Personalize Before Retrieve: LLM-based Personalized Query Expansion for User-Centric Retrieval

Yingyi Zhang[1,2,*], Pengyue Jia[2,*], Derong Xu[2,3], Yi Wen[2], Xianneng Li[1,†], Yichao Wang[4,†], Wenlin Zhang[2], Xiaopeng Li[2], Weinan Gan[4], Huifeng Guo[4], Yong Liu[4], Xiangyu Zhao[2,†]

[1]School of Economics and Management, Dalian University of Technology,
[2]Department of Data Science, City University of Hong Kong,
[3]School of Artificial Intelligence and Data Science, University of Science and Technology of China,
[4]Huawei Technologies Ltd.
xianneng@dlut.edu.cn, wangyichao5@huawei.com, xianzhao@cityu.edu.hk

Abstract

Retrieval-Augmented Generation (RAG) critically depends on effective query expansion to retrieve relevant information. However, existing expansion methods adopt uniform strategies that overlook user-specific semantics, ignoring individual expression styles, preferences, and historical context. In practice, identical queries in text can express vastly different intentions across users. This representational rigidity limits the ability of current RAG systems to generalize effectively in personalized settings. Specifically, we identify two core challenges for personalization: 1) user expression styles are inherently diverse, making it difficult for standard expansions to preserve personalized intent. 2) user corpora induce heterogeneous semantic structures—varying in topical focus and lexical organization—which hinders the effective anchoring of expanded queries within the user's corpora space. To address these challenges, we propose *Personalize Before Retrieve* (**PBR**), a framework that incorporates user-specific signals into query expansion prior to retrieval. PBR consists of two components: **P-PRF**, which generates stylistically aligned pseudo feedback using user history for simulating user expression style, and **P-Anchor**, which performs graph-based structure alignment over user corpora to capture its structure. Together, they produce personalized query representations tailored for retrieval. Experiments on two personalized benchmarks show that PBR consistently outperforms strong baselines, with up to 10% gains on PersonaBench across retrievers. Our findings demonstrate the value of modeling personalization *before* retrieval to close the semantic gap in user-adaptive RAG systems.

Code — https://github.com/Applied-Machine-Learning-Lab/PBR-code

Extended version — https://arxiv.org/abs/2510.08935

1 Introduction

Retrieval-Augmented Generation (RAG) has emerged as a pivotal paradigm for enhancing the capabilities of large language models (LLMs) (Lewis et al. 2020; Zhao et al. 2024; Zhang et al. 2025c), leveraging a two-stage process: first retrieving relevant external corpus, then generating responses conditioned on the retrieved information (Cheng et al. 2023). The effectiveness of this paradigm hinges on the quality of query expansion by using LLM world knowledge (Culpepper et al. 2021; Song and Zheng 2024), which directly impacts the accuracy and relevance of the retrieval stage. Most current strategies generate pseudo answers (Gao et al. 2023) or candidate completions (Jia et al. 2023) using LLMs in a zero-shot or few-shot manner. These generated semantic representations, often combined with the original query, are used to retrieve information from a global external corpus.

Recently, with LLMs widely adopted as personalized assistants, many user queries are required to retrieve from individual histories and contexts, demanding retrieval mechanisms that go beyond generic corpora to leverage user-specific data. Existing query expansion (QE) strategies re-

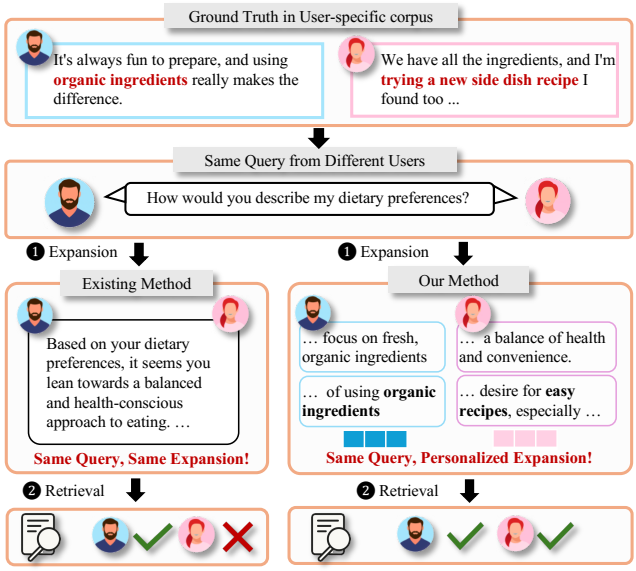


Figure 1: An example comparing generic and personalized query expansion.

*Equal contribution.
†Corresponding author.

main limited in their ability to capture user-specific semantics. In practice, the same textual query may convey vastly different intentions depending on a user's preferences (Zou et al. 2023), background knowledge (Westhofen, Jung, and Neider 2025), or expression style (Neelakanteswara, Chaudhari, and Zamani 2024). As users exhibit diverse personas and contextual needs, it becomes increasingly important for QE methods to be personalized—capable of adapting to user-specific semantics within private corpora—so as to ensure accurate retrieval and enhance downstream generation.

As illustrated in Figure 1 (top), when two users submit the same query—"*How would you describe my dietary preferences?*"—standard expansion methods yield generic outputs that fail to account for individual differences. This results in mismatched retrieval: for example, the health-conscious user (User 1) successfully retrieves content related to "*organic ingredients*," which aligns with their preferences. In contrast, the variety-seeking user (User 2), whose personal content emphasizes "*trying new side dishes*," fails to retrieve relevant information, highlighting the limitations of non-personalized expansion strategies. Such failures stem from neglecting rich signals in user—such as style, intent patterns, and thematic focus—resulting in semantically misaligned expansions and degraded retrieval performance.

Motivated by these limitations, we aim to develop a personalized query expansion framework that adapts to individual user intent and expression as shown in Figure 1. However, introducing personalization into expansion brings two fundamental challenges: **(1) User expression styles are inherently diverse.** Users articulate intent using varied linguistic patterns—ranging from minimal prompts to elaborative reasoning—driven by personal habits, domain familiarity, or rhetorical preference (Neelakanteswara, Chaudhari, and Zamani 2024). These styles are often implicit and non-transferable, making it difficult to construct expansions that faithfully preserve user-specific semantics within a generalizable framework. **(2) User corpora induce heterogeneous semantic structures.** Users' personal corpora often differ significantly in topical coverage, content organization, and linguistic granularity. This high degree of heterogeneity makes it difficult to locate reliable semantic regions related to the user query, and in turn, complicates the alignment of user-specific preferences. Without user-specific semantic anchors to guide expansion, queries are prone to drifting toward irrelevant regions, resulting in mismatched retrieval despite plausible surface semantics. These challenges raise a central question: *How can we personalize query expansion—both in style and structure—before retrieval to align with individual user intent and corpus characteristics?*

