Exploring and Improving Cross- and Multi-Domain Personalized Question Answering via a Dual Retrieval Augmentation Approach

Ozel Yilmazel
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
Amherst, MA, USA
oyilmazel@umass.edu

Hamed Zamani
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
Amherst, MA, USA
zamani@cs.umass.edu

James Allan
Center for Intelligent Information Retrieval
University of Massachusetts Amherst
Amherst, MA, USA
allan@cs.umass.edu

Abstract

Personalized long-form question answering aims to tailor responses based on a user's historical data, which often spans diverse information domains. While prior work typically assumes user profiles are domain-aligned with the users' questions, real-world profiles are often diverse, containing information across multiple domains. In this work, we use the LaMP-QA benchmark, where user profiles span diverse Stack Exchange domains (Arts & Entertainment, Lifestyle & Personal Development, and Society & Culture), to systematically investigate how domain composition of the user profile affects personalization performance. We examine three profile composition settings: *Cross-Domain*, where profiles contain only out-of-domain information; *Multi-Domain*, where profiles span multiple domains including the question's domain; and *In-Domain*, where profiles match the question domain. Our analysis reveals that state-of-the-art personalization methods struggle in *Cross-Domain* settings, often failing to outperform non-personalized baselines. To address this challenge, we introduce Dual Retrieval Augmentation, a framework that jointly retrieves from both the user profile and a public corpus. We apply our framework to existing personalization approaches and demonstrate that it consistently outperforms state-of-the-art methods across all three profile settings, validating its robustness for personalized question answering.

CCS Concepts

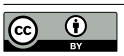
• **Information systems** → **Personalization**; **Question answering**; *Retrieval models and ranking*; • **Computing methodologies** → *Natural language generation.*

Keywords

Personalized Question Answering; Retrieval Augmented Generation; Long-form Text Generation

ACM Reference Format:

Ozel Yilmazel, Hamed Zamani, and James Allan. 2026. Exploring and Improving Cross- and Multi-Domain Personalized Question Answering via a Dual Retrieval Augmentation Approach. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3805712.3809932



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809932

1 Introduction

Personalization is a fundamental component of modern information systems, playing a key role in search [19], recommendation [10], and text generation [15, 18]. By tailoring outputs to a user's intent, background, and preferences, personalized systems can improve user satisfaction and trust. With the rise of Large Language Models (LLMs) that interact with users through natural language, *personalized text generation* has become an increasingly important area of study [15].

Recent efforts have introduced benchmarks for personalized generation, progressing from content-focused tasks such as personalized tweet paraphrasing in LaMP [15] or personalized email generation in LongLaMP [7] to the more recent LaMP-QA benchmark [16] that focuses on personalized question answering (QA). Additionally, a variety of frameworks have been proposed to personalize LLMs, ranging from Retrieval Augmented Generation (RAG) using user information [15] to training with personalized human feedback [13], optimizing LLMs with user-specific supervision [5], creating personalized prompts [8], and parameter-efficient fine-tuning solutions per user [17].

A common assumption in personalized question answering is that a user's historical interactions are topically aligned with their current query. Under this assumption, retrieving relevant items from the user profile provides useful context for generating a personalized response. In practice, however, users often have diverse interests that span multiple topics and domains. A user who asks a question about music production may have a history that includes questions about cooking, finance, and philosophy. This mismatch between the topical scope of the profile and the domain of the current question raises an important but underexplored research question: ***How does the domain composition of a user's profile affect personalization performance?***

Our contributions can be summarized as follows: **(1)** We present the first systematic study of how user profile domain composition affects personalized long-form QA. **(2)** We show that existing personalization approaches degrade substantially in certain profile composition settings, often underperforming non-personalized baselines. **(3)** We propose Dual Retrieval Augmentation, a general framework that jointly retrieves from the user profile and a public corpus containing target-domain content. We demonstrate the generality of this framework by applying it to both RAG-based personalization [15] and planning-based approaches [16], achieving statistically significant improvements over state-of-the-art baselines across all three profile composition settings.

