PrLM: Learning Explicit Reasoning for Personalized RAG via Contrastive Reward Optimization

Kepu Zhang*
Teng Shi*
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
{kepuzhang,shiteng}@ruc.edu.cn

Weijie Yu†
School of Information Technology and Management
University of International Business and Economics
Beijing, China
yu@uibe.edu.cn

Jun Xu
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China
junxu@ruc.edu.cn

Abstract

Personalized retrieval-augmented generation (RAG) aims to produce user-tailored responses by incorporating retrieved user profiles alongside the input query. Existing methods primarily focus on improving retrieval and rely on large language models (LLMs) to implicitly integrate the retrieved context with the query. However, such models are often sensitive to retrieval quality and may generate responses that are misaligned with user preferences. To address this limitation, we propose PrLM, a reinforcement learning framework that trains LLMs to explicitly reason over retrieved user profiles. Guided by a contrastively trained personalization reward model, PrLM effectively learns from user responses without requiring annotated reasoning paths. Experiments on three personalized text generation datasets show that PrLM outperforms existing methods and remains robust across varying numbers of retrieved profiles and different retrievers.

CCS Concepts

• **Information systems** → **Personalization**; • **Computing methodologies** → **Natural language generation**.

Keywords

Personalization, Retrieval augmented generation, Reinforcement learning

ACM Reference Format:
Kepu Zhang*, Teng Shi*, Weijie Yu, and Jun Xu. 2025. PrLM: Learning Explicit Reasoning for Personalized RAG via Contrastive Reward Optimization . In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3746252.3760851

†Weijie Yu is the corresponding author.
* Equal contribution. Work partially done at Beijing Key Laboratory of Research on Large Models and Intelligent Governance, and Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE.

CIKM '25, November 10–14, 2025, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3760851

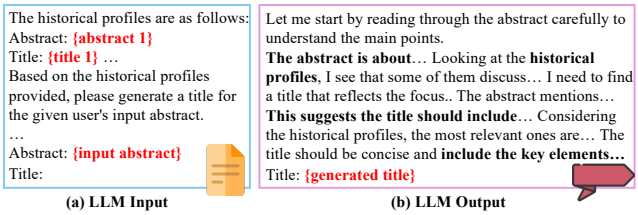


Figure 1: An example of explicit reasoning in personalized RAG from the LaMP-5 dataset. (a) The input includes the user's historical abstracts and titles, along with the target abstract. (b) The LLM trained with PrLM generates a personalized title by explicitly reasoning over the retrieved user profile and the target abstract.

1 Introduction

Personalized retrieval-augmented generation (RAG) [20, 21] aims to generate user-specific responses by incorporating retrieved user profiles alongside the input query. This task has drawn increasing attention due to its broad applicability in personalized dialogue, summarization, recommendation explanation, and beyond [23]. By combining the strengths of retrieval and generation [8, 34, 37], personalized RAG systems can provide responses that are both grounded in external knowledge and tailored to individual users.

Recent studies [3, 18, 26] in personalized RAG have primarily focused on enhancing the retrieval stage. These methods aim to select the most relevant user profiles and concatenate them with the query before inputting the combined context into a large language model (LLM). The LLM is then expected to implicitly integrate the retrieved information and produce a personalized response. Despite achieving promising results, this implicit reasoning paradigm is highly sensitive to retrieval noise. Motivated by recent advances in LLM reasoning [4, 31], we argue that introducing an explicit reasoning process—where the LLM first reflects on the retrieved context before producing the final output—has the potential to improve the robustness, and personalization fidelity. Figure 1 illustrates an example of explicit reasoning in a personalized RAG. The LLM analyzes the target abstract in light of the user's historical profiles and generates an intermediate reasoning trace that guides the final personalized title. This structured process highlights the model's capacity to align more faithfully with user preferences.

However, implementing explicit reasoning in personalized RAG poses two major challenges. **First**, there is a lack of annotated intermediate reasoning paths, making it difficult to directly supervise

the reasoning process during training. **Second**, current training objectives do not account for the degree of personalization in the generated responses. To effectively guide explicit reasoning, it is necessary to design a reward model that can evaluate and optimize for personalization quality.

To address the above challenges, we propose PrLM, a reinforcement learning framework that enables LLMs to explicitly reason over retrieved user profiles in personalized RAG. We leverage GRPO [22] to train the LLM by sampling and optimizing over groups of reasoning paths without requiring intermediate supervision. This approach encourages the model to learn a structured decision process that infers user preferences before generating the final response. By modeling this explicit reasoning, PrLM enhances the model's ability to effectively leverage personalized context and improves the faithfulness of the generated outputs. Moreover, to assess the degree of personalization, we further introduce a contrastively trained reward model. For the same input query, we compare the LLM's responses generated with and without retrieved user profiles, and train the reward model to assign higher scores to responses that better reflect user-specific information. This personalization reward is then integrated into the GRPO framework to guide learning. Experiments on three personalized text generation datasets demonstrate that PrLM consistently outperforms existing baselines and maintains robustness across varying numbers of retrieved profiles and different retrievers.

