcal{V}_{pool}$, a candidate $v \in \mathcal{V}_{pool}$ is retained only if their profile $\mathcal{P}_v$ satisfies $t_q \in \{t_i \mid p_i \in \mathcal{P}_v\}$, yielding the valid candidate set $\mathcal{C}_{valid}$.

4.1.2 Macro-Level User Retrieval

Following the pre-filtering, the macro-level user retrieval module is employed to evaluate the global interest alignment between the target user and the valid candidates. This stage stems from the observation that users sharing similar historical topic distributions tend to exhibit overlapping long-term preferences, making their historical behaviors more likely to provide reliable collaborative insights for the target user. To quantify this alignment, we calculate the topic probability distribution for each user. Let T denote the set of all available topics, and $\mathcal{P}_{u,t}$ be the subset of records belonging to topic t . The probability of topic t for user u is calculated as:

$$\mathbf{P}_u(t) = \frac{|\mathcal{P}_{u,t}|}{|\mathcal{P}_u|} \quad (1)$$

Let $\mathbf{P}_u = [P_u(t_1), \dots, P_u(t_{|T|})]$ and $\mathbf{P}_v$ denote the topic probability distribution vectors across all $|T|$ topics for the target user u and the candidate v , respectively. We evaluate their discrepancy using

Jensen-Shannon Divergence (JSD). By defining the mean distribution $\mathbf{M}$, the JSD is computed via the Kullback-Leibler (D_{KL}) divergence:

$$\mathbf{M} = \frac{1}{2}(\mathbf{P}_u + \mathbf{P}_v) \quad (2)$$

$$\begin{aligned} \text{JSD}(\mathbf{P}_u \parallel \mathbf{P}_v) &= \frac{1}{2}D_{KL}(\mathbf{P}_u \parallel \mathbf{M}) \\ &+ \frac{1}{2}D_{KL}(\mathbf{P}_v \parallel \mathbf{M}) \end{aligned} \quad (3)$$

The global similarity between target user u and candidate $v \in \mathcal{C}_{valid}$ is then formally measured as:

$$S_{macro}(u, v) = 1 - \text{JSD}(\mathbf{P}_u \parallel \mathbf{P}_v) \quad (4)$$

We retain the top 50% of candidates based on S_{macro} to proceed to the fine-grained micro-matching stage.

4.1.3 Micro-Level User Retrieval

The micro-level user retrieval module further refines candidate users by measuring fine-grained record-level alignment within query-relevant topics. While macro-level retrieval captures global interest similarity, users with similar topic distributions may still have different specific needs. Thus, this stage performs topic-constrained max-similarity matching between the target user's historical records and each candidate user's records. A candidate user receives a higher score when their records within query-relevant topics are semantically closer to the target user's records in the same topics, indicating more reliable collaborative evidence for personalized generation.

To formalize this matching, we first map the query and all aspect-free historical records into a dense continuous space using a pre-trained encoder f_{emb} :

$$\begin{aligned} \mathbf{E}_{q_u} &= f_{emb}(q_u) \in \mathbb{R}^{d_{encode}}, \\ \mathbf{E}_{\tilde{p}_i} &= f_{emb}(\tilde{p}_i) \in \mathbb{R}^{d_{encode}}. \end{aligned} \quad (5)$$

We then select the top- s shared topics $\mathcal{T}_s$ that are most relevant to the current query q_u . For each shared topic t , its relevance is computed by averaging the similarities between the query embedding and the candidate user's records under that topic:

$$S_{topic}(t) = \frac{1}{|\mathcal{P}_{v,t}|} \sum_{\tilde{p}_k \in \mathcal{P}_{v,t}} \text{Sim}(\mathbf{E}_{q_u}, \mathbf{E}_{\tilde{p}_k}). \quad (6)$$

Within each selected topic $t \in \mathcal{T}_s$, we perform topic-constrained max-similarity matching. For each target record $\tilde{p}_i \in \mathcal{P}_{u,t}$, we search the candidate user's records $\mathcal{P}_{v,t}$ for the most semantically

similar record. The matching score for topic t is calculated as:

$$S_{match}(t) = \frac{1}{|\mathcal{P}_{u,t}|} \sum_{\tilde{p}_i \in \mathcal{P}_{u,t}} \max_{\tilde{p}_j \in \mathcal{P}_{v,t}} \text{Sim}(\mathbf{E}_{\tilde{p}_i}, \mathbf{E}_{\tilde{p}_j}). \quad (7)$$

where the maximum operation captures the nearest-neighbor alignment for each target record within the same topic.

Finally, the micro-level similarity score S_{micro} is obtained by averaging these matching scores across the selected topics:

$$S_{micro}(u, v) = \frac{1}{|\mathcal{T}_s|} \sum_{t \in \mathcal{T}_s} S_{match}(t). \quad (8)$$

4.1.4 Hybrid Score Fusion

After obtaining the macro-level interest similarity and micro-level record alignment, we combine them to produce the final retrieval ranking. The final score is computed as follows:

$$S_{final}(u, v) = \alpha S_{macro}(u, v) + (1 - \alpha) S_{micro}(u, v) \quad (9)$$

4.2 Latent Soft Prompts Generation

4.2.1 DQ-Former: Intent-Guided Dynamic Feature Extraction

Given the Top- K most similar collaborative users $\mathcal{V}_u$ from the MGCR module, we collect their local records under the query topic t_q , while considering the target user u 's global preferences across all M topics. To represent these multi-source behaviors efficiently without excessive text-level tokens, topic center vectors are computed via mean pooling. Specifically, for a record set $\mathcal{P}_t$ under topic t , its center vector $\mathbf{c}_t$ is defined as:

$$\mathbf{c}_t = \frac{1}{|\mathcal{P}_t|} \sum_{p_i \in \mathcal{P}_t} \mathbf{E}_{\tilde{p}_i} \quad (10)$$

Then, the target user's global topic centers and the K collaborative users' local centers are concatenated to form a unified collective history matrix $\mathbf{X} \in \mathbb{R}^{(M+K) \times d_{encode}}$:

$$\mathbf{X} = [\mathbf{c}_{u,t_1}, \dots, \mathbf{c}_{u,t_M}, \mathbf{c}_{v_1,t_q}, \dots, \mathbf{c}_{v_K,t_q}] \quad (11)$$

To compress and refine this collective history matrix $\mathbf{X}$, we propose the Dynamic Query-Former (**DQ-Former**) architecture. Unlike traditional architectures that rely on static learnable queries—which often suffer from representation

collapse—DQ-Former incorporates a **Dynamic Intent Injection** mechanism to dynamically adapt to the current query context.