To address these challenges, we propose a *Personalize Before Retrieve* (**PBR**) framework that integrates user-specific signals into query understanding prior to retrieval. PBR adopts a two-stage design that combines expression-level expansion with structure-level alignment, enabling queries to reflect both personal style and corpus semantics. ❶ **P-PRF: Personal Style-Aligned Pseudo Relevance Feedback.** This module extracts linguistic patterns from user history to guide LLMs in generating pseudo-utterances and reasoning paths. These signals capture personalized expression style and hidden intent, often missed by generic expansion methods. ❷ **P-Anchor: Personal Structure-Aligned Semantic Anchoring.** We encode user history into a semantic graph that captures localized relevance patterns. Personalized PageRank is applied to identify representative anchor points within this structure to encode the whole user corpora or to express user general preference. The query is then guided toward these anchors to reflect the user's semantic landscape. By jointly modeling stylistic variation and structural relevance, P-PRF and P-Anchor produce personalized query representations that enhance alignment between user intent and retrieval semantics in RAG systems. Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to propose a personalized query expansion framework for RAG systems that adapts to the user for improving retrieval.
- We propose **P-PRF**, a style-aware expansion module that generates pseudo queries and reasoning paths conditioned on user history.
- We introduce **P-Anchor**, a graph-based alignment module that grounds queries within personalized semantic spaces via local corpus structure.
- Experiments on two real-world dialogue datasets show that **PBR** significantly improves personalized retrieval, validating the benefits of pre-retrieval personalization.

2 Problem Definition

We study *personalized query expansion*, which expands underspecified queries to align with both user expression style and corpus heterogeneity informed by past interactions.

Formally, let q represent the user query, and let $H = \{h_1, h_2, \ldots, h_n\}$ denote the user's historical utterances, referred to as the user corpus, where each h_i is a textual segment from prior conversations or interactions. The query is encoded as $\mathbf{q} = \phi(q) \in \mathbb{R}^d$, while the historical utterances are encoded into a set of corpus vectors $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ using a fixed encoder $\phi(\cdot)$:

$$\mathbf{c}_i = \phi(h_i) \in \mathbb{R}^d, \quad \forall i = 1, \ldots, n. \tag{1}$$

The goal is to construct a personalized query representation $\mathbf{q}^* \in \mathbb{R}^d$ that retrieves content from the user corpus $\mathcal{C}$ not only based on lexical-semantic similarity, but also aligned with user-specific latent intent. Formally, the retrieved set is defined as:

$$\mathcal{R}(\mathbf{q}^*) = \text{TopK}\left(\{\text{sim}(\mathbf{q}^*, \mathbf{c}_k) \mid \mathbf{c}_k \in \mathcal{C}\}\right), \tag{2}$$

where $\text{sim}(\cdot)$ denotes a similarity function (e.g., cosine similarity), and $\text{Top}_k(\cdot)$ returns the top-k most relevant items.

The core challenge lies in constructing $\mathbf{q}^*$ from the observable query $\mathbf{q}$ and the latent user-specific context $\mathcal{C}$. We formulate a transformation function $f_\Theta(\mathbf{q}, \mathcal{C})$ such that:

$$\mathbf{q}^* = f_\Theta(\mathbf{q}, \mathcal{C}) = \mathbf{q} + \Delta_{\text{user}}(\mathbf{q}, \mathcal{C}), \tag{3}$$

where $\Delta_{\text{user}}(\mathbf{q}, \mathcal{C})$ encodes personalized adjustments reflecting individual expression style, intent formulation, and structural semantics in the user corpus.

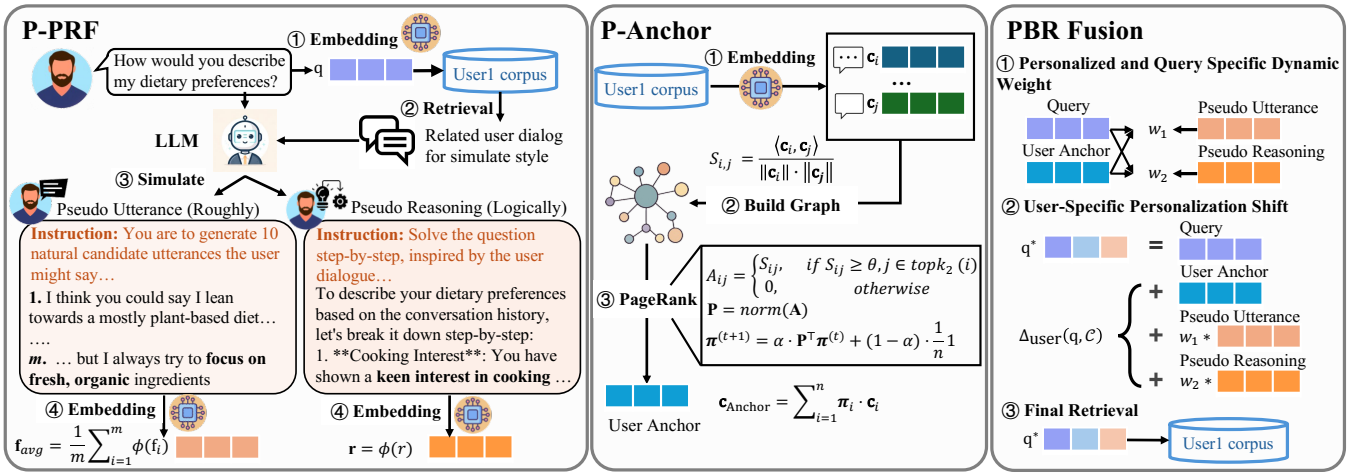


Figure 2: Overview of PBR, which consists of three components: **P-PRF** for stylistic expansion, **P-Anchor** for structural grounding, and **PBR Fusion** module for generating the final query representation.

3 PBR Framework

3.1 Framework Overview

As illustrated in Figure 2, the **PBR** framework enhances personalized retrieval by explicitly incorporating user-specific signals into query representation before retrieval. It consists of three core modules: **P-PRF** simulates two complementary forms of feedback to generate pseudo utterances and pseudo reasoning that reflect the user's stylistic tendencies and underlying reasoning logic. **P-Anchor** constructs a semantic graph over the user's corpus and applies PageRank to identify representative anchor points that capture the corpus-level user anchor. These two sources of personalized signals are integrated via **PBR Fusion**, which computes a personalized and query-specific dynamic weight to balance the contributions of pseudo utterances and pseudo reasoning. The resulting fused representation $\mathbf{q}^*$ reflects both user hidden style and contextual grounding, enabling accurate retrieval from the semantic index constructed over the user corpus C.

3.2 P-PRF

Generic QE methods often overlook the nuanced expression styles (Neelakanteswara, Chaudhari, and Zamani 2024) and implicit reasoning patterns behind user expression (Li et al. 2025), leading to mismatches in personalized retrieval. This challenge is exacerbated when queries are ambiguous or under-specified, requiring systems to infer intent beyond surface forms. Existing solutions typically apply uniform expansion strategies, failing to adapt to user-specific signals.

