Personalized RAG in the Wild: Benchmarking Personalized RAG with Authentic User Behavioral Signals

Anonymous ACL submission

Abstract

Personalized Retrieval-Augmented Generation (Personalized RAG) aims to incorporate user-specific signals into retrieval-augmented generation pipeline to better support individualized information needs. However, existing research in this area relies heavily on synthetic data and lacks unified benchmarks for evaluating the full personalization process across retrieval and generation stages, failing to capture the complexity of real-world user intent. To address these limitations, we introduce the Personalized RAG Benchmark (PRBEN)[1], a benchmark derived from the anonymized logs of a mainstream commercial search engine. We extract multi-level personalization signals, including query histories, user attributes, and click feedback, to formulate realistic personalized retrieval and generation tasks. Furthermore, we design an end-to-end evaluation pipeline supervised by actual user click-through data and human-curated golden standards for generated responses. Specifically, we propose a novel metric system, including Accuracy scores and Personalization Alignment, to measure a model's ability to capture user intent while avoiding hallucinated preferences. Experiments demonstrate that PRBEN supports comprehensive evaluation of personalized RAG models, particularly in their ability to disambiguate user intents, and provides a reliable testbed for advancing user-aware RAG systems.

1 Introduction

Retrieval-Augmented Generation (RAG) has emerged as a dominant paradigm for open-domain Question Answering, mitigating the hallucination issues of Large Language Models (LLMs) by grounding responses in external knowledge (Lewis et al., 2020b; Guu et al., 2020). However, most existing RAG systems operate on a "one-size-fits-all" design, applying identical retrieval and generation strategies to the same query regardless of the user, despite the fact that user intent is often implicit, ambiguous, and highly personalized (Li et al., 2025a). Without incorporating personalization signals such as user histories, these adhoc RAG systems struggle to resolve intent ambiguity, often retrieving evidence weakly aligned with user current intents and consequently producing responses that fail to satisfy the user's tailored information needs (Li et al., 2025b). As illustrated by the query "Golf" (which may refer to the Volkswagen Golf or the outdoor sport) in Figure 1, a conventional RAG system may retrieve mixed results and generate intent-misaligned responses. In contrast, leveraging historical user interactions (e.g., queries about cars) enables personalized RAG systems to disambiguate intent and align both retrieval and generation with individual information needs. Therefore, incorporating personalized signals is essential for effective RAG systems.

Figure 1: Preference incorporation and disambiguation in personalized RAG. The query "Golf" illustrates how personalized information is used to resolve semantic ambiguity and generate outputs aligned with user preferences.

However, despite this growing interest, standardized benchmarks for evaluating personalized RAG remain limited. Existing datasets fall short in supporting systematic evaluation of personalized retrieval-generation pipelines, mainly due to shortcomings in both data construction and evaluation protocols. First, **most benchmarks rely**

[1]Code and benchmark are publicly available at https://github.com/prben-ai/PRBEN.

on LLM-simulated user profiles or manual templates (Zhou et al., 2024; Tan et al., 2025). Consequently, these datasets lack noisy and drift feedback signals (e.g., imperfect clicks or skipped results), making it difficult to evaluate how personalized models behave under realistic user feedback conditions (Ryan et al., 2025). Second, **current studies often evaluate retrieval and generation separately**, ignoring the "butterfly effect" in which personalized retrieval directly affects generation faithfulness (Zhang et al., 2025b). As a result, models optimized on decoupled metrics may score highly on benchmarks but generalize poorly in practice.

As summarized in Table 1, we compare representative benchmarks across three key aspects: personalization signals (search logs, personalization), retrieval-generation coverage (retrieval, generation, end-to-end retrieval-augmented generation), and real-world behavioral grounding (click signals, human annotation). We cover diverse datasets, including AOL Search Log (Guo et al., 2021), MS MARCO (Nguyen et al., 2016), LaMP (Salemi et al., 2024), PersonaBench (Tan et al., 2025), BESPOKE (Kim et al., 2025), and EnronQA (Ryan et al., 2025), spanning real-world log-based and synthetic persona-driven settings. Overall, existing benchmarks focus on isolated retrieval or generation tasks, lacking a unified end-to-end evaluation framework grounded in authentic user behavior.

To bridge this gap, we introduce **PRBEN**, a personalized RAG benchmark driven by real-world user behavioral signals. Unlike prior benchmarks constructed from synthetic user profiles, PRBEN is derived from anonymized search logs of a large-scale commercial search engine, enabling the reconstruction of realistic User-Query-Click interaction contexts (Liu et al., 2023). Such real-world interactions provide naturally occurring personalization signals that reflect authentic user behavior. Furthermore, we design a unified evaluation framework that models personalization as a continuous signal flow, assessing a model's ability to translate implicit user histories into explicit personalized responses (Zhang et al., 2025a). PRBEN supports an end-to-end evaluation setting in which systems must first infer user intent from behavioral signals, then retrieve candidate documents aligned with inferred preferences, and finally generate responses that integrate factual evidence while remaining consistent with user-specific information needs (Lewis et al., 2020b; Zerhoudi and Granitzer, 2024).

We further propose a multi-dimensional metric system that jointly measures retrieval accuracy and personalization alignment (Li et al., 2025a). Retrieval metrics (Recall, MRR, NDCG, MAP, Relevance) assess how personalized signals guide retrieval, while generation metrics (BLEU, ROUGE-L, PTC, PGDC) and factual fidelity metrics (Faithfulness, Factuality) evaluate preference alignment and detect personalization hallucinations. Experiments on PRBEN show that personalized RAG remains challenging for state-of-the-art LLMs: retrieval recall stays below 30% with naive history concatenation, static user attributes may introduce noise, and performance drops sharply under noisy retrieval. These results establish PRBEN as a demanding benchmark requiring robustness, long-context reasoning, and intent disambiguation. In summary, the main contributions of this framework are three-fold:

Dataset	Logs	Pers.	Ret.	Gen.	E2E RAG	Click	Human
AOL Search Log	✓	✓	✓	×	×	✓	×
MS MARCO	×	×	✓	×	×	×	✓
LaMP	×	✓	×	✓	×	×	✓
PersonaBench	×	✓	×	✓	×	×	✓
BESPOKE	×	✓	✓	✓	×	×	✓
EnronQA	✓	✓	✓	✓	Partial	×	×
PRBEN (Ours)	✓	✓	✓	✓	✓	✓	✓

Table 1: Coverage of existing benchmarks across seven dimensions: search logs (Logs), personalization (Pers.), retrieval (Ret.), generation (Gen.), end-to-end retrieval-augmented generation (E2E RAG), click signals, and human annotation.