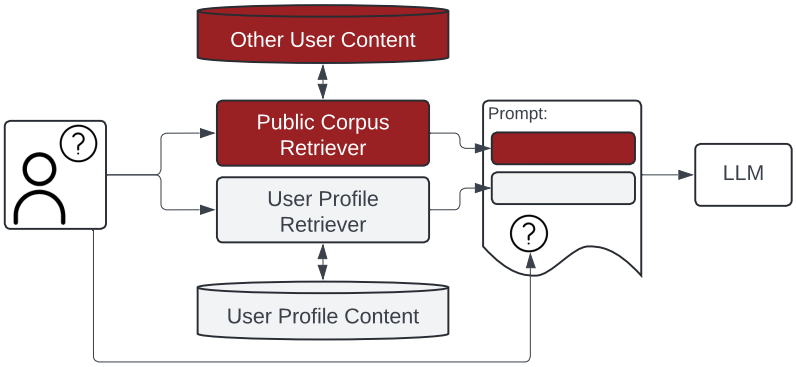


Figure 1: General retrieval architecture comparison. Standard RAG (gray) retrieves only from the user profile. Dual Retrieval Augmentation framework (maroon + gray) performs personalized retrieval and public corpus retrieval.

To improve reproducibility and foster research in this area, we open-source our implementation and publicly release our experiment results.[1]

2 Methodology

We first formalize the problem of personalized QA across different domain composition settings (§2.1), then introduce our core idea (§2.2), and show how it can be applied to two existing personalization approaches (§2.3–§2.4).

2.1 Problem Formulation

We consider a scenario where a user u poses a question x_u. Following prior work that incorporates long-term user history as a user profile [6, 15], we assume the user profile $P_u = \{p_i\}_{i=1}^{n_u}$ consists of n_u previously asked questions along with the detailed narrative of the question written by the user. Each profile entry p_i is associated with a domain label $\delta(p_i) \in \mathcal{D}$, where $\mathcal{D}$ represents the set of all domains. Similarly, the user's current question x_u belongs to a target domain $\delta(x_u)$. We define the domain-specific subset of a user profile as $P_u^{(t)} = \{p \in P_u : \delta(p) = t\}$. Lastly, we define C to be a public corpus that contains posts from other users in the target domain ($\delta(x_u)$). We investigate three user profile settings based on the relationship between the profile domains and the question domain:

Cross-Domain. The user profile contains entries only from domains other than the question domain: $P_u \cap P_u^{(\delta(x_u))} = \emptyset$.

Multi-Domain. The user profile contains all available user entries, without domain filtering.

In-Domain. The user profile contains entries exclusively from the question domain: $P_u = P_u^{(\delta(x_u))}$.

The goal of personalized long-form question answering is to generate a personalized response $y_u = f(x_u, P_u, C)$.

2.2 Dual Retrieval Augmentation

Recent work has shown that RAG provides a strong framework for personalizing LLMs [7, 15, 16]. These methods retrieve relevant entries from the user profile P_u and condition the generator on them. However, our analysis reveals that this approach suffers substantial performance degradation in cross-domain and some multi-domain settings, often underperforming non-personalized baselines (§3.4). We attribute this to a *domain gap*: when the user profile lacks content from the target question domain, retrieved out-of-domain entries provide insufficient and potentially misleading context.

To address this, we propose Dual Retrieval Augmentation, which supplements the user profile retrieval with a second retrieval from a public corpus C containing content in the target domain (Figure 1). The two retrieval sources play complementary roles: the user profile captures individual preferences and communication style, while the public corpus supplies domain-specific language and concepts that bridge the domain gap. Together, they allow the generator to produce responses that are both personalized and grounded in the target domain.

Dual Retrieval Augmentation is a general principle, not a specific model. In the remainder of this section, we apply it to two existing personalization approaches.