To address these limitations, **P-PRF** simulates user-specific signals before retrieval by simulating personalized feedback in two complementary forms: *roughly* via stylistic utterances, and *logically* via intent reasoning. Rather than using the full user history H, we retrieve a task-relevant subset $H_q = \text{Top-}k_1\big(\{\text{sim}(\mathbf{q}, \mathbf{c}_t)\}_{t=1}^{n}\big)$ based on semantic similarity, where $\mathbf{q}$ and $\mathbf{c}_t$ are the embeddings of the query and past utterances. This subset captures salient contextual-aware user traits for downstream simulation.

Pseudo Utterance Generation (Roughly) To simulate user-specific expression patterns, we employ a large language model $\mathcal{G}_{\text{LLM}}^{\text{utt}}$ to generate m pseudo utterances conditioned on the original query q and user dialogue history H_q. Each utterance f_i is designed to reflect how the user might naturally articulate the query, capturing stylistic nuances such as tone, verbosity, and phrasing:

$$\{f_1, \ldots, f_m\} = \mathcal{G}_{\text{LLM}}^{\text{utt}}(\mathbf{q}, H_q), \quad \mathbf{f}_{\text{avg}} = \frac{1}{m}\sum_{i=1}^{m}\phi(f_i), \quad (4)$$

where $\phi(f_i)$ denotes the embedding of the i-th utterance and $\mathbf{f}_{\text{avg}}$ is their mean representation.

Pseudo Reasoning Generation (Logically) Beyond surface expression, effective personalization also requires modeling the user's implicit reasoning process. We define $\mathcal{G}_{\text{LLM}}^{\text{rea}}$ as a parallel simulating pipeline to elicit a step-by-step rationale r of user expression:

$$r = \mathcal{G}_{\text{LLM}}^{\text{rea}}(\mathbf{q}, H_q), \quad \mathbf{r} = \phi(r). \quad (5)$$

This rationale introduces logical cues that are often absent from surface queries, thereby providing a complementary semantic signal to guide expansion alignment.

3.3 P-Anchor

Existing methods are often designed to retrieve from a unified RAG corpus, lacking explicit grounding in the structural organization of individual user corpora (Tan et al. 2025; Wu et al. 2025). To address this, we propose **P-Anchor**, a structure-aware module that captures corpus-level preferences by identifying semantically central regions within a graph-structured user space.

Graph Construction from User Corpus. We represent the user's prior corpus—such as interaction records with AI assistants—as a semantic graph $G = (\mathcal{C}, \mathcal{E})$, following previous study (Tang et al. 2025), where $\mathcal{C}$ and $\mathcal{E}$ denote the sets of nodes and edges, respectively. The set of nodes $\mathcal{C}$ consists of history corpus vectors, where each node $\mathbf{c}_i$ as in Eq. (1)

Method	Base		HyDE		Query2Term		MILL		CoT		ThinkQE		PBR (Ours)	
Metrics	R@5	N@5	R@5	N@5	R@5	N@5	R@5	N@5	R@5	N@5	R@5	N@5	R@5	N@5
multi-qa-MiniLM-L6-cos-v1														
Overall	0.4484	0.3669	0.3464	0.2945	0.3584	0.3060	0.3200	0.3254	0.3627	0.3000	0.4791	0.3902	**0.5035**	**0.4201**
Basic information	0.4515	0.3088	0.3106	0.2395	0.3424	0.2614	0.2879	0.2337	0.3424	0.2454	0.4606	0.3265	**0.5091**	**0.3647**
Preference (hard)	0.3659	0.3759	0.3122	0.3079	0.3659	0.3491	0.2927	0.4250	0.3171	0.3335	**0.4195**	**0.4310**	0.4049	0.4175
Social	0.4852	0.4356	0.3909	0.3259	0.3494	0.3144	0.3808	0.3923	0.4009	0.3320	0.5503	0.4554	**0.5541**	**0.4914**
Preference (easy)	0.4904	0.4582	0.4615	0.4422	0.4327	0.4095	0.3750	0.4195	0.4423	0.4133	0.5064	0.4626	**0.5321**	**0.5129**
all-MiniLM-L6-v2														
Overall	0.3783	0.3074	0.3747	0.3144	0.3908	0.3256	0.3030	0.3059	0.3721	0.3098	0.3861	0.3393	**0.4516**	**0.3855**
Basic information	0.3515	0.2644	0.4015	0.2975	0.3894	0.2914	0.3030	0.2547	0.3455	0.2696	0.3455	0.2677	**0.4515**	**0.3485**
Preference (hard)	0.4000	0.3547	0.3512	0.3561	0.3805	0.3798	0.3220	0.4171	0.3463	0.3490	0.3366	0.3486	**0.4341**	**0.4352**
Social	0.3921	0.3048	0.2805	0.2435	0.3984	0.3163	0.2638	0.2829	0.4160	0.3046	**0.4780**	**0.4334**	0.4494	0.3777
Preference (easy)	0.4295	0.4199	**0.4904**	0.4646	0.3974	0.4038	0.3526	0.3944	0.4359	0.4285	0.4487	0.4357	0.4840	**0.4800**
bge-base-en-v1.5														
Overall	0.3738	0.3015	0.3108	0.2597	0.3007	0.2497	0.2791	0.2869	0.3199	0.2642	0.3643	0.3156	**0.4029**	**0.3402**
Basic information	0.3970	0.2748	0.2955	0.2051	0.3000	0.1955	0.2879	0.2319	0.3152	0.2052	0.3152	0.2430	**0.4121**	**0.3057**
Preference (hard)	0.3268	0.3343	0.3073	0.3148	0.2976	0.3186	0.2244	0.3566	0.3073	0.3281	0.3463	0.3635	**0.3707**	**0.3889**
Social	0.3204	0.2799	0.2714	0.2443	0.2852	0.2503	0.2752	0.3004	0.2965	0.2647	**0.4261**	**0.3735**	0.3657	0.3089
Preference (easy)	0.4583	0.4065	0.4615	0.4356	0.3397	0.3693	0.3365	0.3819	0.4071	0.4117	0.4744	0.4289	**0.4904**	**0.4734**
Overall Average	0.4002	0.3253	0.3440	0.2895	0.3500	0.2938	0.3007	0.3061	0.3295	0.3121	0.4098	0.3484	**0.4527***	**0.3819***

Table 1: Retrieval performance in **PersonaBench**. The best results are in **bold**, and the second-best results are underlined. "*" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline.

represents an encoded segment. The edges $\mathcal{E}$ are established based on pairwise similarity between the nodes. For each pair, we compute cosine similarity $S_{ij} = \frac{\langle \mathbf{c}_i, \mathbf{c}_j \rangle}{\|\mathbf{c}_i\| \cdot \|\mathbf{c}_j\|}$. We construct a sparse adjacency matrix A to represent the structural links within the user corpus, which retains only the top-k_2 neighbors exceeding a similarity threshold θ:

$$A_{ij} = \begin{cases} S_{ij}, & \text{if } S_{ij} \geq \theta \text{ and } j \in \text{Top-}k_2(i), \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Graph-Based Anchor Representation. We follow PageRank (Page et al. 1999; Haveliwala 2002) over G to estimate a stationary distribution $\boldsymbol{\pi}$ reflecting node centrality. Let $\mathbf{P} = norm(\mathbf{A})$ be the row-normalized transition matrix, then $\boldsymbol{\pi}^{(t+1)} = \alpha \cdot \mathbf{P}^\top \boldsymbol{\pi}^{(t)} + (1-\alpha) \cdot \frac{1}{n}\mathbf{1}$. The user anchor is computed as a weighted aggregation:

$$\mathbf{c}_{\text{Anchor}} = \sum_{i=1}^{n} \boldsymbol{\pi}_i \cdot \mathbf{c}_i. \tag{7}$$

This representation encodes structurally central semantics from the user's corpus, anchoring the query in a context-aware and personalized semantic space.