Specifically, we initialize a set of N learnable latent queries $\mathbf{Q}_{latent} \in \mathbb{R}^{N \times d_{encode}}$ to capture personalized features from multiple perspectives. To make these query slots adaptive to the current input, we inject the current query embedding $\mathbf{E}_{q_u}$ into each latent query:

$$\mathbf{Q}_{dy} = \text{LayerNorm}(\mathbf{Q}_{latent} + \mathbf{E}_{q_u} \mathbf{W}_{inj}), \quad (12)$$

where $\mathbf{W}_{inj} \in \mathbb{R}^{d_{encode} \times d_{encode}}$ is a trainable projection matrix and the projected query embedding is broadcast to all latent queries.

The dynamic queries then attend to the collective history matrix $\mathbf{X}$ through a multi-head cross-attention layer. For clarity, we present the attention computation in its single-head form:

$$\mathbf{A} = \text{Softmax} \left(\frac{(\mathbf{Q}_{dy} \mathbf{W}_Q)(\mathbf{X} \mathbf{W}_K)^T}{\sqrt{d_{encode}}} \right) (\mathbf{X} \mathbf{W}_V), \quad (13)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are trainable projection matrices. In implementation, this operation is extended to an 8-head multi-head attention layer. The final extracted latent features are obtained with a residual connection and layer normalization:

$$\mathbf{H} = \text{LayerNorm}(\mathbf{Q}_{dy} + \mathbf{A}). \quad (14)$$

4.2.2 MLP Projection for Semantic Alignment

The MLP projection module is designed to map the extracted latent features into the continuous embedding space of the backbone language model. By aligning this representation with the LLM's internal understanding, the model can seamlessly interpret the injected features as context-aware soft prompts. Specifically, we project the dimensionality of $\mathbf{H}$ to match d_{model} using a two-layer MLP:

$$\mathbf{S} = \text{MLP}(\mathbf{H}) = \text{GELU}(\mathbf{H} \mathbf{W}_1 + \mathbf{b}_1) \mathbf{W}_2 + \mathbf{b}_2 \quad (15)$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_{encode} \times d_{model}}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_{model} \times d_{model}}$. The resulting soft prompts $\mathbf{S} \in \mathbb{R}^{N \times d_{model}}$ are then prepended to the original textual input sequence, guiding the frozen LLM implicitly.

4.2.3 Optimization and Auxiliary Losses

The optimization process is used to update the parameters of the DQ-Former and the MLP projector through supervised fine-tuning. We freeze the base

generative model and distill pseudo-labels y_{pseudo} from a teacher model using only the question and its description, ensuring no leakage of the evaluation aspects. To ensure accuracy, feature diversity, and training stability, the overall objective $\mathcal{L}_{total}$ integrates a primary generation loss $\mathcal{L}_{gen}$ with two crucial auxiliary constraints: $\mathcal{L}_{orth}$ and $\mathcal{L}_{InfoGain}$.

Primary Generation Loss ($\mathcal{L}_{gen}$): This is the standard auto-regressive cross-entropy loss used to guide the generated soft prompts toward the pseudo-labels:

$$\mathcal{L}_{gen} = - \sum \log \pi_{\theta}(y_{pseudo} \mid [\mathbf{S}; I_{task}; q_u]) \quad (16)$$

Orthogonal Penalty Loss ($\mathcal{L}_{orth}$). We apply an orthogonal penalty to the cross-attention outputs H to encourage different latent slots to capture non-redundant personalized signals. Let $\tilde{H}$ be the row-wise L2-normalized matrix of H :

$$\mathcal{L}_{orth} = \|\tilde{H}\tilde{H}^T - \mathbf{I}\|_F \quad (17)$$

Information Gain Loss ($\mathcal{L}_{InfoGain}$): To prevent the MLP projector from memorizing noise or generating static vectors, we introduce a contrastive margin loss across three projections: $\bar{P}_{fusion}$ (target and collaborative users), $\bar{P}_{self}$ (target only), and $\bar{P}_{noise}$ (target and random users). Let $\mathbf{e}_{q_u}^{LLM}$ denote the mean-pooled LLM input embeddings of the current query. This objective makes $\bar{P}_{fusion}$ more query-relevant than $\bar{P}_{self}$ and $\bar{P}_{noise}$, promoting collaborative gains while suppressing irrelevant signals:

$$\begin{aligned} \mathcal{L}_{InfoGain} = & \max(0, \text{Sim}(\mathbf{e}_{q_u}^{LLM}, \bar{\mathbf{P}}_{self}) \\ & - \text{Sim}(\mathbf{e}_{q_u}^{LLM}, \bar{\mathbf{P}}_{fusion}) + m_1) \\ & + \max(0, \text{Sim}(\mathbf{e}_{q_u}^{LLM}, \bar{\mathbf{P}}_{noise}) \\ & - \text{Sim}(\mathbf{e}_{q_u}^{LLM}, \bar{\mathbf{P}}_{fusion}) + m_2). \end{aligned} \quad (18)$$

The final optimization objective is a linearly weighted combination of the aforementioned losses, where λ_1 and λ_2 are tuning hyperparameters:

$$\mathcal{L}_{total} = \mathcal{L}_{gen} + \lambda_1 \mathcal{L}_{orth} + \lambda_2 \mathcal{L}_{InfoGain} \quad (19)$$

4.3 Personalized Answer Generation

The final stage directly integrates the extracted soft prompts $\mathbf{S}$, the explicit task instructions I_{task} , and the target query q_u to guide the frozen LLM. By seamlessly aligning with the LLM’s input space, the soft prompts $\mathbf{S}$ serve as personalized control

signals that actively intervene in the model’s internal prompting trajectory, implicitly steering the prompting process toward the user’s specific intent.

The final personalized response $\hat{y}_u$ is:

$$\hat{y}_u = \pi_{\theta}([\mathbf{S}; I_{task}; q_u]) \quad (20)$$

5 Experiments

5.1 Datasets

We evaluate on LaMP-QA (Salemi and Zamani, 2025), a recent and, to the best of our knowledge, the only benchmark specifically designed for personalized long-form question answering. It covers three domains: Arts & Entertainment, Lifestyle & Personal Development, and Society & Culture. To support the collaborative retrieval paradigm, the original training set is partitioned into two disjoint subsets: 800 instances are randomly sampled for model training, while the remaining samples are utilized to construct a global candidate pool.