2.3 Dual Retrieval Augmented Generation (D-RAG)

We denote a language model M and a retrieval function $R(q, S, k)$ that returns k entries from corpus S given query q. Standard RAG-based personalization generates a response as $\hat{y}_u = M(x_u, R(x_u, P_u, k))$ [15]. D-RAG extends this by additionally conditioning on public retrieval:

$$\hat{y}_u = M\big(x_u,\ R(x_u, P_u, k),\ R(x_u, C, k)\big) \tag{1}$$

2.4 Planned Dual Retrieval Augmented Generation (Plan-D-RAG)

Recent work has shown that planning-based generation, in which a model first predicts the key aspects a response should address, improves long-form personalized QA [16]. We apply Dual Retrieval Augmentation to this approach, yielding Plan-D-RAG.

Plan-D-RAG uses a fine-tuned planner M_{plan} that produces a structured aspect list before the generator writes the final answer. Critically, both the planner and the generator are conditioned on dual retrieval context, enabling them to synthesize personalized

[1] Visit https://github.com/oz03-hub/Dual-Retrieval-Augmentation

preferences with target-domain content during training and inference. Generation proceeds in two stages:

$$plan = M_{\text{plan}}\big(x_u,\ R(x_u, P_u, k),\ R(x_u, C, k)\big) \tag{2}$$

$$\hat{y}_u = M\big(x_u,\ R(x_u, P_u, k),\ R(x_u, C, k),\ plan\big) \tag{3}$$

The planner is trained following the framework introduced in LaMP-QA [16], we refer readers there for full details on planner training and aspect prediction.

3 Experiments

3.1 Datasets and Evaluation Metric

We conduct our experiments on the LaMP-QA benchmark [16], a recent personalized long-form QA benchmark consisting of user profiles from the Stack Exchange platform. Each user profile contains questions from three broad Stack Exchange categories: Arts and Entertainment (A&E), Lifestyle and Personal Development (L&P), and Society and Culture (S&C). This multi-domain structure makes LaMP-QA particularly well suited for our study, as it enables systematic investigation of in-, cross-, and multi-domain personalization settings. For evaluating personalization performance, we follow the aspect-based evaluation framework from LaMP-QA [9, 14, 16], which extracts personalization aspects from question narratives and grades generated answers on each aspect using an LLM evaluator (Qwen 2.5 32B [12]). Note that the narratives and extracted aspects are hidden from generator models during inference to ensure unbiased evaluation. Specifically, each aspect is scored on a scale from 0 to 2 and then normalized by dividing by 2. The final evaluation score for a generated response is the average of the normalized aspect scores. We refer readers to LaMP-QA [16] for additional details on the prompts used in evaluation.

Public Corpus Construction. Our framework requires a public corpus C that excludes the current user's content and always contains target-domain information. We construct C by pooling all target domain profile entries from other users in the dataset split. However, LaMP-QA contains duplicate user profiles where the same real-world user appears with different user IDs. To prevent leakage, we implement checksum-based filtering: for each user u, we compute checksums for all posts and exclude any profile sharing matching checksums with u during retrieval, this ensures any retrieved content from C is not from the current user.

3.2 Experimental Setup

We use Qwen2.5 (3B, 7B, 14B) [12] and GPT-4o-mini [11] as the generator LLM M. For the planner model M_{plan}, we fine-tune Qwen2.5-7B using LoRA [3] with a sequence-to-sequence cross-entropy objective. All models use a maximum context length of 8192 tokens and generate with temperature 0 [2]. For retrieval, we use Contriever [4] fine-tuned on MS MARCO [1] to retrieve $k = 10$ entries from both the user profile and public corpus.

3.3 Baselines

We compare against four baselines that represent a progression from no personalization to planned personalization. All baselines except for Public-Only are drawn from the publicly available LaMP-QA Benchmark [16].

No-Personalization. The question x_u is provided directly to the LLM without additional context: $\hat{y}_u = M(x_u)$.