Authentic Resource: We release PRBEN, the first Personalized RAG benchmark derived from real-world search logs, effectively mitigating the biases inherent in synthetic data.

End-to-End Framework: We establish a full-process evaluation pipeline that correlates retrieval personalization with generation quality, supervised by actual user click-through data.

Multi-Dimensional Metrics: We introduce multi-dimensional metrics of accuracy and personalization to assess models' alignment with user intent and behavior, providing a standardized testbed for user-centric AI.

2 The PRBEN Framework

To overcome the limitations of synthetic user simulation and isolated retrieval tasks in existing benchmarks (Park et al., 2025), we present PRBEN, a comprehensive evaluation framework derived from authentic commercial search logs. PRBEN constructs a full-process evaluation framework span-

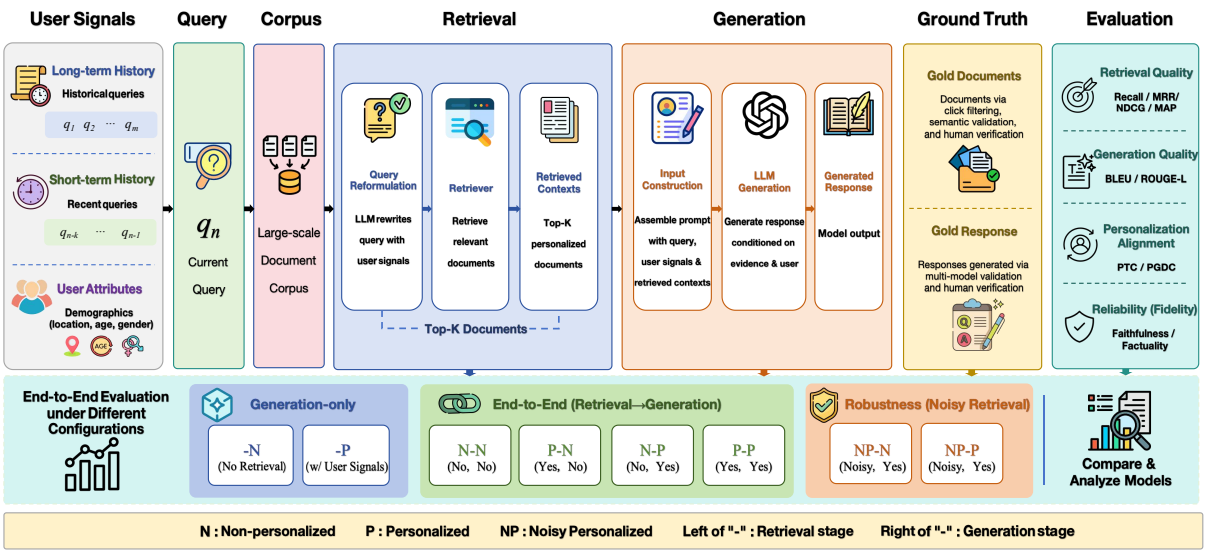

Figure 2: The **PRBEN** pipeline. The benchmark constructs realistic personalized RAG scenarios by injecting user profile information into the retrieval and generation pipeline, establishes a two-stage golden standard driven by real-world user logs, integrates multi-dimensional evaluation metrics, and ultimately forms an end-to-end evaluation framework.

ning document retrieval and augmented generation, and employs multi-dimensional metrics that balance accuracy and personalization (Liu et al., 2024b), enabling systematic assessment of models in practical scenarios (Hagström et al., 2025). The PRBEN framework is illustrated in Figure 2.

2.1 Raw Data Collection

The PRBEN dataset is built from anonymized search logs collected over a three-month period from a large-scale web search engine. Each record includes the user query, clicked URLs, users' historical queries, and profile information. The raw logs are noisy and contain many non-personalized queries, making them unsuitable for direct evaluation. To address this, we design a fine-grained filtering pipeline to construct high-quality samples that better capture users' personalized information needs, enabling realistic personalized retrieval and generation within the PRBEN framework. Details are provided in Appendix A.

2.2 User Profile Construction

To quantitatively evaluate a model's ability to respond to personalized needs, PRBEN constructs a multi-dimensional user profile as the evaluation context, encompassing rich personalized signals (Au et al., 2025; Liang et al., 2025). The profile includes both static attributes and modeled historical behavior patterns to capture the dynamic evolution of user interests and needs, more faithfully reflecting real-world personalized retrieval and generation scenarios (Karatzoglou and Hidasi, 2017; Huang et al., 2023). Specifically, for a given user u, the user profile P_u is defined as:

$$P_u = \{A_u, H_u\} = \Big\{\{\text{gen}_u, \text{age}_u, \text{loc}_u\}, \{q_{u,1}, q_{u,2}, \ldots, q_{u,t-1}\}\Big\} \quad (1)$$

where A_u represents the static attributes of user u, including gender gen_u, age age_u and location loc_u, and H_u represents the sequence of historical behaviors from the first query $q_{u,1}$ to the previous query $q_{u,t-1}$, capturing the dynamic trajectory of user interests.

2.3 Ground-Truth Construction

To objectively quantify the personalized RAG capabilities of models, PRBEN constructs golden-standard responses that explicitly reflect users' personalized needs. Each evaluation instance consists of a golden retrieval set D_u^* and a golden response Y_u^*, which together serve as the reference for preference alignment.