5.2 Evaluation Metrics

Following the standard LaMP-QA protocol, we employ Qwen-2.5-32B-Instruct¹ as an automated judge to assess how well the generated responses cover the hidden personalized aspects a_i (Section 3.1). The final performance metric is calculated as the average coverage score across all aspects. For detailed scoring rubrics and prompts, we refer readers to the original papers (Salemi and Zamani, 2025).

5.3 Baselines

To evaluate the effectiveness of our approach, we compare CLP against representative baselines: (1) **Zero-Shot LLM:** The base LLM generates responses using its internal parametric knowledge without access to any external user history. (Salemi and Zamani, 2025) (2) **Standard RAG:** A conventional retrieval-augmented generation pipeline that provides the LLM with the top- k most relevant historical records from the user’s local profile. (Salemi and Zamani, 2025) (3) **Long-Context:** This approach utilizes the full capacity of the LLM by concatenating the user’s history in descending order of semantic relevance to the current query, filling the context window up to its maximum limit. (Yang et al., 2025a) (4) **Plan-RAG:** A two-stage

¹<https://huggingface.co/Qwen/Qwen2.5-32B-Instruct>

baseline that employs a planning module to explicitly draft personalized aspects, which subsequently guide the final generation process. (Salemi and Zamani, 2025) (5) **CEP-RAG**: The method uses the same collaborative users retrieved by MGCR as CLP, but injects their histories through explicit textual concatenation rather than latent soft prompts. (Shi et al., 2025) (6) **Prompt Tuning**: A parameter-efficient baseline that optimizes continuous soft prompts via distillation from a teacher model. (Li and Liang, 2021) (7) **PPlug**: A plug-and-play personalized framework that steers generation through record-level weight fusion over the user’s historical interactions. (Liu et al., 2025a)

5.4 Implementation Details

Model Configuration. We employ Qwen2.5-7B-Instruct², Gemma-2-9B-it³ and Llama-3.1-8B-Instruct⁴ as backbones. For DQ-Former, we set the number of latent queries to $N = 16$ and the collaborative group size to $K = 5$. For MGCR, we set $s = 3$ in micro-level retrieval and $\alpha = 0.5$ in hybrid score fusion. We utilize Qwen3-Max as the teacher model to generate pseudo-gold labels. Models are optimized via AdamW with a learning rate of $2e-5$, a batch size of 2, and 4 gradient accumulation steps. The auxiliary loss parameters ($m_1 = 0.15$, $m_2 = 0.3$, $\lambda_1 = 0.05$, and $\lambda_2 = 0.15$) are determined via a coarse grid search.

Baseline Setup. All baselines are evaluated under a unified experimental setting with the same backbone LLMs, decoding configuration, and Contriever (Izacard et al., 2021) text encoder whenever applicable. Implementation details and reproducibility configurations are provided in Appendix B.

Hardware Configurations. All experiments are conducted on a server equipped with two NVIDIA RTX A6000 GPUs (approximately 48 GB of VRAM each).

5.5 Main Results

Table 1 reports the results on the LaMP-QA test sets across three domains and three backbone LLMs. CLP achieves the best average score on all three backbones, improving over the strongest reproduced baseline by 0.0464, 0.0771, and 0.0396

²<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

³<https://huggingface.co/google/gemma-2-9b-it>

⁴<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

points on Qwen-2.5, Gemma-2, and Llama-3.1, respectively. This indicates that CLP is not tied to a specific backbone model, but provides stable gains across different LLM families and personalized QA domains.

Compared with PPlug, which assigns query-dependent weights to target-user profiles and maps them into the LLM semantic space, CLP further introduces relevant collaborative histories through MGCR. The improvements over PPlug show that properly selected peer histories provide complementary personalization signals beyond isolated user records. Meanwhile, Prompt Tuning trains the same number of continuous prompt vectors using identical teacher-distilled pseudo-labels, but these vectors are static and do not contain external user knowledge. This comparison suggests that the gain of CLP does not simply come from adding trainable soft prompts or teacher-distilled knowledge, but stems intrinsically from the joint effect of collaborative knowledge and latent feature extraction.

Retrieval-based baselines such as CEP-RAG directly place retrieved histories into the textual prompt. Although this strategy may provide useful personal information, it also introduces irrelevant or weakly related content, which can distract the LLM. In contrast, CLP uses DQ-Former to extract compact latent soft prompts from target and collaborative histories. This design separates personalized feature extraction from answer generation, allowing the frozen LLM to use distilled signals without processing long, noisy histories. The consistent gains over Long-Context and CEP-RAG further support the effectiveness of latent collaborative prompting.

To further validate automatic evaluation, we provide an alternative-evaluator robustness check and human evaluation in Appendices A.5 and A.6. Human judgments confirm this trend: CLP achieves a 52.0% win rate against a 22.0% lose rate, with 87.3% agreement with the automatic evaluator.

5.6 Ablation Study

To examine the contribution of each component, we conduct ablation studies on the Qwen2.5 backbone by systematically removing MGCR retrieval stages, DQ-Former, auxiliary objectives, and different input sources. The results are shown in Table 2. **MGCR Ablation.** Discarding any retrieval stage leads to performance drops, confirming that the three levels provide complementary signals. Specifically, omitting Topic-Level Retrieval causes the