3.4 PBR Fusion

To synthesize the signals captured in P-PRF and P-Anchor, we design a fusion mechanism that integrates stylistic pseudo-feedback with the query and user anchor. Recognizing that the utility of pseudo-feedback varies across users and queries—depending on individual expression patterns and the semantic openness of the query—we introduce a **personalized and query-specific dynamic weight** that balances the contribution of P-PRF according to its semantic proximity to both the original query $\mathbf{q}$ and the corpus anchor $\mathbf{c}_{\text{Anchor}}$. Specifically, we compute two weights w_1 and w_2 to quantify the semantic alignment of the pseudo-utterance and reasoning feedback, respectively:

$$w_1 = 1 + \text{sim}\left(\frac{\mathbf{q} + \mathbf{c}_{\text{Anchor}}}{2}, \mathbf{f}_{\text{avg}}\right), \tag{8}$$

$$w_2 = 1 + \text{sim}\left(\frac{\mathbf{q} + \mathbf{c}_{\text{Anchor}}}{2}, \mathbf{r}\right), \tag{9}$$

where $\text{sim}(\cdot)$ denotes cosine similarity. The additive constant ensures positivity and smooth interpolation.

Then, we define a **user-specific personalization shift** as:

$$\Delta_{\text{user}}(\mathbf{q}, \mathcal{C}) = \mathbf{c}_{\text{Anchor}} + w_1 \cdot \mathbf{f}_{\text{avg}} + w_2 \cdot \mathbf{r}, \tag{10}$$

and construct the final personalized query embedding as $\mathbf{q}^* = \mathbf{q} + \Delta_{\text{user}}(\mathbf{q}, \mathcal{C})$. The resulting query $\mathbf{q}^*$ incorporates user-specific semantics from both style and structure. It encodes not only the immediate surface form of the input query, but also latent stylistic preferences, goal-driven reasoning, and historically salient corpus signals.

We perform the **final retrieval** from the user corpus $\mathcal{C}$ using `faiss`-based (Douze et al. 2024) nearest neighbor search with $\mathbf{q}^*$. This process leverages embedding-level alignment to recover user-relevant content.

4 Experiment

In this section, we conduct experiments to address the following research questions:

- **RQ1**: *Does the proposed PBR method improve retrieval performance on personalized corpus?*
- **RQ2**: *What are the individual contributions of the P-PRF and P-Anchor modules to the overall performance?*
- **RQ3**: *How does information propagation within the P-Anchor module affect the alignment between the query and user-specific corpora?*

Dataset	LongMenEval-s						LongMenEval-m					
Metrics	R@1	N@1	R@3	N@3	R@5	N@5	R@1	N@1	R@3	N@3	R@5	N@5
multi-qa-MiniLM-L6-cos-v1												
Base	0.1885	0.7924	0.6730	0.8020	0.8067	0.8301	0.1098	0.5251	0.4057	0.5559	0.5394	0.6038
HyDE	0.1957	0.7852	0.6945	0.8173	0.8115	0.8436	0.1146	0.4964	0.4129	0.5411	0.5179	0.5828
Query2Term	0.1814	0.7446	0.6396	0.7684	0.7470	0.7959	0.1146	0.4749	0.3747	0.5091	0.4821	0.5531
MILL	0.1838	0.7709	0.6635	0.7845	0.6635	0.7712	0.1122	0.5155	0.4129	0.5552	0.4129	0.5468
CoT	0.1862	0.7470	0.6516	0.7707	0.7542	0.8006	0.1098	0.4654	0.3866	0.5110	0.4726	0.5475
ThinkQE	0.1933	0.8019	0.7064	0.8214	0.8162	0.8424	0.1169	0.5227	0.4168	0.5647	0.5517	0.6156
PBR	**0.2100***	**0.8210***	**0.7279***	**0.8444***	**0.8282***	**0.8598***	**0.1217***	**0.5537***	**0.4320***	**0.5776***	**0.5537***	**0.6177***
all-MiniLM-L6-v2												
Base	0.2196	0.8234	0.7399	0.8477	0.8568	0.8731	0.1265	0.5322	0.4415	0.5857	0.5346	0.6191
HyDE	0.2172	0.8138	0.7208	0.8372	0.8377	0.8599	0.1074	0.4988	0.4081	0.5483	0.5131	0.5884
Query2Term	0.2196	0.8019	0.6778	0.8000	0.8138	0.8351	0.1193	0.4940	0.3747	0.5209	0.4964	0.5689
MILL	0.2100	0.8162	0.7208	0.8382	0.7208	0.8257	0.1217	0.5179	0.3962	0.5530	0.3962	0.5450
CoT	0.2124	0.7709	0.6730	0.7932	0.7780	0.8187	0.1193	0.4487	0.3389	0.4920	0.4344	0.5273
ThinkQE	0.2196	0.8234	0.7232	0.8332	0.8162	0.8527	0.1241	0.5585	0.4461	0.5871	0.5537	0.6229
PBR	**0.2267***	**0.8592***	**0.7780***	**0.8754***	**0.8640***	**0.8902***	**0.1408***	**0.5800***	**0.4463***	**0.5949***	**0.5847***	**0.6424***
bge-base-en-v1.5												
Base	0.2267	0.8616	0.7900	0.8863	0.8926	0.9007	0.1217	0.5943	0.5322	0.6605	0.6444	0.6929
HyDE	0.2196	0.8759	0.7995	0.8974	0.8998	0.9044	0.1313	0.5943	0.5298	0.6538	0.6659	0.7010
Query2Term	0.2100	0.8544	0.7637	0.8739	0.8926	0.8989	0.1360	0.5776	0.5155	0.6290	0.6348	0.6754
MILL	0.2267	0.8663	0.7804	0.8819	0.7804	0.8680	0.1289	0.5967	0.5107	0.6362	0.5107	0.6278
CoT	0.2196	0.8425	0.7637	0.8707	0.8783	0.8918	0.1384	0.5609	0.4964	0.6148	0.5967	0.6570
ThinkQE	0.2124	0.8807	0.8108	0.9012	0.9021	0.9126	0.1384	0.6072	0.5080	0.6321	0.6445	0.7013
PBR	**0.2315***	**0.8902***	**0.8138***	**0.9072***	**0.9045***	**0.9191***	**0.1408***	**0.6205***	**0.5489***	**0.6612***	**0.6874***	**0.7060***

Table 2: Retrieval performance in **LongMemEval**. The best results are in **bold**, and the second-best results are underlined. "*****" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline.