For D_u^*, PRBEN extracts personalized context from real-world search logs. The user's historical queries H_u and static attributes A_u are utilized to construct an augmented personalized golden query $q_{u,t}^*$, which is used to retrieve a candidate document set D_u'. The final golden retrieval set D_u^*

is constructed based on the union of the user's clicked documents D_{clicked} and the retrieved candidate documents D'_u. To further improve document quality and mitigate noise introduced by implicit click signals, both D_{clicked} and D'_u are further verified through a secondary filtering stage using the high-performance model Claude Sonnet 4.6. This process removes documents that are inconsistent with user intent, irrelevant to the personalized query context, or semantically misaligned with the underlying information need, thereby enhancing the personalization validity and semantic reliability of the document set. Furthermore, 5% of the filtered instances are randomly sampled for human verification to ensure the overall quality, reliability, and rationality of the personalized document collection.Details are provided in Appendix C.1.

Finally, D^*_u is defined as the union of the documents actually clicked by the user, D_{clicked}, and the candidate document set D'_u:

$$D^*_u = D_{\text{clicked}} \cup D'_u, \quad D'_u = r(q^*_{u,t}, \mathcal{D}) \tag{2}$$

where $r(\cdot)$ denotes the process for retrieving documents from the corpus $\mathcal{D}$.

During the construction of Y^*_u, personalized information is integrated with the golden retrieval set D^*_u to produce the standard response. Specifically, golden responses are first generated by the high-performance closed-source model GPT-4.1 (Achiam et al., 2023). To further improve factual consistency and personalization reliability, the generated responses are subsequently examined and refined through cross-model verification using Claude Sonnet 4.6, forming a multi-model collaborative validation pipeline. This process verifies whether the responses are fully aligned with user queries, personalized attributes, historical preferences, clicked evidence, and retrieved documents. In addition, 5% of the finalized responses are randomly sampled for human verification to ensure response quality, factual correctness, and personalization consistency. Details are provided in Appendix C.2.This design guides the model to perform fine-grained knowledge and preference selection from the golden documents rather than relying solely on the retriever, enabling the final golden response Y^*_u to more accurately reflect authentic personalized user needs. Overall, the construction process jointly balances factuality, personalization, and semantic consistency, formalized as:

$$Y^*_u = g(q_{u,t}, A_u, H_u, D^*_u) \tag{3}$$

where $g(\cdot)$ represents the generation process for the golden response.

The dual-stage golden standards can be flexibly applied at the process or result level, allowing independent evaluation of retrieval and generation or assessment of the entire pipeline, demonstrating the rigorous value of our benchmark in authenticity, preference alignment, and generative interpretability (Wang et al., 2025b).

2.4 Full-Process Personalized Evaluation

PRBEN evaluates personalized RAG systems from both retrieval and generation perspectives, enabling end-to-end assessment of personalized information seeking. Specifically, we examine whether models can effectively incorporate user-specific information into retrieval and ranking, and generate responses that are both factually grounded and aligned with user preferences.

2.4.1 Personalized Retrieval Evaluation

In this stage, based on user personalized information A_u and H_u, the LLM under evaluation is utilized to perform personalized rewriting on the original query $q_{u,t}$, generating an augmented user query $q'_{u,t}$. The BM25 algorithm (Robertson et al., 2009) is then employed to retrieve documents from the corpus $\mathcal{D}$. Specifically, we compute the BM25 score between $q'_{u,t}$ and each document in $\mathcal{D}$, and then select the Top-K highest-scoring documents as the personalized candidate set D_u, which is evaluated against the final golden retrieval set D^* (Zhang et al., 2019):

$$\begin{aligned} D_u &= \text{Top-K}_{d\in\mathcal{D}} \left(\text{Score}_{\text{BM25}}(q'_{u,t}, d)\right), \\ q'_{u,t} &= l_r(q_{u,t}, A_u, H_u) \end{aligned} \tag{4}$$

where $l_r(\cdot)$ represents the LLM under evaluation for the query rewriting task. This process introduces $q'_{u,t}$ as an evaluative variable and uses retrieval recall as the primary metric to assess $l_r(\cdot)$'s ability to inject personalized user information into retrieval, and to quantify its preference for documents aligned with user interests and historical context during ranking. The retrieved documents are further used to support evaluation of the generation stage (Lewis et al., 2020b).

2.4.2 Personalized Generation Evaluation

The phase is defined as a validation process for the personalized reconstruction and alignment of retrieval results, characterized by significant continuity with the retrieval evaluation phase. We take

Statistic	Value
Sample size	10408
History query length (range)	2–823
History query length (avg.)	211
Gender (Female)	54.80%
Gender (Male)	45.20%
Age range	0–65+
Provinces/regions	32

Table 2: Statistics of the PRBEN benchmark

the user query $q_{u,t}$ and the personalized candidate set D_u as input, combined with user personalized information A_u and H_u, and utilize the LLM under evaluation to generate the response Y_u. This phase is formally defined as:

$$Y_u = l_g(q_{u,t}, D_u, A_u, H_u) \quad (5)$$

where $l_g(\cdot)$ denotes the LLM under evaluation for the response generation task. By comparing Y_u with the golden response Y_u^*, we examine the capabilities of the model $l_g(\cdot)$ from two primary perspectives: **(1) In-depth utilization of candidate evidence**: We assess whether the model effectively grounds its response in D_u by extracting query-relevant facts conditioned on latent user preferences, ensuring factuality and faithfulness (Saad-Falcon et al., 2024; Ming et al., 2024). **(2) Guided utilization of personalized features**: We assess whether A_u and H_u steer the model to reorganize evidence from D_u and generate responses aligned with user preferences, reflected in personalized keyword coverage and semantic consistency (Saad-Falcon et al., 2024; Brown et al., 2025).

3 Experiment

The basic statistics of the proposed PRBEN benchmark are shown in Table 2, which is released under the CC BY-NC-SA 4.0 license.

3.1 Multi-Dimensional Evaluation

We propose a comprehensive evaluation for personalized RAG methods along two dimensions: **Accuracy**, assessing relevance, factuality, and faithfulness, and **Personalization**, evaluating the model's adaptation to user preferences and its ability to generate preference-oriented outputs (Yu et al., 2024; Es et al., 2024).