Public-Only. The top k retrieved posts from the public corpus C and provided as context: $\hat{y}_u = M(x_u, R(x_u, C, k))$. This represents standard RAG without personalization.

RAG. The top-k posts are retrieved from the user profile P_u: $\hat{y}_u = M(x_u, R(x_u, P_u, k))$. This is the single-source counterpart (gray segments in Figure 1) of D-RAG.

Plan-RAG. A fine-tuned planner M_{plan} generates aspects based on the question and retrieved profile content, which then guide response generation: $\hat{y}_u = M(x_u, R(x_u, P_u, k), M_{\text{plan}}(x_u, R(x_u, P_u, k)))$. This is the single-source counterpart to Plan-D-RAG and is the current state-of-the-art method for personalization.

3.4 Empirical Findings

Table 1 presents results across the three profile settings, revealing several consistent patterns. First, Plan-D-RAG achieves the best average performance in the majority of experimental configurations, demonstrating the effectiveness of combining dual retrieval with planning. Second, methods incorporating user profile retrieval substantially outperform public-only retrieval in nearly all cases, highlighting the critical importance of user-specific context for personalization. Third, D-RAG and Plan-D-RAG methods outperform their single retrieval baselines in the majority of configurations and consistently improves with model capacity. Larger models appear better equipped to synthesize information from multiple retrieval sources without distraction from potentially irrelevant content.

Cross-Domain Results. The Cross-Domain setting represents a challenging scenario, where user profiles contain no information from the target question domain. RAG and Plan-RAG perform poorly, consistently falling below the No-Personalization baseline across all model sizes, indicating that retrieving from misaligned profile content actively harms generation quality. Our analysis reveals that dual augmentation with public data in the target domain is essential across all domains: when retrieved profile content is misaligned with the user's question, it guides the generator LLM to form incorrect conclusions and spurious connections. This is particularly prominent in sub-categories, such as hobbies and work, where users often reference specific content (e.g., particular TV series or video games) from previous discussions. Without target domain context, the generator cannot accurately identify which specific interests the user is referring to. We verify the importance of target domain content by measuring retrieval precision for each user, treating posts from the target domain as relevant and others as non-relevant, then averaging across users. The average P@10 was 0.7834 (A&E: 0.8661, L&P: 0.6581, S&C: 0.8261), indicating that more than half of retrieved posts from user profiles belong to the target domain. In the Cross-Domain setting, we deliberately exclude these relevant posts, explaining why Cross-Domain personalization presents such a significant challenge.

Our proposed D-RAG method substantially improves over standard RAG by incorporating public information containing the target domain, recovering much of the lost performance. Notably, Public-Only often outperforms RAG in cross-domain settings, suggesting that when user profiles are domain-misaligned, personalization attempts can be counterproductive. D-RAG effectively combines

Table 1: Test set performance across different domain settings. A&E, L&P, and S&C denote the target domain of the question. Bold indicates best results. † denotes statistical significance against all other baselines, and * denotes statistical significance against RAG counterpart method (paired t-test, Bonferroni Correction applied, $p < 0.05$).