4.1 Experiment Setting

Datasets. We evaluate PBR on two benchmarks tailored for personalized retrieval. **PersonaBench** (Tan et al. 2025) simulates user-specific queries over synthetic private data. It includes both explicit and implicit user traits across diverse domains. **LongMemEval** (Wu et al. 2025) targets long-term interactive memory, with two subsets: *LongMemEval-s* and *LongMemEval-m*, to evaluate the retrieval performance under both short- and long-horizon personalization.

Evaluation Metrics. We report performance using two standard retrieval metrics: **Recall@K (R@K)**, which measures the proportion of relevant items ranked within the top-K results, and **NDCG@K (N@K)** to reflect retrieval quality.

Baselines. We compare PBR with six baselines: **Base**, a static query retriever without expansion; **HyDE** (Gao et al. 2023), which generates hypothetical expansion; **Query2Term** (Jagerman et al. 2023), which performs keyphrase-based query expansion; **MILL** (Jia et al. 2023), which leverages mutual varified expansion; **CoT** (Wei et al. 2022), which applies chain-of-thought reasoning for better ranking; and **ThinkQE** (Lei, Shen, and Yates 2025).

Implement details. To verify the robustness of PBR, all methods are evaluated under three retrieval backbones: `multi-qa-MiniLM-L6-cos-v1` (Wang et al. 2020), `all-MiniLM-L6-v2` (Wang et al. 2020), and `bge-base-en-v1.5` (Chen et al. 2024). For a fair comparison, all methods are using GPT-4o-mini (Achiam et al. 2023), and run 5 times. In **PBR** framework, we set $k_1 = 5$,$m = 5$, $\theta = 0.75$, and $k_2 = 10$ across *PersonaBench* and *LongMemEval-s*. The larger $k_2 = 50$ is particularly suited for *LongMemEval-m*, which contains richer user histories.

4.2 Overall Performance (RQ1)

To comprehensively evaluate the effectiveness of PBR, we conduct experiments on two benchmarks. **PersonaBench** contains 6 users, each with a dedicated corpus and about 50 user queries characterized by ambiguous intent and highly personalized language. In contrast, **LongMemEval** simulates a large-scale factual QA scenario with 500 distinct user queries, each paired with a relevant memory corpus. The former highlights soft personalization challenges, while the latter emphasizes hard factual grounding under long-context retrieval.

PBR resolving semantic ambiguity through user-specific expansion. Queries in PersonaBench often exhibit semantic openness and personal references—such as "Where is my hometown?" or "What is my favorite color?"—which cannot be resolved without grounding in the user's unique corpus. As shown in Table 1, our PBR framework achieves the best overall performance across all retrievers (R@5: **0.4527**, N@5: **0.3819**), significantly surpassing the best baseline (ThinkQE: 0.4098 / 0.3484). This gain stems from PBR's ability to generate user-style pseudo-feedback that anchors the query into a meaningful semantic space, reducing ambiguity and improving retrieval accuracy.

PBR excels in factual retrieval by boosting top-ranked precision. In contrast to PersonaBench, queries in LongMemEval are fact-oriented with specific personalization requirements, each associated with a specific memory corpus. In this setting, PBR boosts top-ranked retrieval performance. In Table 2, PBR also outperforms all baselines on LongMemEval, particularly on R@1, where it achieves **0.2315** on the -s version and **0.1408** on the -m version using `bge`.

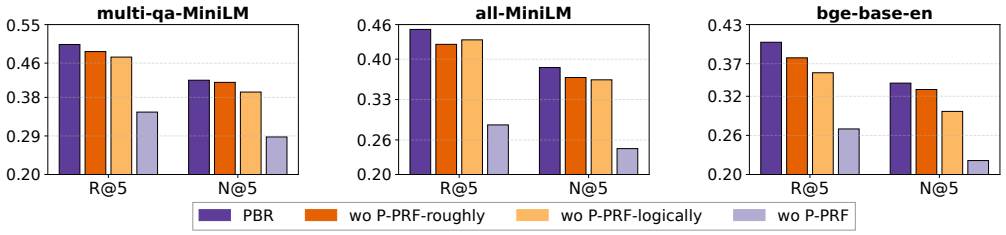


Figure 3: Overall performance in **PersonaBench** comparison across three retrieval models.

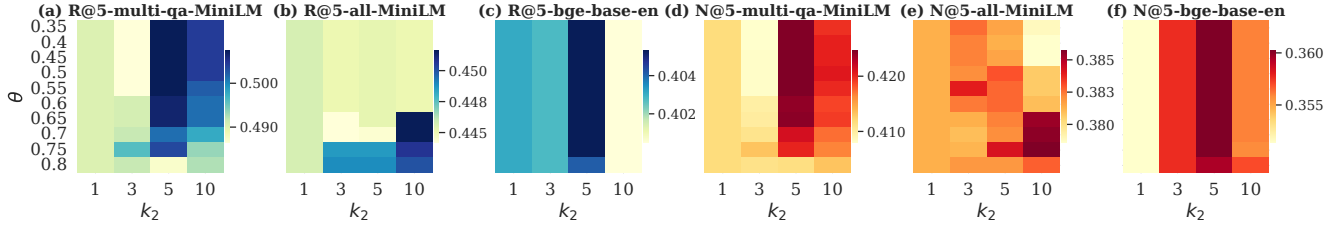


Figure 4: Parameter sensitivity analysis on P-Anchor of θ and k_2 settings.

Compared to the best-performing baseline (0.2267 / 0.1384), PBR achieves absolute gain on top-1 precision. This suggests that our style simulation and structural anchoring more precisely locate relevant memory, elevating correct answers to top-1 even in long-context settings.