Accuracy Evaluation Accuracy measures the consistency between retrieved documents and the query, and whether the generated outputs are grounded in evidence and factually correct. This dimension includes three metrics:

(a) Relevance Relevance measures the alignment between the retrieved documents and the query intent. Specifically, we convert both the query and documents into vectors using TF-IDF and compute the cosine similarity between them, where a higher similarity indicates better retrieval quality.

(b) Faithfulness Faithfulness quantifies how well the generated content is supported by retrieved evidence, using NLI-based entailment scores from `microsoft/deberta-xlarge-mnli` (He et al., 2020). The calculation is performed as follows, where d_i denotes the i-th document in the personalized candidate set D_{u} and y_j denotes the j-th sentence in the response Y_u.

$$s(d_i, y_j) = P_{\text{NLI}}(\text{entailment} \mid d_i, y_j),$$
$$\text{Faithfulness} = \frac{1}{|Y_u|} \sum_{y_j \in Y_u} \frac{1}{|D_u|} \sum_{d_i \in D_u} s(d_i, y_j) \quad (6)$$

(c) Factuality Factuality evaluates the objective accuracy of generated content using the large language model GPT-4.1 to automatically verify factual and logical consistency. It outputs a confidence score in $[0, 1]$, with the final score computed as the average of M independent assessments.

$$\text{Factuality} = \frac{1}{M} \sum_{m=1}^{M} \text{LLM}(Y_u, \text{Prompt}_{\text{fact}}) \quad (7)$$

Personalization Evaluation Personalization evaluation measures the system's adaptive capability in learning user-specific preferences during both the retrieval and generation phases. This dimension comprises three metrics:

(a) Retrieval Metrics We use standard ranking metrics such as Recall(Manning, 2008), MRR(Voorhees and Dang, 2001), NDCG(Järvelin and Kekäläinen, 2002), and MAP(Robertson et al., 2009) to evaluate the coverage and ranking of preference-aligned documents, ensuring relevant evidence is effectively retrieved and prioritized for personalized generation.

(b) Similarity Metrics For evaluating generation, we use textual similarity metrics, including BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004), to measure lexical overlap and sequence-level consistency between generated responses and personalized ground-truth responses, capturing precision and semantic completeness.

(c) Personalized Topic/Keyword Coverage (PTC) PTC measures how well generated content reflects user preferences by quantifying the coverage of

Method	Recall@10	Recall@5	MRR@10	MRR@5	NDCG@10	NDCG@5	MAP	Relevance
Non-Personalized Retrieval								
BM25	0.1902	0.1117	0.3746	0.3589	0.2070	0.2335	0.1218	0.0519
Q2D	0.0459	0.0274	0.1273	0.1176	0.0514	0.0591	0.0220	0.0796
BART	0.1837	0.1076	0.3619	0.3462	0.1997	0.2248	0.1171	0.0390
Gemini-2.5	0.1541	0.0929	0.3129	0.3014	0.1691	0.1939	0.0997	0.0833
DeepSeek-R1	0.1726	0.1029	0.3499	0.3344	0.1901	0.2168	0.1123	0.0727
Qwen-2.5-7B-Instruct	0.1547	0.0921	0.3220	0.3076	0.1702	0.1940	0.0980	0.0281
Personalized Retrieval								
• ***Long History***								
Gemini-2.5	0.2455	0.1478	0.4493	0.4358	0.2695	0.3083	0.1807	0.0902
DeepSeek-R1	0.2440	0.1464	0.4440	0.4305	0.2671	0.3048	0.1788	0.0665
Qwen-2.5-7B-Instruct-SFT	0.2472	0.1486	0.4508	0.4375	0.2717	0.3105	0.1843	0.0857
• ***Short History***								
Gemini-2.5	0.2708	0.1623	0.4827	0.4693	0.2964	0.3379	0.2032	0.0827
DeepSeek-R1	0.2638	0.1570	0.4673	0.4543	0.2883	0.3272	0.1986	0.0597
Qwen-2.5-7B-Instruct-SFT	0.2747	0.1658	0.4889	0.4763	0.3023	0.3463	0.2117	0.0778
• ***Long + Short History***								
Gemini-2.5	0.2859	0.1721	0.5037	0.4910	0.3130	0.3575	0.2175	0.0867
DeepSeek-R1	0.2857	0.1715	0.4979	0.4851	0.3125	0.3563	0.2191	0.0585
Qwen-2.5-7B-Instruct-SFT	0.2972	0.1794	0.5227	0.5106	0.3272	0.3748	0.2317	0.0828
• ***User Attributes***								
Gemini-2.5	0.2032	0.1225	0.3929	0.3787	0.2236	0.2563	0.1416	0.0845
DeepSeek-R1	0.1820	0.1087	0.3541	0.3307	0.1998	0.2278	0.1239	0.0529
Qwen-2.5-7B-Instruct-SFT	0.1905	0.1152	0.3697	0.3562	0.2099	0.2411	0.1322	0.0815

Table 3: **Retrieval results on the PRBEN dataset.** Personalized methods incorporating user history consistently outperform the original query baselines, with short-term history performing best, long-term history second, and their combination achieving the highest overall results. Incorporating user attribute information may introduce noise and reduce performance.

personalized keywords. Let T_u^{gold} be the user's golden preference keywords and T_u the keywords in the generated response y_u. The PTC score is defined as:

$$\text{PTC} = \frac{|T_u \cap T_u^{\text{gold}}|}{|T_u^{\text{gold}}|} \tag{8}$$

(d) Personalized Generation Direction Consistency (PGDC) PGDC measures the consistency between the generated response Y_u and the personalized golden response Y_u^* in terms of intent and preference orientation. An LLM assigns a discrete consistency score in $\{1, \ldots, 5\}$, which is linearly normalized to $[0, 1]$:

$$\text{PGDC} = \frac{1}{5}\,\text{LLM}(Y_u, Y_u^*) \tag{9}$$

where $\text{LLM}(\cdot) \in \{1, 2, 3, 4, 5\}$.