		Cross Domain				Multi Domain				In Domain			
	Method	A&E	L&P	S&C	Avg	A&E	L&P	S&C	Avg	A&E	L&P	S&C	Avg
Qwen 2.5 (3B)	No-Personalization	0.2810	**0.4364**	0.4518	0.3897	0.2810	0.4364	0.4518	0.3897	0.2810	0.4364	0.4518	0.3897
	Public-Only	0.2275	0.3759	0.3952	0.3328	0.2275	0.3759	0.3952	0.3328	0.2275	0.3759	0.3952	0.3328
	RAG	0.1907	0.3080	0.3092	0.2693	0.2757	0.3708	0.4107	0.3524	0.2684	0.3718	0.4106	0.3503
	Plan-RAG	0.2844	0.4123	0.4338	0.3768	0.3318	0.4465	0.4679	0.4154	0.3322	0.4386	**0.4812**	**0.4173**
	D-RAG	0.2659*	0.4002*	0.4318*	0.3659*	0.2858	0.4231*	0.4458*	0.3849*	0.2823	0.4214*	0.4503*	0.3847*
	Plan-D-RAG	**0.3149**†*	0.4295*	**0.4595***	**0.4013**†*	**0.3329**	**0.4478**	**0.4760**	**0.4189**	**0.3362**	**0.4406**	0.4719	0.4162
Qwen 2.5 (7B)	No-Personalization	0.3196	0.4662	0.4794	0.4217	0.3196	0.4662	0.4794	0.4217	0.3196	0.4662	0.4794	0.4217
	Public-Only	**0.3684**	0.4890	0.5193	**0.4589**	0.3684	0.4890	0.5193	0.4589	0.3684	0.4890	0.5193	0.4589
	RAG	0.2615	0.3996	0.4346	0.3652	0.3424	0.4486	0.5011	0.4307	0.3400	0.4451	0.5027	0.4293
	Plan-RAG	0.3264	0.4642	0.4958	0.4288	0.3653	0.4880	0.5305	0.4612	0.3648	0.4827	0.5350	0.4608
	D-RAG	0.3319*	0.4803†*	0.4958*	0.4360*	0.3558	0.4855*	0.4963	0.4458*	0.3524	0.4837*	0.5016	0.4459*
	Plan-D-RAG	0.3607*	**0.4910**†*	**0.5243**†*	0.4587*	**0.3745**	**0.5008***	**0.5398**	**0.4717**†*	**0.3801***	**0.5068**†*	**0.5361**	**0.4743**†*
Qwen 2.5 (14B)	No-Personalization	0.3325	0.4882	0.5058	0.4421	0.3325	0.4882	0.5058	0.4421	0.3325	0.4882	0.5058	0.4421
	Public-Only	0.3399	0.4883	0.5047	0.4443	0.3399	0.4883	0.5047	0.4443	0.3399	0.4883	0.5047	0.4443
	RAG	0.3043	0.4311	0.4526	0.3960	0.3902	0.4866	0.5379	0.4715	0.3897	0.4805	0.5358	0.4687
	Plan-RAG	0.3513	0.4614	0.5060	0.4395	0.4019	0.4934	0.5639	0.4864	0.4054	0.4888	0.5612	0.4851
	D-RAG	0.3371*	0.4848*	0.5184*	0.4467*	0.3665	0.4915	0.5309	0.4629	0.3671	0.4871	0.5308	0.4617
	Plan-D-RAG	**0.3978**†*	**0.5038**†*	**0.5621**†*	**0.4879**†*	**0.4092**	**0.5113**†*	**0.5705**	**0.4970**†*	**0.4070**	**0.5060**†*	**0.5733***	**0.4954**†*
GPT-4o-mini	No-Personalization	0.3797	0.5159	0.5252	0.4736	0.3797	0.5159	0.5252	0.4736	0.3797	0.5159	0.5252	0.4736
	Public-Only	0.4272	0.5280	0.5761	0.5104	0.4272	0.5280	0.5761	0.5104	0.4272	0.5280	0.5761	0.5104
	RAG	0.3089	0.4401	0.4479	0.3989	0.3954	0.4778	0.5308	0.4680	0.4042	0.4818	0.5262	0.4707
	Plan-RAG	0.3949	0.5275	0.5546	0.4923	0.453	0.5484	0.6091	0.5368	**0.4604**	0.5475	0.6095	0.5391
	D-RAG	0.3960*	0.5469†*	0.5771*	0.5066*	0.4223*	0.5557*	0.6067*	0.5282*	0.4239*	0.5561*	0.6069*	0.5289*
	Plan-D-RAG	**0.4494**†*	**0.5523**†*	**0.6093**†*	**0.5370**†*	**0.4559**†	**0.5645**†*	**0.6226**†*	**0.5476**†*	0.4600	**0.5573***	**0.6211**†*	**0.5461**†*

the benefits of both approaches, matching or exceeding Public-Only performance while retaining the ability to leverage user-specific signals. Plan-D-RAG further extends these gains, consistently matching or exceeding all baselines, with larger models showing the most substantial improvements.