4.3 Ablation Study (RQ2)

	PBR		w/o P-Anchor		w/o P-PRF	
Metrics	R@5	N@5	R@5	N@5	R@5	N@5
multi-qa-MiniLM-L6-cos-v1						
Overall	**0.5035**	**0.4201**	0.4871	0.4100	0.3457	0.2877
Basic info	**0.5091**	**0.3647**	0.4833	0.3570	0.3318	0.2242
Pref (hard)	0.4049	0.4175	**0.4390**	**0.4295**	0.2683	0.2981
Social	**0.5541**	**0.4914**	0.5164	0.4529	0.4277	0.3851
Pref (easy)	**0.5321**	**0.5129**	0.5192	0.5160	0.3590	0.3414
all-MiniLM-L6-v2						
Overall	**0.4516**	**0.3855**	0.4382	0.3729	0.2860	0.2449
Basic info	**0.4515**	**0.3485**	0.4379	0.3393	0.1985	0.1617
Pref (hard)	**0.4341**	**0.4352**	0.4244	0.4201	0.2878	0.2718
Social	**0.4494**	**0.3777**	0.4333	0.3639	0.4179	0.3387
Pref (easy)	**0.4840**	**0.4800**	0.4712	0.4587	0.3846	0.3634
BAAI/bge-base-en-v1.5						
Overall	**0.4029**	**0.3402**	0.3921	0.3256	0.2699	0.2214
Basic info	**0.4121**	**0.3057**	0.3970	0.2958	0.2576	0.1804
Pref (hard)	0.3707	**0.3889**	**0.3902**	0.3778	0.2146	0.2173
Social	**0.3657**	**0.3089**	0.3431	0.2822	0.3261	0.2816
Pref (easy)	**0.4904**	**0.4734**	0.4744	0.4582	0.2949	0.2785
Average	**0.4527**	**0.3819**	0.4391	0.3695	0.3005	0.2513

Table 3: Ablation study of PBR in **PersonaBench**.

To investigate the individual roles of **P-PRF** and **P-Anchor** in PBR, we perform a module-wise ablation across three retrievers in PersonaBench, as shown in Table 3. Each model variant is evaluated under both R@5 and N@5.

P-PRF drives major performance gains in personalized retrieval. Removing P-PRF leads to a substantial performance drop across all retrievers (e.g., on all-MiniLM, R@5 falls from 0.4516 to 0.2860), with the most severe decline observed on personalization-heavy subsets. This underscores the importance of P-PRF in generating semantically rich and stylistically aligned pseudo queries that effectively capture user intent.

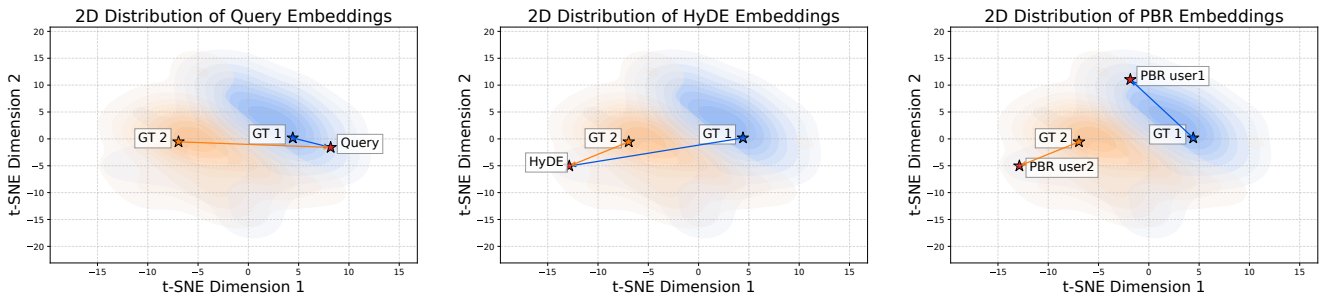
Effectiveness of P-PRF Components. To assess how **P-PRF** enhances intent modeling, we perform ablation studies in PersonaBench and three retrievers (Figure 3). Removing either the *roughly*-aligned pseudo-utterance or *logically*-structured pseudo-reasoning consistently degrades performance in R@5 and N@5. The former harms lexical coverage, while the latter impairs goal-oriented reasoning, reducing ranking precision on inference-heavy datasets.

P-Anchor promotes semantic alignment in structured corpora. Removing P-Anchor leads to a consistent drop (e.g., N@5 from **0.3819** to 0.3695), especially in *Basic Information* and *Social*, where user corpora exhibit clearer structure. In contrast, in non-clustered tasks like *Preference (hard)*, it offers limited gain and may overfit. P-Anchor is most effective when user context is semantically coherent.

4.4 Parameter Sensitivity Analysis (RQ3)

To verify how information propagation within the P-Anchor module influences the alignment between queries and user-specific semantic space, we conduct a sensitivity analysis over two structural hyperparameters: the semantic threshold θ and the neighbor size k_2. The results are shown in Figure 4.

Moderate propagation improves alignment. When $\theta \in [0.65, 0.75]$ and $k_2 = 5$, both R@5 and N@5 reach consistent peaks across datasets, indicating that controlled expansion of semantic neighborhoods facilitates better anchoring and personalization. **Over-propagation causes semantic drift.** As θ and k_2 continue to increase, performance either saturates or slightly declines (e.g., at $\theta = 0.8$, $k_2 = 10$), suggesting that excessive propagation introduces noise and



(a) Query-GT Comparison (b) HyDE-GT Comparison (c) PBR-GT Comparison (Ours)

Figure 5: Visualization of retrieval distance of the query in the Figure 1 compared to ground truth (GT) under different methods: (a) vanilla query, (b) HyDE, and (c) our proposed PBR.

weakens the semantic specificity of the alignment.

4.5 Visualization Case Study

To verify the practical effectiveness of our proposed PBR framework, we conduct a retrieval visualization case study. Figure 5 visualizes the t-SNE distribution of user queries, ground truths (GT), and generated queries under different methods. The results show that traditional expansion methods produce generic results misaligned with different users. Measuring cosine similarity to user-specific ground truth, PBR achieves higher separation between users (0.31 vs. 0.27), outperforming HyDE (0.04 vs. –0.01) and the original query (0.05 vs. 0.10). This highlights PBR's effectiveness in embedding user-personalized semantics.

P-PRF enables semantically diverse and intent-aligned queries. Compared to vanilla queries or HyDE-generated expansions, PBR-generated queries are semantically closer to ground truth responses in both user clusters. The trajectory of PBR vectors more accurately captures user-relevant preferences, confirming its effectiveness in generating expressive and personalized query variants.

Improved alignment leads to more successful retrieval. Unlike HyDE, which generates a generic QE of all users, PBR yields multiple expansions (e.g., PBR user1 and PBR user2) that better cover the user-specific semantic space. This personalized QE allows the retriever to more precisely locate relevant results near each GT, reducing retrieval distance and improving semantic alignment.

5 Literature Review

5.1 Query Expansion in RAG System

Query expansion (QE) improves retrieval performance by simulating PRF. With the rise of LLMs, recent studies have explored LLMs as zero-shot QE engines. HyDE (Gao et al. 2023) and Query2Doc (Wang, Yang, and Wei 2023) demonstrate that prompting LLMs can yield precise expansions with PRF. Other works enhance semantic diversity or reasoning depth (Zhang et al. 2025b): MILL (Jia et al. 2023) employs mutual verification, and GRF (Mackie, Chatterjee, and Dalton 2023) applies LLM-based relevance feedback. ThinkQE (Lei, Shen, and Yates 2025) models expansion as a reasoning process, while LLMlingua (Jiang et al. 2023) focuses on compressing expansion prompts. GAR (Xia et al. 2024) further introduces a knowledge graph-aware QE framework.