3.2 Retrieval Baselines and Results

In this study, we conduct a comparative analysis of various retrieval methods, which are categorized into two primary groups: non-personalized methods and personalized methods. The non-personalized baselines include Q2D (Wang et al., 2023), BART (Lewis et al., 2020a), Gemini-2.5 (Comanici et al., 2025) DeepSeek-R1 (Liu et al., 2024a), and Qwen-2.5-7B-Instruct (Team et al., 2024). For personalized methods, we systematically evaluate the impact of different personalized signals on query augmentation, including user historical queries, user attributes, and their combinations. Here, H_u represents two types of historical behaviors: the short-term history consists of the five queries closest in time to the current query, while the long-term history selects the five queries most semantically similar to the current query across the entire history, capturing the complete trajectory of user interests.Across both closed-source models Gemini (Comanici et al., 2025) and open-source models Qwen (Team et al., 2024) and DeepSeek (Guo et al., 2025).

The results summarized in Table 3 indicate that personalized methods incorporating user history generally outperform the original query baselines across all evaluation metrics, demonstrating that our PRBEN exhibits significant personalization effects. Among the various personalized signals, short-term query history shows the strongest performance, followed by long-term query history, while their combination achieves the best overall results, highlighting the effectiveness of integrating multiple forms of personalized information. Notably, incorporating user attribute information may even degrade overall performance. This is because, in real-world data distributions, only a small propor-

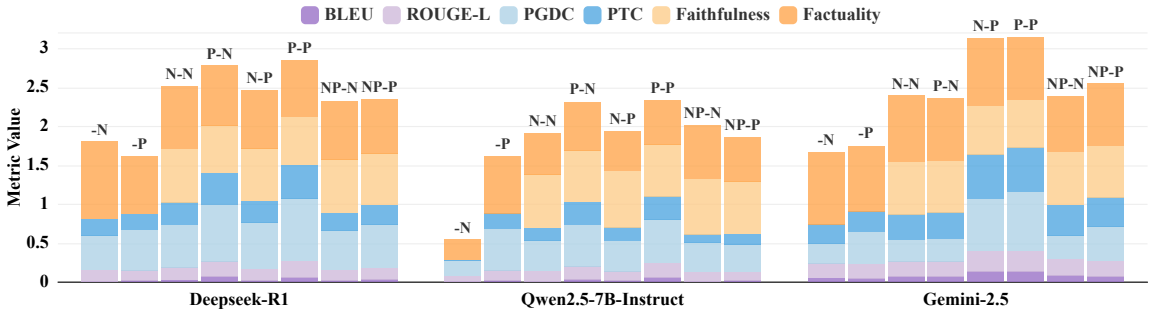

Figure 3: **Evaluation results on the PRBEN dataset.** PRBEN successfully differentiates between models, showing that personalization metrics steadily improve as personalized information is progressively introduced, with retrieval-stage personalization being particularly critical, while blindly injecting user history may introduce noise.

tion of samples can benefit from user attributes, and directly using static attributes introduces excessive noise, leading to overgeneralization. Improvement is observed only for specific attribute-based samples. These results indicate that PERBEN fundamentally differs from traditional query-correction (Su et al., 2025) or simple query rewriting datasets, constituting a challenging preference-driven personalization benchmark.

3.3 End-to-End Baselines and Results

To evaluate the full personalized RAG pipeline, we design three categories of evaluation configurations on PRBEN: generation-only settings (-N, -P), end-to-end settings (N-N, P-N, N-P, P-P), and robustness settings (NP-N, NP-P). These configurations are designed to assess, respectively, (i) the ability of the generation model to exploit personalized signals, (ii) end-to-end personalization across retrieval and generation stages, and (iii) model robustness under noisy personalized retrieval. In our notation, the hyphen (-) separates the retrieval and generation stages. Symbols N, P, and NP indicate the absence of personalization, the presence of personalization, and noisy personalization signals, respectively.

We perform evaluations using Gemini, Qwen, and DeepSeek in three dimensions: overall generation quality (BLEU, ROUGE-L), personalization alignment (PGDC, PTC), and reliability (Faithfulness, Factuality). As shown in Figure 3, significant performance differences are observed across different configurations. As personalized information is progressively introduced at each stage, personalization-related metrics exhibit a consistent improvement trend, with the P-P setting achieving the best overall performance. In contrast, reliability metrics show more pronounced variations across models. Under robustness settings that more closely reflect real-world scenarios, model performance remains lower than that of the P-P configuration, indicating that effectively leveraging personalized information throughout the full personalized RAG pipeline remains a challenging task. We also found that personalization at the retrieval stage is particularly crucial, and it even exerts a significant impact on personalization at the generation stage. Overall, Gemini demonstrates superior performance compared to Qwen and DeepSeek. These results indicate that PRBEN can provide a fine-grained and reliable characterization of the impact of personalization strategies at different stages.

3.4 Human Verification and Validation

To verify the authenticity of the benchmark and the effectiveness of the metrics, we randomly sampled 5% of the instances from the test set for human evaluation. Three annotators conducted a double-blind evaluation, scoring the generated results on a Likert scale (1-5) in three dimensions: textual similarity, satisfaction of personalized preferences, and consistency of generation direction.Details are provided in Appendix D.

The experimental results are shown in Figure 4. The metric evaluations exhibit a high degree of consistency with human ratings. Specifically, the Spearman correlation (ρ) between ROUGE-L, PGDC, Faithfulness and human ratings all exceed 0.7. This further demonstrates that the benchmark can effectively reflect real-world human preferences for personalized retrieval and generation.

3.5 Case Study

We take the query "Oracle" from PRBEN as a case study. The user's historical queries include NVIDIA, Xiaomi, Kingsoft Cloud, and GPU rankings, with the clicked result title being "Oracle |

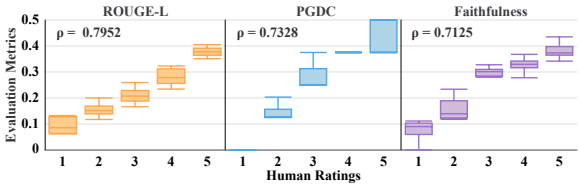

Figure 4: Human verification of **PRBEN**: 5% of test instances were double-blind rated, with Spearman correlations for ROUGE-L, PGDC, and Faithfulness all above 0.7, confirming the benchmark reflects real-world personalized preferences.