Multi-Domain and In-Domain Results. The Multi-Domain setting represents the most realistic scenario for real-world personalization systems, where user profiles span multiple domains including the target. Interestingly, performance here is comparable to the In-Domain setting, and in some cases slightly exceeds it, suggesting that diverse user profiles are not inherently detrimental; rather, additional context from related domains may provide valuable supplementary information. The relative benefit of Dual Retrieval Augmentation remains substantial across all model sizes, with Plan-D-RAG consistently outperforming both standard RAG and Plan-RAG. In the In-Domain setting, where user profiles contain exclusively information from the target domain, standard RAG performs substantially better, validating that domain alignment is crucial for effective personalization. The contrast between RAG performance in Cross-Domain versus In-Domain settings underscores a central challenge this work addresses. Plan-D-RAG continues to outperform all baselines even in this favorable scenario, indicating that supplementing user profiles with public knowledge provides complementary information even when profiles are already domain-aligned. The performance gap between Plan-RAG and Plan-D-RAG is notably smaller in this setting compared to Cross-Domain, suggesting that the planning mechanism with Dual Retrieval Augmentation provides greater marginal benefit when compensating for missing domain knowledge.

Instance-Level Analysis. While aggregate metrics demonstrate the effectiveness of Dual Retrieval Augmentation, Figure 2 reveals important findings at the instance level. Comparing D-RAG against RAG using GPT-4o-mini, D-RAG wins on substantially more instances than it loses across all settings, with the advantage most pronounced in the Cross-Domain setting. This win rate narrows in Multi-Domain and In-Domain settings, suggesting that dual retrieval provides the greatest marginal benefit when user profiles lack target domain content. However, a non-trivial number of instances exhibit performance degradation with dual retrieval when comparing planning-based personalization methods. We hypothesize that in these cases, the planner generates aspects favoring public knowledge at the expense of more valuable user-specific information, producing plans that diverge from optimal aspects. This suggests that while planning is beneficial overall, it may be more sensitive to the balance between personalization and target domain knowledge. Developing adaptive retrieval strategies that dynamically weight sources based on query characteristics remains a promising direction for future work.

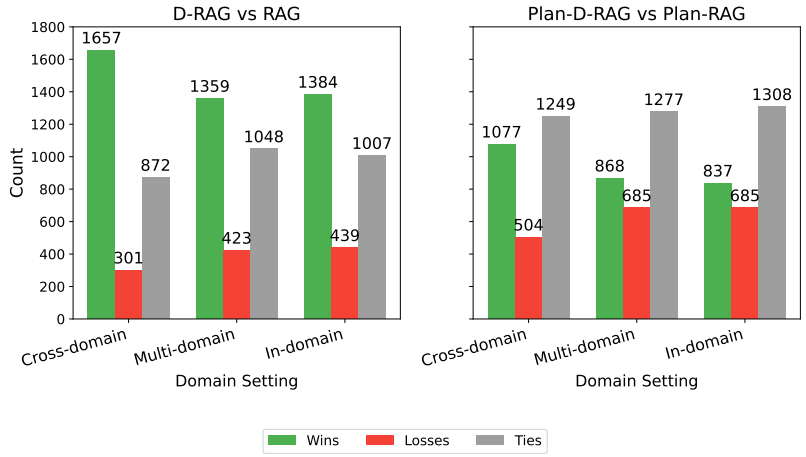


Figure 2: Win/Loss/Tie Counts of Dual Retrieval Augmentation against the baseline model (GPT-4o-mini).