However, existing work overlooks personalization and fails to align QE with user-specific semantic spaces. In contrast, our PBR generates style-aware and structure-aware expansions, bridging this gap for personalized retrieval.

5.2 Personalized RAG System

Personalized RAG system has recently emerged as a crucial direction for adapting LLMs to individual users. A recent survey by Li et al. (Li et al. 2025, 2023b) provides a comprehensive overview, highlighting the importance of personalization in RAG. Approaches such as UniMS-RAG (Wang et al. 2024a) integrate multi-source signals for dialogue personalization, while PersonaRAG (Zerhoudi and Granitzer 2024) enhances personalization via role-based agent frameworks. Salemi et al. (Salemi et al. 2024) further investigate LLM personalization by modeling language preferences and user goals. Complementary work like Wu et al. (Wu et al. 2024) explores the role of explicit user profiles, and Cohn et al. (Cohn et al. 2025) leverage interaction logs for educational agent adaptation. Domain-specific applications, such as personalized care systems (Yang et al. 2025; Liu et al. 2024) and memory-based assistant evaluation (Wu et al. 2025; Xu et al. 2025c,a), highlight the importance of long-term memory and evolving user states in effective RAG.

Despite recent advances, little attention has been given to improving retrieval accuracy at the pre-retrieval stage. PBR enables fine-grained personalization by jointly modeling the user's style and corpus structure before retrieval.

6 Conclusion

This paper introduces **PBR**, a *Personalize-Before-Retrieve* framework for personalized retrieval. PBR expands user query prior to retrieval by integrating two components: **P-PRF**, a style-aware pseudo query generator, and **P-Anchor**, a graph-based corpus alignment module. Our results highlight the importance of pre-retrieval personalization for improving user-centric retrieval performance. Future work will investigate generalizing PBR to broader information-seeking tasks beyond query-based retrieval.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (NSFC) under Grant 72071029, 72231010, 62502404, and the Graduate Research Fund of the School of Economics and Management of Dalian University of Technology (No. DUTSEMDRFKO1). This research was partially supported by Hong Kong Research Grants Council (Research Impact Fund No.R1015-23, Collaborative Research Fund No.C1043-24GF, General Research Fund No.11218325), Institute of Digital Medicine of City University of Hong Kong (No.9229503), and Huawei (Huawei Innovation Research Program).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bogdan, P. C.; Macar, U.; Nanda, N.; and Conmy, A. 2025. Thought Anchors: Which LLM Reasoning Steps Matter? *arXiv preprint arXiv:2506.19143*.

Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; and Liu, Z. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Cheng, X.; Luo, D.; Chen, X.; Liu, L.; Zhao, D.; and Yan, R. 2023. Lift yourself up: Retrieval-augmented text generation with self-memory. *Proc. of NeurIPS*, 43780–43799.

Cohn, C.; Rayala, S.; Snyder, C.; Fonteles, J.; Jain, S.; Mohammed, N.; Timalsina, U.; Burriss, S. K.; Srivastava, N.; Deweese, M.; et al. 2025. Personalizing Student-Agent Interactions Using Log-Contextualized Retrieval Augmented Generation (RAG). *arXiv preprint arXiv:2505.17238*.

Culpepper, J. S.; Faggioli, G.; Ferro, N.; and Kurland, O. 2021. Topic difficulty: Collection and query formulation effects. *ACM Transactions on Information Systems (TOIS)*, 1–36.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Fu, Z.; Li, X.; Wu, C.; Wang, Y.; Dong, K.; Zhao, X.; Zhao, M.; Guo, H.; and Tang, R. 2025. A unified framework for multi-domain ctr prediction via large language models. *ACM Transactions on Information Systems*, 1–33.

Gao, J.; Chen, B.; Zhao, X.; Liu, W.; Li, X.; Wang, Y.; Zhang, Z.; Wang, W.; Ye, Y.; Lin, S.; et al. 2024. Llm-enhanced reranking in recommender systems.

Gao, L.; Ma, X.; Lin, J.; and Callan, J. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proc. of ACL*, 1762–1777.

Haveliwala, T. H. 2002. Topic-sensitive PageRank. *Proceedings of the 11th international conference on World Wide Web*, 517–526.

Jagerman, R.; Zhuang, H.; Qin, Z.; Wang, X.; and Bendersky, M. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653*.

Jia, P.; Liu, Y.; Zhao, X.; Li, X.; Hao, C.; Wang, S.; and Yin, D. 2023. Mill: Mutual verification with large language models for zero-shot query expansion. *arXiv preprint arXiv:2310.19056*.

Jiang, H.; Wu, Q.; Lin, C.-Y.; Yang, Y.; and Qiu, L. 2023. LLMLingua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736*.

Langville, A. N.; and Meyer, C. D. 2006. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press.

Lei, Y.; Shen, T.; and Yates, A. 2025. ThinkQE: Query Expansion via an Evolving Thinking Process. *arXiv preprint arXiv:2506.09260*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Proc. of NeurIPS*, 9459–9474.

Li, C.; Wang, Y.; Liu, Q.; Zhao, X.; Wang, W.; Wang, Y.; Zou, L.; Fan, W.; and Li, Q. 2023a. STRec: Sparse transformer for sequential recommendations. In *Proceedings of the 17th ACM conference on recommender systems*, 101–111.

Li, X.; Jia, P.; Xu, D.; Wen, Y.; Zhang, Y.; Zhang, W.; Wang, W.; Wang, Y.; Du, Z.; Li, X.; et al. 2025. A survey of personalization: From rag to agent. *arXiv preprint arXiv:2504.10147*.

Li, X.; Qiu, Z.; Zhao, X.; Wang, Z.; Zhang, Y.; Xing, C.; and Wu, X. 2022. Gromov-wasserstein guided representation learning for cross-domain recommendation. In *Proc. of CIKM*, 1199–1208.

Li, X.; Su, L.; Jia, P.; Zhao, X.; Cheng, S.; Wang, J.; and Yin, D. 2023b. Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llm. *arXiv preprint arXiv:2312.15450*.

Li, X.; Yan, F.; Zhao, X.; Wang, Y.; Chen, B.; Guo, H.; and Tang, R. 2023c. Hamur: Hyper adapter for multi-domain recommendation. In *Proc. of CIKM*, 1268–1277.

Liu, Q.; Wu, X.; Zhao, X.; Zhu, Y.; Zhang, Z.; Tian, F.; and Zheng, Y. 2024. Large language model distilling medication recommendation model. *arXiv preprint arXiv:2402.02803*.

Mackie, I.; Chatterjee, S.; and Dalton, J. 2023. Generative relevance feedback with large language models. In *Proc. of SIGIR*, 2026–2031.

Meyer, C. D. 2000. *Matrix analysis and applied linear algebra*. SIAM.