Cloud Applications and Cloud Platform." Based on this clicked title, the golden documents and their corresponding gold responses all focus on "Oracle Corporation."

Taking the DeepSeek-R1 model as an example, the experimental results show that: in the non-personalized (N-N) setting, retrieval returns 3/5 documents related to "ancient bone scripts" and 2/5 related to "Oracle Corporation," leading to a final generated response about ancient bone scripts. Conversely, under the fully personalized configuration (P-P), all retrieved documents are related to "Oracle Corporation," and the generated response is closer to the golden responses in the benchmark. However, in the more realistic noisy setting (NP-P), the generated response deviates from "Oracle Corporation." These results not only validate the effectiveness of our benchmark but also demonstrate that personalization remains a challenging task for existing LLMs.

4 Related Work

4.1 Personalized Retrieval and Generation

Personalized Information Retrieval enhances retrieval relevance by leveraging users' historical behaviors, preference patterns, and contextual signals. Early methods mainly relied on query history or static user attributes to model interests and re-rank results (Zhou et al., 2024; Ge et al., 2018; Zhou et al., 2020). Recent deep learning approaches adopt sequence modeling and contextual representations to better capture long-term interests and short-term dynamic preferences (Liu et al., 2025). However, many existing datasets are either manually constructed or derived from domain-specific logs (Kim et al., 2025; Salemi et al., 2024; Rekabsaz et al., 2021), limiting their ability to reflect the complexity and diversity of real-world open-domain user behaviors.

Retrieval-Augmented Generation (RAG) further integrates external document retrieval with generative models to improve factuality and coverage in question answering and text generation (Lewis et al., 2020b; Guu et al., 2020). In personalized settings, some studies incorporate user preferences into retrieval or generation modules to adapt outputs to individual needs (Li et al., 2025a; Salemi et al., 2024). However, existing works mainly focus on generation quality and textual relevance, while overlooking systematic modeling of personalized behaviors, ambiguous intents, and end-to-end decision chains from retrieval to generation.

4.2 Personalized RAG Benchmark

Robust benchmarks are essential for advancing personalized RAG research. Traditional evaluation primarily relies on standard retrieval metrics (e.g., NDCG, MAP) and human-annotated generation quality scores (Nguyen et al., 2016; Karpukhin et al., 2020). More recent studies leverage LLM-based automatic evaluation frameworks (Es et al., 2024; Saad-Falcon et al., 2024).

Despite these efforts, existing benchmarks remain limited in process-level assessment, chained generation consistency, and real user feedback coverage. Most datasets are domain-specific or synthetically constructed, introducing bias and failing to capture the complexity and hierarchical nature of authentic user behaviors (Wang et al., 2025a).

5 Conclusion

In this study, we focused on personalized RAG in real-world user scenarios and introduced **PRBEN**, a comprehensive benchmark for systematically evaluating personalized RAG methods. PRBEN addresses two major limitations of prior work: the scarcity of authentic user preference signals and the lack of standardized evaluation frameworks. Leveraging multi-level user features from large-scale search logs, including query histories, demographics, and click feedback, we design a multi-dimensional evaluation pipeline covering retrieval quality, generation faithfulness, and personalization effectiveness. Experimental and human evaluation results demonstrate that PRBEN effectively differentiates methods in capturing and utilizing user preferences, providing a reliable foundation for future personalized RAG research.

6 Limitations

Although PRBEN is constructed from real-world search logs and covers the full pipeline of personalized search from retrieval to generation, thereby reflecting realistic user behavioral signals in practical systems, it still has several limitations. First, the dataset is primarily derived from a single commercial search platform, which may lead to differences in user behavior distributions and interaction patterns across different search engines or application scenarios, leaving its cross-platform generalization capability to be further validated. Second, PRBEN is mainly built from static historical logs, focusing on modeling observed user behaviors, while the long-term and continuously evolving nature of user interests is not explicitly captured.

In addition, the current data modality is primarily text-based search logs, without systematically incorporating multilingual queries or multimodal interaction signals (e.g., images and videos). Therefore, its applicability to more complex cross-lingual and multimodal personalized search scenarios remains an open direction for future extension. Finally, the benchmark mainly focuses on personalized modeling in standard text-based search settings, and its performance in more complex interactive and online decision-making environments remains to be further explored.

7 Ethical Considerations

The dataset is derived from anonymized search logs. To protect user privacy, we applied k-anonymity filtering (k=5) and masked all personally identifiable information (PII) using named entity recognition (NER). All users consented to research use of their data via the service's Terms of Service. The dataset is released under a strict license for research purposes only, prohibiting any attempts to re-identify users.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Steven Au, Cameron J Dimacali, Ojasmitha Pedirappagari, Namyong Park, Franck Dernoncourt, Yu Wang, Nikos Kanakaris, Hanieh Deilamsalehy, Ryan A Rossi, and Nesreen K Ahmed. 2025. Personalized graph-based retrieval for large language models. *arXiv preprint arXiv:2501.02157*.

Andrew Brown, Muhammad Roman, and Barry Devereux. 2025. A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *arXiv preprint arXiv:2508.06401*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Songwei Ge, Zhicheng Dou, Zhengbao Jiang, Jian-Yun Nie, and Ji-Rong Wen. 2018. Personalizing search results using hierarchical rnn with query-aware attention. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 347–356.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Qian Guo, Wei Chen, and Huaiyu Wan. 2021. Aol4ps: A large-scale data set for personalized search. *Data Intelligence*, 3(4):548–567.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-wei Chang. 2020. Realm: Retrieval-augmented language model pre. *Training*.