Method	Arts & Ent.	Life. & Dev.	Soc. & Cul.	Avg.
Qwen-2.5-7B-Instruct				
Zero-Shot LLM	0.3164	0.4629	0.5574	0.4456
Standard RAG	0.3360	0.4464	0.5061	0.4295
Long-Context	<u>0.3873</u>	<u>0.5212</u>	<u>0.5375</u>	<u>0.4820</u>
Plan-RAG	0.3638	0.4812	0.5321	0.4590
CEP-RAG	0.3753	0.5233	0.5673	<u>0.4886</u>
Prompt Tuning	0.3201	<u>0.5282</u>	<u>0.5427</u>	<u>0.4637</u>
PPlug	0.3377	0.5196	<u>0.5842</u>	0.4805
CLP(Ours)	0.3985*	0.5924*	0.6142*	0.5350*
Gemma-2-9B-Instruct				
Zero-Shot LLM	0.1863	0.4194	0.4995	0.3684
Standard RAG	0.2681	0.4241	0.4769	0.3897
Long-Context	0.2810	0.4510	0.4862	0.4061
Plan-RAG	0.3424	0.4627	<u>0.5642</u>	0.4564
CEP-RAG	0.2899	0.4564	0.5123	0.4195
Prompt Tuning	0.2485	0.4610	0.4788	0.3961
PPlug	<u>0.3475</u>	<u>0.5523</u>	0.4995	0.4664
CLP(Ours)	0.3982*	0.6009*	0.6315*	0.5435*
Llama-3.1-8B-Instruct				
Zero-Shot LLM	0.2852	0.4574	0.4981	0.4136
Standard RAG	0.3152	0.4099	0.4894	0.4048
Long-Context	0.2978	0.4486	0.5149	0.4204
Plan-RAG	<u>0.3402</u>	0.4997	0.5473	0.4624
CEP-RAG	0.3302	0.4849	0.5301	0.4484
Prompt Tuning	0.3125	0.4519	0.5057	0.4234
PPlug	0.3261	<u>0.5293</u>	<u>0.5769</u>	<u>0.4774</u>
CLP(Ours)	0.3709*	0.5658*	0.6143*	0.5170*

Table 1: Overall performance comparison on the LaMP-QA benchmark across three domains and three backbone LLMs. The best results under each backbone are highlighted in bold, and the best baseline results are underlined. * indicates statistically significant improvement over the strongest baseline under the same setting with a one-sided paired t-test ($p < 0.05$).

largest decline (0.5350 $\rightarrow$ 0.5090), indicating that coarse topic filtering is crucial for pruning the irrelevant user search space. Meanwhile, removing Micro-Level Retrieval notably hurts performance (0.5350 $\rightarrow$ 0.5173), indicating that fine-grained record alignment is essential for capturing query-specific intents. The drop without Macro-Level Retrieval (0.5350 $\rightarrow$ 0.5225) further proves that global preference similarity serves as a valuable prior.

Latent Module Ablation. Removing DQ-Former and using only a Vanilla MLP reduces the average score from 0.5350 to 0.4940, indicating that a simple projection module is insufficient for extracting useful signals from multi-source histories. We further remove the dynamic intent injection mechanism and use static learnable queries instead, which decreases the average score from 0.5350 to 0.5162. This result shows that modulating latent queries with the current query intent helps DQ-Former extract more query-relevant personalized features, rather than relying on fixed and query-agnostic latent prompts.

Objective Ablation. Removing either auxiliary objective weakens the model, confirming the impor-

Method	Arts & Ent.	Life & Dev.	Soc. & Cul.	Avg.
CLP(Ours)	0.3985	0.5924	0.6142	0.5350
<i>MGCR Ablation</i>				
w/o Topic-Level Retrieval	0.3756	0.5547	0.5967	0.5090
w/o Macro-Level Retrieval	0.3845	0.5788	0.6043	0.5225
w/o Micro-Level Retrieval	0.3786	0.5793	0.5939	0.5173
<i>Latent Module Ablation</i>				
w/o DQ-Former (Vanilla MLP)	0.3420	0.5668	0.5731	0.4940
w/o Dynamic Intent Injection	0.3692	0.5812	0.5981	0.5162
<i>Objective Ablation</i>				
w/o $\mathcal{L}_{orth}$	0.3591	0.5755	0.5944	0.5097
w/o $\mathcal{L}_{InfoGain}$	0.3201	0.5704	0.5771	0.4892
w/o Auxiliary Losses	0.3244	0.5689	0.5091	0.4675
<i>Input Source Ablation</i>				
Target Only	0.3645	0.5660	0.5799	0.5035
Collaborative Only	0.3517	0.5689	0.5877	0.5028
Target + Random Users	0.3584	0.5607	0.5921	0.5037

Table 2: Ablation study of CLP on the Qwen2.5 backbone. We ablate MGCR retrieval stages, DQ-Former, auxiliary losses, and input sources to analyze the contribution of each component.

tance of both training constraints. Both auxiliary objectives improve over the generation-only objective, with $\mathcal{L}_{InfoGain}$ contributing a larger gain than $\mathcal{L}_{orth}$. Their combination achieves the best performance, suggesting that feature diversity and contrastive signal preservation provide complementary benefits for stabilizing soft-prompt optimization.

Input Source Ablation. Relying solely on target or collaborative histories yields similar but sub-optimal performance (0.5035 and 0.5028, respectively), indicating that they offer complementary signals that are both indispensable. Notably, the “Target + Random Users” setting performs comparably to Target Only; this suggests that DQ-Former is relatively robust to irrelevant collaborative inputs, which reduces the impact of irrelevant noise under suboptimal retrieval conditions. However, it remains far below our full CLP method, suggesting that the performance gains mainly come from relevant collaborative knowledge rather than merely expanding the context.

6 Conclusion

In this paper, we proposed CLP for personalized question answering. CLP combines MGCR for multi-grained collaborative user retrieval with DQ-Former for query-aware latent soft prompt generation. Experiments across three domains and three LLM backbones show that CLP consistently improves personalized response quality over strong baselines. The results suggest that collaborative knowledge is more effective when retrieved with intent alignment and injected through compact latent prompts rather than explicit text concatenation.

Limitations

Although CLP consistently improves personalized question answering performance, it still has several limitations.

First, CLP does not explicitly model the temporal evolution of user preferences. The current framework retrieves and represents user histories mainly based on topic and semantic relevance, without directly distinguishing recent records from older ones. In real-world scenarios, however, user interests and information needs may gradually change over time. Incorporating time-aware user modeling could further improve the framework's ability to capture evolving preferences.

Second, CLP adopts a fixed number of latent soft prompt tokens for all inputs. We set the prompt length to $N = 16$ according to the hyperparameter analysis, which achieves the best overall performance in our experiments. However, different users and queries may require different levels of personalized evidence. Future work can explore adaptive prompt-length selection, allowing the model to allocate latent tokens dynamically according to query complexity and the amount of useful profile information.