4 Conclusion

This paper systematically analyzed personalized question answering using the LaMP-QA benchmark across three domain composition settings: Cross-Domain, Multi-Domain, and In-Domain. We demonstrated that state-of-the-art personalization methods struggle when user profiles lack target domain content, often performing worse than non-personalized baselines in Cross-Domain scenarios. To address this challenge, we introduced Dual Retrieval Augmentation, a framework that jointly retrieves from both user profiles and a public corpus to influence the model distribution towards the target domain. Experiment results demonstrate that methods using Dual Retrieval Augmentation consistently improves performance across all domain composition settings, with particularly substantial gains in challenging Cross-Domain scenarios. We further showed that combining dual retrieval with planning-based generation (Plan-D-RAG) achieves the best overall performance. Our findings highlight the importance of considering domain composition in personalization and demonstrate that leveraging multiple information sources provides a robust solution for personalized question answering with diverse user profiles.

Our study has several limitations that suggest directions for future work. Mainly, our experiments mostly use models from the Qwen 2.5 family for both generation and evaluation. While this provides a controlled experimental setting, further research is needed to verify whether the observed failure of personalization in cross-domain and multi-domain settings, as well as the effectiveness of Dual Retrieval Augmentation, generalize across other model families and scales.

Second, while the cross-domain setting may appear extreme or artificial, it mirrors a realistic cold-start scenario in which a user has no prior experience in the target domain, or too little relevant history for meaningful retrieval. Investigating how Dual Retrieval Augmentation performs under varying degrees of domain overlap ranging from full cold-start to sparse in-domain coverage, would provide a more complete picture of its practical applicability.

5 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:1611.09268 [cs.CL] https://arxiv.org/abs/1611.09268

[2] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rygGQyrFvH

[3] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[4] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=jKN1pXi7b0

[5] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging. arXiv:2310.11564

[6] Pranav Kasela, Marco Braga, Gabriella Pasi, and Raffaele Perego. 2024. SE-PQA: Personalized Community Question Answering. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1095–1098. doi:10.1145/3589335.3651445

[7] Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A. Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, Nedim Lipka, Chien Van Nguyen, Thien Huu Nguyen, and Hamed Zamani. 2024. LongLaMP: A Benchmark for Personalized Long-form Text Generation. arXiv:2407.11016 [cs.CL] https://arxiv.org/abs/2407.11016

[8] Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2024. Learning to Rewrite Prompts for Personalized Text Generation. In *Proceedings of the ACM on Web Conference 2024 (WWW '24)*. ACM. doi:10.1145/3589334.3645408

[9] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12076–12100. doi:10.18653/v1/2023.emnlp-main.741

[10] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. arXiv:1906.00091 [cs.IR]

[11] OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] https://arxiv.org/abs/2410.21276

[12] Qwen. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[13] Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization Methods for Personalizing Large Language Models through Retrieval Augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 752–762. doi:10.1145/3626772.3657783

[14] Alireza Salemi, Julian Killingback, and Hamed Zamani. 2025. ExPerT: Effective and Explainable Evaluation of Personalized Long-Form Text Generation. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Vienna, Austria, 17516–17532. doi:10.18653/v1/2025.findings-acl.900

[15] 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)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7370–7392. doi:10.18653/v1/2024.acl-long.399

[16] Alireza Salemi and Hamed Zamani. 2025. LaMP-QA: A Benchmark for Personalized Long-form Question Answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 1139–1159. doi:10.18653/v1/2025.emnlp-main.60

[17] Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. Personalized Pieces: Efficient Personalized Large Language Models through Collaborative Efforts. arXiv:2406.10471

[18] Yiyan Xu, Jinghao Zhang, Alireza Salemi, Xinting Hu, Wenjie Wang, Fuli Feng, Hamed Zamani, Xiangnan He, and Tat-Seng Chua. 2025. Personalized Generation In Large Model Era: A Survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 24607–24649. doi:10.18653/v1/2025.acl-long.1201

[19] Gui-Rong Xue, Jie Han, Yong Yu, and Qiang Yang. 2009. User Language Model for Collaborative Personalized Search. *ACM Trans. Inf. Syst.* 27, 2, Article 11 (mar 2009), 28 pages. doi:10.1145/1462198.1462203