Lovisa Hagström, Sara Vera Marjanovic, Haeun Yu, Arnav Arora, Christina Lioma, Maria Maistro, Pepa Atanasova, and Isabelle Augenstein. 2025. A reality check on context utilisation for retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19691–19730.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Qiushi Huang, Shuai Fu, Xubo Liu, Wenwu Wang, Tom Ko, Yu Zhang, and Lilian Tang. 2023. Learning retrieval augmentation for personalized dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2523–2540.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Alexandros Karatzoglou and Balázs Hidasi. 2017. Deep learning for recommender systems. In *Proceedings of the eleventh ACM conference on recommender systems*, pages 396–397.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Hyunseo Kim, Sangam Lee, Kwangwook Seo, and Dongha Lee. 2025. Bespoke: Benchmark for search-augmented large language model personalization via diagnostic feedback. *arXiv preprint arXiv:2509.21106*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Xiaopeng Li, Pengyue Jia, Derong Xu, Yi Wen, Yingyi Zhang, Wenlin Zhang, Wanyu Wang, Yichao Wang, Zhaocheng Du, Xiangyang Li, and 1 others. 2025a. A survey of personalization: From rag to agent. *arXiv preprint arXiv:2504.10147*.

Xiaopeng Li, Lixin Su, Pengyue Jia, Suqi Cheng, Junfeng Wang, Dawei Yin, and Xiangyu Zhao. 2025b. Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llms. *ACM Transactions on Information Systems*, 43(6):1–33.

Siqi Liang, Yudi Zhang, and Yue Guo. 2025. Personaagent with graphrag: Community-aware knowledge graphs for personalized llm. *arXiv preprint arXiv:2511.17467*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jintao Liu, Ruixue Ding, Linhao Zhang, Pengjun Xie, and Fie Huang. 2024b. Cofe-rag: A comprehensive full-chain evaluation framework for retrieval-augmented generation with enhanced data diversity. *arXiv preprint arXiv:2410.12248*.

Jiongnan Liu, Zhicheng Dou, Jian-Yun Nie, and Ji-Rong Wen. 2023. Integrated personalized and diversified search based on search logs. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):694–707.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025. Llms+ persona-plug= personalized llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9373–9385.

Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if" the moon is made of marshmallows". *arXiv preprint arXiv:2410.03727*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Chanhee Park, Hyeonseok Moon, Chanjun Park, and Heui-Seok Lim. 2025. Mirage: A metric-intensive benchmark for retrieval-augmented generation evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2883–2900.

Navid Rekabsaz, Oleg Lesota, Markus Schedl, Jon Brassey, and Carsten Eickhoff. 2021. Tripclick: the log files of a large health web search engine. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2507–2513.

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

Michael J Ryan, Danmei Xu, Chris Nivera, and Daniel Campos. 2025. Enronqa: Towards personalized rag over private documents. *arXiv preprint arXiv:2505.00263*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 338–354.

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.

Jinbo Su, Lingzhe Gao, Wei Li, Shihao Liu, Haojie Lei, Xinyi Wang, Yuanzhao Guo, Ke Wang, Daiting Shi, and Dawei Yin. 2025. Racqc: Advanced retrieval-augmented generation for chinese query correction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 675–689.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh RN, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, and 1 others. 2025. Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 878–893.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Ellen M Voorhees and Hoa Trang Dang. 2001. Overview of the trec 2001 question answering track. In *Trec*, pages 42–51.

Kevin Shukang Wang, Karel Joshua Harjono, and Ramon Lawrence. 2025a. Towards reliable agents: Benchmarking customized llm-based retrieval-augmented generation frameworks with deployment validation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 655–661.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2025b. Omnieval: An omnidirectional and automatic rag evaluation benchmark in financial domain. In *Proceedings of the 2025 conference on empirical methods in natural language processing*, pages 5737–5762.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, pages 102–120. Springer.

Saber Zerhoudi and Michael Granitzer. 2024. Personarag: Enhancing retrieval-augmented generation systems with user-centric agents. *arXiv preprint arXiv:2407.09394*.

Kepu Zhang, Teng Shi, Weijie Yu, and Jun Xu. 2025a. 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.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025b. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. *arXiv preprint arXiv:2506.08938*.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38.

Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2020. Encoding history with context-aware representation learning for personalized search. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1111–1120.

Yujia Zhou, Qiannan Zhu, Jiajie Jin, and Zhicheng Dou. 2024. Cognitive personalized search integrating large language models with an efficient memory mechanism. In *Proceedings of the ACM Web Conference 2024*, pages 1464–1473.

A Appendix: Data Construction and Filtering Details

A.1 Data Source

The dataset is collected from real-world user interaction logs of a large-scale commercial search engine, aiming to capture the distribution of real-world search behavior and user information needs. The data includes multiple types of behavioral signals, such as user queries, user profiles, historical interactions, and click records.

A.2 Raw Data Characteristics

The raw data exhibits typical characteristics of real-world logs, including:

- **Incomplete information**: user profiles and historical behaviors are partially missing or sparsely observed;
- **Sparse personalization signals**: many queries lack sufficient contextual information for reliable intent modeling;
- **Broken click chains**: some clicked links are expired or no longer accessible;
- **Noisy interactions**: logs contain duplicated, irrelevant, or low-quality behavioral records.

A.3 Data Filtering Pipeline

We adopt a coarse-to-fine multi-stage filtering strategy to progressively clean the raw data.

Completeness Filtering We remove instances with missing user profiles and historical behavior signals, ensuring basic feasibility for intent modeling.

Personalization Filtering We filter out generic or weakly personalized queries that fail to reflect user-specific preferences.

Click Consistency Filtering We remove samples with invalid click signals, including expired links and clicks that cannot be aligned with the original query semantics.

Quality Filtering We further eliminate duplicated, noisy, or semantically ambiguous samples.

A.4 Safety Filtering

We additionally filter out samples containing inappropriate content, including but not limited to violent, pornographic, hateful, or illegal material. This ensures that the dataset complies with ethical and safety standards.

A.5 Privacy and Anonymization

All user-related information is strictly anonymized. Any identifiers that could directly or indirectly reveal user identity are removed or obfuscated. The dataset does not contain any personally identifiable information (PII) and is used solely for research purposes.

A.6 LLM-based Iterative Filtering

In addition to rule-based filtering, we introduce a large language model as an auxiliary multi-round refinement module.

At each iteration, the model assesses samples from multiple dimensions, including semantic clarity, personalization consistency, and overall data quality. Low-quality samples are filtered out, while borderline cases are reconsidered in subsequent rounds.