References

- Seyyede Zahra Aftabi and Saeed Farzi. 2026. [Enhancing reliability in community question answering with an expert-oriented RAG system](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2551–2569, Rabat, Morocco. Association for Computational Linguistics.
- Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2023. In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945*.
- Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182.
- Zhanghao Hu, Hanqi Yan, Qinglin Zhu, Zhenyi Shen, Yulan He, and Lin Gui. 2025. Beyond prompting: An efficient embedding framework for open-domain question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19975–19990.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.
- Pranav Kasela, Marco Braga, Gabriella Pasi, and Raffaele Perego. 2024. [Se-pqa: Personalized community question answering](#). In *Companion Proceedings of the ACM Web Conference 2024, WWW '24*, page 1095–1098. ACM.
- Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. 2024. [Longlamp: A benchmark for personalized long-form text generation](#). *Preprint*, arXiv:2407.11016.
- 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Xiang Lisa Li and Percy Liang. 2021. [Prefix-tuning: Optimizing continuous prompts for generation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.
- Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025a. Llm+ persona-plug= personalized llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9373–9385.
- Langming Liu, Shilei Liu, Yujin Yuan, Yizhen Zhang, Bencheng Yan, Zhiyuan Zeng, Zihao Wang, Jiaqi Liu, Di Wang, Wenbo Su, Pengjie Wang, Jian Xu, and Bo Zheng. 2025b. [Uqabench: Evaluating user embedding for prompting llms in personalized question answering](#). In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD '25*, page 5652–5661, New York, NY, USA. Association for Computing Machinery.
- Jinming Nian, Zhiyuan Peng, Qifan Wang, and Yi Fang. 2025. W-rag: Weakly supervised dense retrieval in rag for open-domain question answering. In *Proceedings of the 2025 International ACM SIGIR conference on innovative concepts and theories in information retrieval (ICTIR)*, pages 136–146.
- Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. 2025a. Latent inter-user

- difference modeling for llm personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10610–10628.
- Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. 2025b. Measuring what makes you unique: Difference-aware user modeling for enhancing llm personalization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21258–21277.
- Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*.
- Alireza Salemi, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Zhuowan Li, Spurthi Amba Hombaiah, Weize Kong, Tao Chen, Hamed Zamani, and Michael Bendersky. 2026. *Pathways of thoughts: Multi-directional thinking for long-form personalized question answering*. In *Proceedings of the ACM Web Conference 2026, WWW '26*, page 5822–5833, New York, NY, USA. Association for Computing Machinery.
- Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. *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, Bangkok, Thailand. Association for Computational Linguistics.
- Alireza Salemi and Hamed Zamani. 2025. *LaMP-QA: A benchmark for personalized long-form question answering*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1159, Suzhou, China. Association for Computational Linguistics.
- Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Han Li. 2025. Retrieval augmented generation with collaborative filtering for personalized text generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1294–1304.
- Yiheng Shu, Saisri Padmaja Jonnalagedda, Xiang Gao, Bernal Jiménez Gutiérrez, Weijian Qi, Kamalika Das, Huan Sun, and Yu Su. 2026. *REMem: Reasoning with episodic memory in language agent*. In *The Fourteenth International Conference on Learning Representations*.
- Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2025. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 281–296.
- Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. 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.
- Thibaut Thonet, Germán Kruszewski, Jos Rozen, Pierre Erbacher, and Marc Dymetman. 2025. Fast: Feature-aware sampling and tuning for personalized preference alignment with limited data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9352–9381.
- Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. 2025. *Chain-of-retrieval augmented generation*. In *Advances in Neural Information Processing Systems*, volume 38, pages 59888–59915. Curran Associates, Inc.
- 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.
- Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. 2025. *Personalized generation in large model era: A survey*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24607–24649, Vienna, Austria. Association for Computational Linguistics.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025a. *Qwen2.5 technical report*. *Preprint*, arXiv:2412.15115.
- Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Yang Wang, and Liangming Pan. 2025b. How is llm reasoning distracted by irrelevant context? an analysis using a controlled benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13340–13358.
- Jinghao Zhang, Yuting Liu, Wenjie Wang, Qiang Liu, Shu Wu, Liang Wang, and Tat-Seng Chua. 2025a. Personalized text generation with contrastive activation steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7128–7141.
- Kepu Zhang, Teng Shi, Weijie Yu, and Jun Xu. 2025b. Prlm: Learning explicit reasoning for personalized rag via contrastive reward optimization. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5484–5488.

- Linhai Zhang, Jialong Wu, Deyu Zhou, and Yulan He. 2025c. 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.
- Yingyi Zhang, Pengyue Jia, Derong Xu, Yi Wen, Xianneng Li, Yichao Wang, Wenlin Zhang, Xiaopeng Li, Weinan Gan, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2026. Personalize before retrieve: Llm-based personalized query expansion for user-centric retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 16406–16414.
- Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. [Hydra: Model factorization framework for black-box llm personalization](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 100783–100815. Curran Associates, Inc.

A Datasets and Evaluation

To reliably evaluate the generated responses, we use the aspect-based evaluation method from the LaMP-QA benchmark (Salemi and Zamani, 2025). Since using Large Language Models (LLMs) for automatic evaluation has known challenges, this section provides experimental evidence to support our method. We explain the reliability of the extracted aspects, the validity of the scoring method, the rationale for using a 32B-parameter evaluator, and additional robustness checks with alternative and human evaluations.

A.1 Dataset Details

In this section, we provide a detailed overview of the LaMP-QA benchmark utilized in our experiments. To comprehensively evaluate the personalization capabilities of our proposed framework, the dataset is constructed based on the SE-PQA dataset sourced from the community-based StackExchange platform. The benchmark encompasses three major domains: Arts & Entertainment, Lifestyle & Personal Development, and Society & Culture.

Following rigorous filtering processes to ensure that all retained questions strictly require personalization and contain sufficient narrative details, we reorganize the data splits to fit our collaborative retrieval setting. Specifically, we randomly sample 800 queries from the original training set to form our actual training set. The remaining training instances are utilized to construct a vast candidate collaborative user pool, which is shared across both the training and testing phases for user retrieval. The candidate pool is constructed only from the training split and excludes all validation/test queries, answers, and extracted aspects. During evaluation, the current test query and its aspects are never included in either the target profile or the collaborative pool. The detailed statistics regarding the original number of questions across different domains and splits are summarized in Table 3.

A.2 Reliability of the Extracted Aspects

Our evaluation extracts user-specific information needs from the question descriptions to form personalized aspects. The LaMP-QA benchmark shows that these automatically extracted aspects highly match the users’ real needs, getting an average human score of 4.9 out of 5. Also, the agreement between different human reviewers is very high (Cohen’s kappa is 0.87). Since these aspects

Domain	Train	Val.	Test	Total
Arts & Ent.	9,349	801	767	10,917
Life. & Dev.	7,370	892	989	9,251
Soc. & Cul.	7,614	810	1,074	9,498
Overall	24,333	2,503	2,830	29,666

Table 3: Statistics of the dataset utilized in our experiments. The original Train set is further split in our setup: 800 samples are used for actual model training, while the rest form the shared collaborative user pool.

come directly from the users’ own words, they are naturally personalized and provide a solid foundation for our evaluation.