Neelakanteswara, A.; Chaudhari, S.; and Zamani, H. 2024. RAGs to style: Personalizing LLMs with style embeddings. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, 119–123.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. *Stanford InfoLab*.

Salemi, A.; Mysore, S.; Bendersky, M.; and Zamani, H. 2024. LaMP: When Large Language Models Meet Personalization. In *Proc. of ACL*, 7370–7392.

Song, M.; and Zheng, M. 2024. A Survey of Query Optimization in Large Language Models. *arXiv preprint arXiv:2412.17558*.

Tan, J.; Yang, L.; Liu, Z.; Liu, Z.; Murthy, R.; Awalgaonkar, T. M.; Zhang, J.; Yao, W.; Zhu, M.; Kokane, S.; et al. 2025. Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data. *arXiv preprint arXiv:2502.20616*.

Tang, R.; Zhu, C.; Chen, B.; Zhang, W.; Zhu, M.; Dai, X.; and Guo, H. 2025. LLM4Tag: Automatic Tagging System for Information Retrieval via Large Language Models. *arXiv preprint arXiv:2502.13481*.

Wang, H.; Huang, W.; Deng, Y.; Wang, R.; Wang, Z.; Wang, Y.; Mi, F.; Pan, J. Z.; and Wong, K.-F. 2024a. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *arXiv preprint arXiv:2401.13256*.

Wang, H.; Liu, X.; Fan, W.; Zhao, X.; Kini, V.; Yadav, D.; Wang, F.; Wen, Z.; Tang, J.; and Liu, H. 2024b. Rethinking large language model architectures for sequential recommendations. *arXiv preprint arXiv:2402.09543*.

Wang, L.; Yang, N.; and Wei, F. 2023. Query2doc: Query Expansion with Large Language Models. In *Proc. of EMNLP*, 9414–9423.

Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Proc. of NeurIPS*, 5776–5788.

Wang, Y.; Wang, Y.; Fu, Z.; Li, X.; Wang, W.; Ye, Y.; Zhao, X.; Guo, H.; and Tang, R. 2024c. Llm4msr: An llm-enhanced paradigm for multi-scenario recommendation. In *Proc. of CIKM*, 2472–2481.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Proc. of NeurIPS*, 24824–24837.

Westhofen, L.; Jung, J. C.; and Neider, D. 2025. Temporal Conjunctive Query Answering via Rewriting. In *Proc. of AAAI*, 15221–15229.

Wu, B.; Shi, Z.; Rahmani, H. A.; Ramineni, V.; and Yilmaz, E. 2024. Understanding the Role of User Profile in the Personalization of Large Language Models. *arXiv preprint arXiv:2406.17803*.

Wu, D.; Wang, H.; Yu, W.; Zhang, Y.; Chang, K.-W.; and Yu, D. 2025. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. In *Proc. of ICLR*.

Xia, Y.; Wu, J.; Kim, S.; Yu, T.; Rossi, R. A.; Wang, H.; and McAuley, J. 2024. Knowledge-aware query expansion with large language models for textual and relational retrieval. *arXiv preprint arXiv:2410.13765*.

Xu, D.; Jia, P.; Li, X.; Zhang, Y.; Wang, M.; Liu, Q.; Zhao, X.; Wang, Y.; Guo, H.; Tang, R.; et al. 2025a. Align-GRAG: Reasoning-Guided Dual Alignment for Graph Retrieval-Augmented Generation. *arXiv preprint arXiv:2505.16237*.

Xu, D.; Li, X.; Zhang, Z.; Lin, Z.; Zhu, Z.; Zheng, Z.; Wu, X.; Zhao, X.; Xu, T.; and Chen, E. 2025b. Harnessing large language models for knowledge graph question answering via adaptive multi-aspect retrieval-augmentation. In *Proc. of AAAI*, 25570–25578.

Xu, D.; Wen, Y.; Jia, P.; Zhang, Y.; Wang, Y.; Guo, H.; Tang, R.; Zhao, X.; Chen, E.; Xu, T.; et al. 2025c. Towards Multi-Granularity Memory Association and Selection for Long-Term Conversational Agents. *arXiv preprint arXiv:2505.19549*.

Yang, Y.; Xu, C.; Guo, J.; Feng, T.; and Ruan, C. 2025. Improving the RAG-based personalized discharge care system by introducing the memory mechanism. In *2025 IEEE 17th International Conference on Computer Research and Development (ICCRD)*, 316–322.

Zerhoudi, S.; and Granitzer, M. 2024. PersonaRAG: Enhancing Retrieval-Augmented Generation Systems with User-Centric Agents. *arXiv preprint arXiv:2407.09394*.

Zhang, C.; Zhang, H.; Wu, S.; Wu, D.; Xu, T.; Zhao, X.; Gao, Y.; Hu, Y.; and Chen, E. 2025a. Notellm-2: Multimodal large representation models for recommendation. In *Proc. of KDD*, 2815–2826.

Zhang, W.; Li, X.; Dong, K.; Wang, Y.; Jia, P.; Li, X.; Zhang, Y.; Xu, D.; Du, Z.; Guo, H.; et al. 2025b. Process vs. Outcome Reward: Which is Better for Agentic RAG Reinforcement Learning. *arXiv preprint arXiv:2505.14069*.

Zhang, Y.; Jia, P.; Li, X.; Xu, D.; Wang, M.; Wang, Y.; Du, Z.; Guo, H.; Liu, Y.; Tang, R.; et al. 2025c. LSRP: A Leader-Subordinate Retrieval Framework for Privacy-Preserving Cloud-Device Collaboration. *arXiv preprint arXiv:2505.05031*.

Zhang, Z.; Liu, S.; Liu, Z.; Zhong, R.; Cai, Q.; Zhao, X.; Zhang, C.; Liu, Q.; and Jiang, P. 2025d. Llm-powered user simulator for recommender system. In *Proc. of AAAI*, 13339–13347.

Zhang, Z.; Liu, S.; Yu, J.; Cai, Q.; Zhao, X.; Zhang, C.; Liu, Z.; Liu, Q.; Zhao, H.; Hu, L.; et al. 2024. M3oe: Multi-domain multi-task mixture-of experts recommendation framework. In *Proc. of SIGIR*, 893–902.

Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; and Cui, B. 2024. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Zhao, X.; Xia, L.; Zhang, L.; Ding, Z.; Yin, D.; and Tang, J. 2018a. Deep reinforcement learning for page-wise recommendations. In *Proceedings of the 12th ACM conference on recommender systems*, 95–103.

Zhao, X.; Zhang, L.; Ding, Z.; Xia, L.; Tang, J.; and Yin, D. 2018b. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proc. of KDD*, 1040–1048.

Zou, J.; Aliannejadi, M.; Kanoulas, E.; Pera, M. S.; and Liu, Y. 2023. Users meet clarifying questions: Toward a better understanding of user interactions for search clarification. *ACM transactions on information systems*, 1–25.