This iterative process follows a coarse-to-fine strategy, progressively improving dataset quality. Importantly, the large language model is used solely for filtering and is not involved in data generation or annotation, thereby minimizing potential bias and information leakage. leakage.

A.7 Final Dataset Construction

After applying the above filtering stages, we construct the final dataset used in all experiments. Each instance contains structured fields such as user query, anonymized user profile, historical interactions, and task-specific response or retrieval information.

The final dataset achieves a balance between real-world complexity and data quality, providing a reliable foundation for evaluating personalized retrieval and generation systems.

B Appendix: Prompt Templates Used in Our Framework

B.1 Query Reformulation

We design a prompt for personalized query reformulation based on multi-source user behavioral signals, including recent search history, long-term search history, and user profile information. By integrating these heterogeneous signals, our approach captures both dynamic shifts in user interests and stable long-term preferences, enabling more accurate disambiguation and intent understanding in query reformulation.

Personalized Query Reformulation Prompt

You are a professional personalized search query rewriting assistant.
Given a user query and associated user behavioral signals, rewrite the query to better reflect the user's underlying intent.
User recent search history: {recent}
User long-term search history: {long}
User profile information: {profile}
Current query: {query}
Generate a single rewritten query that is semantically complete, natural, and aligned with the user's personalized intent.
Output only the rewritten query.
Example
User recent search history: Volkswagen, Tesla, Hongqi
User long-term search history: car reviews, EV comparisons
User profile information: user interested in automotive content
Current query: Xiaomi
Output: Xiaomi car

B.2 Personalized RAG Generation

We design a prompt for personalized retrieval-augmented generation, which leverages user behavioral signals to guide answer generation while strictly grounding responses in retrieved documents. Our approach integrates multi-granular user signals, including recent search history, long-term search history, and user profile information, to jointly model user intent and preferences. This enables the generation process to produce responses that are both factually grounded and personalized.

Personalized RAG Assistant Prompt

You are a personalized retrieval-augmented generation assistant.
Your task is to generate an answer to the user's question based strictly on the provided retrieved documents. You may optionally use user behavioral signals (recent search history, long-term search history, and user profile information) to better align the response with the user's intent and preferences.
You must not fabricate or introduce any information that is not supported by the retrieved documents. Personalization should only affect the focus and presentation of the response.
User Query: {query}
Retrieved Documents (Top-5): {docs}
User Recent Search History: {recent}
User Long-term Search History: {long}
User Profile Information: {profile}
If the retrieved documents are insufficient to answer the question, respond with: "No relevant information found in the retrieved documents."
Output a single, coherent, and informative answer.

C Appendix: Ground Truth Verification

C.1 Retrieval

To ensure dataset quality, we adopt a hybrid evaluation protocol combining LLM-based verification and human annotation.

First, we use Claude Sonnet 4.6 to perform semantic consistency scoring based on relevance and personalization signals, retaining instances with scores ≥ 3.

We then randomly sample 5% of the filtered data for human evaluation. Three annotators independently assign 5-point consistency scores (1: inconsistent, 5: highly consistent), with instances scoring ≥ 3 considered valid. The human evaluation yields a validity rate of 90.63%.

This process ensures the reliability and quality of the constructed dataset.

The full prompt template is presented as follows:

Semantic Consistency Verification Prompt

Task: Judge whether a candidate document is consistent with the user's intent in Personalized RAG.
Inputs:

- Personal context: user profile, search history, clicked documents
- Query: user query
- Document: candidate document

Score (1–5):
5: Highly consistent
4: Mostly consistent
3: Partially consistent
2: Weakly consistent
1: Inconsistent
Focus: Query relevance; Personalization consistency; Alignment with clicked documents; Retrieval context consistency.
Output:

```
{
  "consistency_score": 1-5,
  "reason": "brief explanation"
}
```

C.2 Generation

We adopt a two-stage filtering process to ensure data quality. First, we use Claude Sonnet 4.6 to refine and verify each generated answer, followed by automatic scoring across five dimensions: faithfulness, personalization, completeness, fluency, and hallucination risk. We retain instances with scores ≥ 3 for faithfulness, personalization, completeness, and fluency, and ≤ 3 for hallucination risk.

We then randomly sample 5% of the filtered data for human evaluation to validate the reliability of the LLM-based filtering. Three annotators independently assess each instance using the same 5-point scale, and an instance is considered valid only if all three annotators judge it as acceptable. The human evaluation yields a validity rate of 93.75%.

The full prompt template is presented as follows:

Personalized Answer Refinement Prompt

You are a Personalized RAG model for evaluating and improving answers.
Given an answer, refine it using retrieved documents and user context.
Tasks: faithfulness; hallucination check; personalization; completeness; fluency
Constraints: no external knowledge; no generation from scratch; minimal edits if high quality
Inputs:

- Query
- Retrieved documents
- User profile
- Search history
- Current answer

Output:

```
{
  "faithfulness": 1-5,
  "personalization": 1-5,
  "completeness": 1-5,
  "fluency": 1-5,
  "hallucination_risk": 1-5,
  "issues": [],
  "final_answer": ""
}
```

D Appendix: Human Evaluation Procedure

This section describes the detailed human evaluation procedure for reproducibility.

Step 1: Sampling. We randomly sample 5% of instances from the test set to construct the evaluation subset.

Step 2: Annotation Setup. Three independent annotators are recruited to perform a double-blind evaluation. Annotators are not exposed to system identity, model source, or ground-truth labels.

Step 3: Evaluation Interface. Each instance contains the user query, retrieved context (if applicable), and a model-generated response. Annotators are asked to evaluate each response independently.

Step 4: Scoring Criteria. Each response is rated on a 5-point Likert scale (1–5) along three dimensions:

- Textual Similarity: semantic alignment with reference content.
- Personalized Preference Satisfaction: degree of alignment with user intent and preferences.
- Generation Consistency: coherence and relevance of the generated response.

Step 5: Aggregation. Final human scores are obtained by averaging annotations across three annotators.

Step 6: Metric Correlation. We compute Spearman correlation (ρ) between automatic metrics (ROUGE-L, PGDC, and Faithfulness) and human judgments.