A.3 Validity of Aspect-Based Evaluation

We choose the aspect-based evaluation rather than direct scoring or pairwise comparison (comparing two responses side-by-side). Experimental results show that the aspect-based method has the highest agreement with human judgments. It matches human preferences in 73% of cases, which is much better than the direct scoring method (58%). Furthermore, we do not use pairwise evaluation because it has a serious position bias. Studies show that simply changing the order of the two responses can change the LLM’s choice in 78% of cases. This high sensitivity makes pairwise evaluation unstable. Instead, the aspect-based method provides a detailed assessment of how well the response meets the user’s specific needs.

A.4 Rationale for the Evaluator Model Size

In our experiments, we use an instruction-tuned 32B-parameter model (e.g., Qwen 2.5-32B) as the evaluator. This choice is supported by a clear trend: as the model size increases, its agreement with human judgment improves. For example, a 0.5B model only has a 48% agreement with humans, while the 32B model achieves 73%. Smaller models cannot easily tell the difference in response quality and often act like simple binary classifiers. This leads to unfairly high scores and poor human agreement. Therefore, larger LLMs are much better at capturing detailed quality differences that match human expectations.

A.5 Robustness to Alternative Evaluator

To further examine whether the observed improvements depend on a single automated evaluator,

Method	Arts & Ent.	Life. & Dev.	Soc. & Cul.	Avg.
Zero-Shot LLM	0.1372	0.3277	0.4515	0.3055
Standard RAG	0.1795	0.3401	0.4156	0.3117
Long-Context	0.2110	0.3329	0.4662	0.3367
Plan-RAG	0.2595	0.3368	0.4742	0.3568
CEP-RAG [†]	0.2114	0.3864	0.5123	0.3700
Prompt Tuning	0.1837	0.3759	0.4387	0.3328
PPlug	0.2828	0.5413	0.4745	0.4329
CLP (Ours)	0.3618	0.5835	0.6404	0.5286

Table 4: Robustness evaluation on Gemma-2-9B-Instruct outputs using Llama-3.1-70B-Instruct as an alternative evaluator. [†] denotes the strengthened explicit collaborative RAG baseline using the same MGCR-retrieved users as CLP.

we re-evaluate all outputs generated by Gemma-2-9B-Instruct using Llama-3.1-70B-Instruct as an alternative judge. The results are reported in Table 4. Although the absolute scores change under this stronger evaluator, CLP still achieves the best performance across all three domains and obtains the highest average score. Specifically, CLP reaches an average score of 0.5286, outperforming the strongest baseline PPlug by 0.0957 points and the explicit collaborative RAG baseline by 0.1586 points.

Compared with the main evaluation using Qwen-2.5-32B-Instruct, the score of CLP remains relatively stable, changing from 0.5435 to 0.5286. This suggests that the advantage of CLP is not an artifact of a specific evaluator, but remains consistent when judged by a different large instruction-tuned model. The results further support the effectiveness of collaborative latent prompting for improving personalized response quality.

A.6 Human Evaluation on Preference and Alignment

To rigorously validate our automatic evaluation metrics, we conduct a human preference evaluation.

Setup and Annotation. We randomly sample 150 queries (50 per domain) generated by Qwen-2.5-7B-Instruct. Our CLP is paired against the strongest baseline in each domain: Long-Context for *Arts & Entertainment*, Prompt Tuning for *Life & Development*, and PPlug for *Society & Culture*. Two independent annotators perform a blind A/B test, choosing which response better captures the user’s intent and provides a more helpful answer (**Win**, **Tie**, or **Lose** for CLP). A third expert resolves any disagreements. The inter-annotator agreement is substantial (Cohen’s $\kappa = 0.768$).

Results and Alignment. As shown in Table 5, CLP consistently outperforms the strongest baselines, achieving an overall Win rate of 52.0% against a 22.0% Lose rate. To compute Human-LLM agreement, we convert the automatic aspect-based scores into pairwise preference labels with a threshold $\tau = 0.05$: CLP is preferred if its score exceeds the baseline by more than τ , the baseline is preferred if the score is lower by more than τ , and the case is treated as a tie otherwise. The Human-LLM agreement rate reaches 87.3%, showing that our automatic evaluation is largely consistent with human judgments.

Domain	Strongest Baseline	Human Preference (%)			Human-LLM Agreement
		Win	Tie	Lose	
Arts & Ent.	Long-Context	42.0 (21)	32.0 (16)	26.0 (13)	86.0%
Life. & Dev.	Prompt Tuning	62.0 (31)	20.0 (10)	18.0 (9)	88.0%
Soc. & Cul.	PPlug	52.0 (26)	26.0 (13)	22.0 (11)	88.0%
Overall	–	52.0 (78)	26.0 (39)	22.0 (33)	87.3%

Table 5: Human evaluation results on 150 sampled queries (50 per domain) generated by Qwen-2.5-7B-Instruct. CLP is compared against the strongest domain-specific baseline. “Win” indicates human annotators preferred CLP. Values in parentheses denote the absolute number of samples.

B Baseline Implementation Details

To ensure a consistent comparative setup, all baseline models were evaluated within a unified experimental framework utilizing the exact same backbone LLMs. Furthermore, all text encoding and retrieval processes across applicable baselines and our method uniformly adopt the Contriever encoder backbone (Izacard et al., 2021). To ensure a fair evaluation of generation quality and consistency across different paradigms, we strictly control the generation configurations for all models, setting the decoding temperature to 0.7 and capping the maximum generation length at 2048 tokens.

Before detailing the specific configurations, we elaborate on the architectural distinctions summarized in Table 6. We evaluate the baselines across four critical capabilities: (1) **Global History Utilization**: integrating a user’s entire historical trajectory to capture a global persona, rather than relying on truncated or locally retrieved subsets; (2) **Latent Representation**: compressing verbose contexts into dense embeddings (soft prompts) to mitigate attention dispersion; (3) **Collaborative Knowledge**: dynamically retrieving and leveraging peer intelligence to overcome individual data sparsity; and (4) **PEFT Required**: utilizing trainable modules

or parameter-efficient fine-tuning (e.g., LoRA or projection layers) during any stage of the framework. As illustrated in Table 6, our CLP framework addresses the trade-off of "partial retrieval versus context overflow" by effectively unifying all four paradigms.

To systematically introduce the baselines, we categorize them into two major technical paradigms based on whether the final answer generation directly uses explicit textual contexts or implicit latent features.

B.1 Explicit Text-based Methods

These baselines rely on constructing explicit textual prompts to feed into the generative LLM. This category includes both frozen, training-free methods and methods that require separate fine-tuning for text-based planning.

Zero-Shot LLM The base LLM acts purely on its internal parametric knowledge. The input strictly consists of the standard task instructions and the target query q , formulated as $I = [Instruction; q]$. It has zero access to external user profiles, serving as the lower bound for personalization.

Standard RAG Following standard retrieval-augmented generation (Salemi and Zamani, 2025), we retrieve the top- K ($K = 10$) most relevant historical records from the target user’s local profile $\mathcal{P}_u$. These records are explicitly concatenated into the context window: $I = [Instruction; \mathcal{P}_{u, top-10}; q]$. While it provides personal context, it is affected by data sparsity.

Long-Context Instead of utilizing a strict chronological order, this approach prioritizes semantic relevance by concatenating the user’s historical interactions in descending order of similarity to the current query. This retrieval-priority concatenation continues until the model’s maximum context limit is reached. Specifically, the context window is truncated at 32,768 (32K) tokens for Qwen2.5-7B-Instruct, and 8,192 (8K) tokens for both Gemma-2-9B-Instruct and Llama-3.1-8B-Instruct. Although this method maximizes the inclusion of historical information, it may suffer from attention dilution and the “lost-in-the-middle” effect.

Strengthened CEP-RAG To isolate the contribution of latent prompting from the effect of collaborative user retrieval, we construct a strengthened explicit collaborative baseline. Specifically, we

use the same Top-5 collaborative users retrieved by MGCR as in CLP, but replace the latent soft prompt generation module with explicit text concatenation. To control the input length, the context is constructed by concatenating 5 personal records and 5 collaborative records, with one record selected from each retrieved peer. This setting provides a stronger comparison than using a weaker collaborative retriever, because both CLP and this baseline access the same retrieved users, and the key difference lies in how collaborative information is injected into the LLM.

Plan-RAG Utilizing the official implementation (Salemi and Zamani, 2025), this is a two-stage pipeline that requires parameter-efficient fine-tuning. In the first stage, a separate language model is fine-tuned via LoRA to act as a planner, explicitly predicting discrete personalized aspects based on $\mathcal{P}_u$. In the second stage, these predicted text aspects are injected into the final prompt to guide the frozen backbone LLM. Compared with CLP, Plan-RAG requires maintaining a separate fine-tuned text generator, increases inference latency, and can be affected by cascading hallucination errors if the planner predicts incorrect aspects.

B.2 Implicit Latent-based Methods

These methods introduce lightweight trainable parameters (such as soft prompts or specialized plugins) to implicitly steer the LLM through continuous embedding spaces, keeping the backbone weights frozen.

Prompt Tuning We optimize a set of continuous soft prompt vectors $S \in \mathbb{R}^{N \times d_{model}}$ utilizing the same pseudo-labels distilled from the teacher model. For fairness, the soft prompt length is set to $N = 16$, identical to our CLP. However, unlike our DQ-Former, these soft prompts are static and cannot dynamically adapt to external user histories.

PPlug Since the official source code for PPlug is not publicly available, we strictly re-implemented the model adhering to the architectural details specified in the original paper (Liu et al., 2025a). Following its core design, PPlug assigns query-dependent weights to all historical interactions of the target user and maps them into a single plug-and-play user embedding. Notably, this method requires fine-tuning the input text encoder, a projection MLP, and a set of trainable instruction embeddings for task optimization, while keeping the backbone LLM

Table 6: Comparison of our proposed CLP framework and representative baselines across four key architectural capabilities. Notably, CLP is the only approach that effectively unifies global history utilization, collaborative knowledge, and latent representation. **PEFT** denotes Parameter-Efficient Fine-Tuning (i.e., optimizing lightweight modules while keeping the backbone LLM frozen).

Methods ($\downarrow$)	Global History	Latent Representation	Collaborative Knowledge	PEFT Required
Zero-Shot LLM	✗	✗	✗	✗
Standard RAG	✗	✗	✗	✗
Long-Context	✗	✗	✗	✗
Plan-RAG	✗	✗	✗	✓
CEP-RAG	✗	✗	✓	✗
Prompt Tuning	✗	✓	✗	✓
PPlug	✓	✓	✗	✓
CLP(Ours)	✓	✓	✓	✓

frozen. Although PPlug leverages global history and latent representation, it is limited to the target user’s own history. Therefore, it does not directly address data sparsity when individual records are insufficient or when the user encounters a new topic, suggesting the potential value of collaborative knowledge.

C Hyperparameter Sensitivity Analysis

We evaluate the impact of the latent soft prompt length ($N \in \{8, 16, 32, 48\}$) and the collaborative user group size ($K \in \{5, 10, 20\}$) using the Qwen-2.5 backbone (Figure 3).

Impact of Soft Prompt Length (N): Performance exhibits a distinct inverted-U shape. A small $N = 8$ creates an information bottleneck, struggling to compress dense multi-source histories. Conversely, excessive tokens ($N \geq 32$) introduce feature redundancy and distract the LLM’s attention.

Impact of Collaborative Group Size (K): A highly-filtered, compact collaborative group ($K = 5$) consistently yields superior and stable performance. Increasing K universally degrades accuracy, as retrieving too many users inevitably introduces long-tail semantic noise, counteracting the benefits of collaborative prompting.

Overall, the results demonstrate that $N = 16$ and $K = 5$ achieve the optimal balance between information capacity and noise filtering.

D Case Study

To intuitively illustrate how our CLP framework leverages implicit collaborative knowledge to guide response generation, we present a mechanical QA

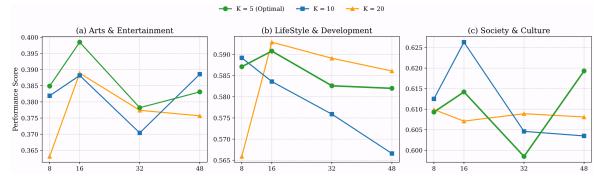


Figure 3: Hyperparameter sensitivity analysis of the latent soft prompt length (N) and the collaborative group size (K) on the Qwen-2.5 backbone.

case in Table 7. The user inquires about upgrading a specific bicycle hub component (*Shimano FH-M756*).

The Standard RAG baseline fails to grasp the technical specificity of the query. Relying solely on sparse individual records, it retrieves irrelevant personal history and erroneously advises the user to “*buy a completely new wheelset.*”

Conversely, the collaborative explicit prompting baseline (CEP-RAG) suffers from severe attention dispersion. Diluted by the explicit concatenation of multi-source texts, it fails to accurately isolate the user’s true intent. Instead of addressing the mechanical problem, it generates irrelevant text regarding the user’s background and ultimately provides a generic cop-out suggestion to “*consult a professional mechanic.*”

In contrast, our CLP framework successfully resolves the query through effective noise filtering and latent feature extraction. Rather than forcing the LLM to process verbose and noisy contexts, CLP utilizes the DQ-Former to distill collaborative behavioral patterns into context-aware soft prompts. These latent signals act as a precise intent compass, steering the LLM to bypass generic advice and directly activate its relevant internal parametric

Target Query: “Shimano FH-M756 freehub body compatibility - can I upgrade?”

Rubric Aspects: (1) Compatibility of freehub bodies with FH-M756-A hub (2) Avoiding a full wheel rebuild

CLP (Ours)	CEP-RAG	Standard-RAG
“Yes, you can upgrade the Shimano FH-M756... Here is a detailed mechanical breakdown: The FH-M756 utilizes a specific 10mm fixing bolt ... Upgrading to newer models or other brands requires specific modifications, so verifying exact compatibility is essential... This approach ensures compatibility while effectively avoiding a full wheel rebuild .”	“ Based on the target user’s history and similar peers’ history , it seems like they have a strong interest in bike maintenance... [... generates irrelevant text about user background ...] ... Before upgrading, ensure it is compatible with your current setup. If you’re unsure, it’s recommended to consult a professional bike mechanic .”	“Based on your previous questions... Given your situation with the Shimano 6-speed freewheel [... hallucinates from irrelevant personal history ...], I would recommend considering upgrading to a completely new wheelset instead of replacing the freehub body...If you decide to upgrade, consider purchasing a new chainring to ensure proper shifting.”
Intent Steering: Accurately captures intent and delivers precise solutions.	Attention Dispersion: Fails to capture intent and gives generic answer.	Data Sparsity: Hallucinates irrelevant context and provides incorrect advice.

Table 7: Qualitative comparison of different personalized paradigms. Our CLP method successfully captures implicit technical intents (highlighted in green), whereas baselines suffer from attention dispersion and data sparsity (hallucinations/failures highlighted in red).

knowledge. Guided by this clear intent, the model accurately identifies the hub’s physical constraints (*e.g.*, *10mm fixing bolt*, *cup-and-cone bearings*). This case clearly demonstrates that latent collaborative prompting provides superior intent alignment and context management compared to explicit text concatenation.

E Prompt Templates

To ensure reproducibility, we provide the prompt templates used for response generation in our experiments. All methods follow the same output requirement: generating a clear and personalized answer to the user’s current question. The main difference lies in the form of personalization signals available to each method, including explicit textual histories for retrieval-based methods and continuous soft prompts for latent-based methods. No prompt exposes the hidden evaluation aspects during generation.

Prompt Template: Long-Context

You are a helpful assistant designed to generate personalized responses to user questions. Your task is to answer a user’s question from a post in a personalized way by considering this user’s past post questions and detailed descriptions of these questions.

Your input:

- The user’s current question from a post.
- The user’s past post questions and detailed descriptions of these questions to learn about the user’s preferences.

Your output:

You should generate a clear and personalized text answer to the user’s current question.

Figure 4: Prompt template for the Long-Context baseline.

Prompt Template: CEP-RAG

You are a helpful assistant designed to generate personalized responses to user questions. Your task is to answer a user’s question from a post in a personalized way by considering the target user’s past post questions and detailed descriptions, as well as the historical posts of similar users as a collaborative reference.

Your input:

- The target user’s current question from a post.
- The target user’s past post questions and detailed descriptions of these questions.
- The past post questions and detailed descriptions from similar users.

Your output: You should generate a clear and personalized text answer to the user’s current question.

Figure 5: Prompt template for the CEP-RAG baseline.

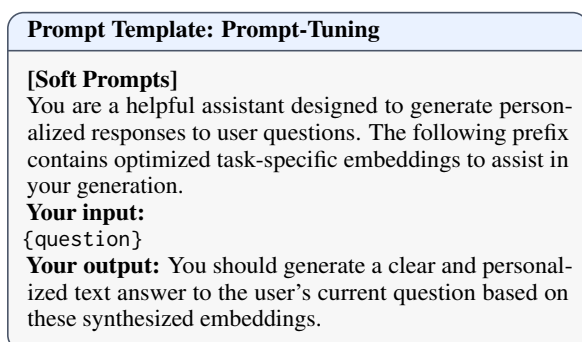


Figure 6: Prompt template for the Prompt-Tuning baseline.

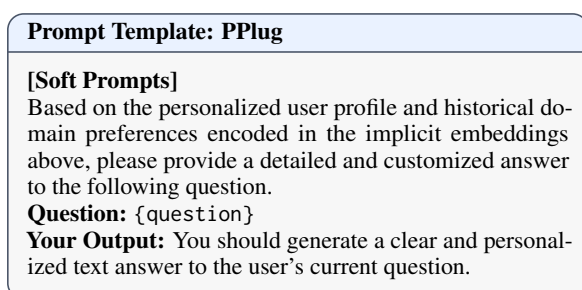


Figure 7: Prompt template for the PPlug baseline.

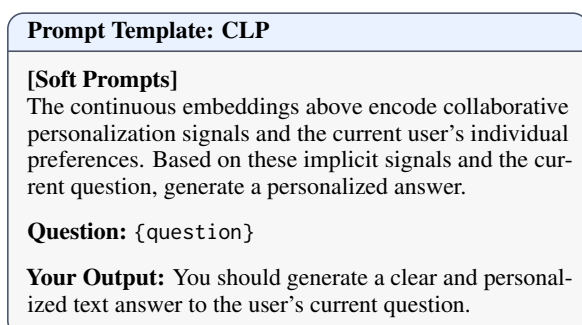


Figure 8: Prompt template for our CLP framework.