To summarize, our contributions are as follows:

- We propose PrLM, a novel reinforcement learning framework that enables large language models to perform explicit reasoning over retrieved user profiles for personalized RAG.
- We introduce a contrastive personalization reward model that provides supervision for personalization by comparing outputs with and without retrieved profiles, guiding LLMs to favor more personalized outputs.
- Extensive experiments on three personalized RAG datasets show PrLM improves both personalization quality and robustness across different retrieval settings and profile quantities.

2 Related Work

2.1 Personalized RAG

In recent years, personalization has gradually drawn people's attention [25, 27, 33]. The personalized RAG task [13–15, 20, 21], based on a user's historical behavior, can generate personalized outputs for the same query according to different users. Existing personalized RAG methods [10, 26, 39] mainly focus on how to integrate the information of user historical behavior into the prompt of the LLM, so that the LLM can implicitly summarize the user's preferences. In this paper, we mainly explore the integration of reasoning with personalized RAG and investigate whether the proposed method can be adapted to various retrieval methods.

2.2 LLM Reasoning

Recently, reasoning LLMs have emerged with the advent of models like Deepseek-R1 [6]. Numerous studies have begun to explore various aspects of enhancing the reasoning capabilities of LLMs [1, 32, 35, 36] and leveraging these capabilities to solve a wide range of problems [2, 11, 12, 24, 28, 30, 38]. In this paper, we primarily investigate reasoning in the context of personalized RAG. We explore how to implement reasoning in personalized RAG and utilize reasoning to help LLMs achieve better personalized outputs.

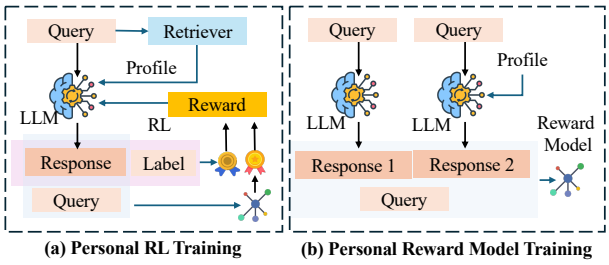


Figure 2: The architecture of PrLM.

Table 1: Statistics of used datasets.

Dataset	#User	#Train	Dev	Test
LaMP-4	1,643	12,500	1,500	1,800
LaMP-5	14,682	14,682	1,500	1,500
LaMP-7	13,437	13,437	1,498	1,500

3 Method

3.1 Task Formulation

The goal of personalized RAG is to generate a user-specific response y based on a user query x and relevant historical profile H:

$$y = \text{LLM}(x, H). \tag{1}$$

The historical profile H is retrieved from a user-specific corpus D, which contains the user's past behavior records (e.g., previous queries, responses, or interactions) given user query x:

$$H = \text{Retrieve}(x, D). \tag{2}$$

While prior work largely targets improving the retrieval step—i.e., obtaining higher-quality H—our focus lies in the generation process itself. We aim to train the LLM to reason explicitly over retrieved profiles and to produce personalized responses that faithfully reflect user preferences, even in the absence of explicit reasoning supervision. To this end, we introduce a group-based reinforcement learning strategy to guide reasoning (§3.2) and a contrastive reward model to assess personalization quality (§3.3).

3.2 Personalized Reasoning Model Training

To enable the LLM to learn explicit reasoning in the absence of annotated reasoning paths, we adopt Group Relative Policy Optimization (GRPO) [22], a reinforcement learning algorithm well-suited for optimizing over multiple sampled reasoning trajectories without relying on a value function. GRPO encourages the model to explore diverse reasoning strategies and preferentially reinforce those that yield higher-quality personalized outputs. Formally, Given a user query x and the retrieved user profile H, the LLM generates a response o composed of two parts: an intermediate reasoning path R enclosed in <think>...</think> and a final personalized output y following the reasoning. During training, multiple such outputs are sampled, and GRPO updates the model using their relative ranking based on a composite reward signal. The total reward assigned to a

Table 2: The main experimental results on the LaMP-4, LaMP-5, and LaMP-7 datasets. Bold indicates the best results.

	LaMP-4				LaMP-5				LaMP-7			
Model	**Rouge-1**	**Rouge-2**	**Rouge-L**	**BLEU**	**Rouge-1**	**Rouge-2**	**Rouge-L**	**BLEU**	**Rouge-1**	**Rouge-2**	**Rouge-L**	**BLEU**
Zero-Shot	1.29	0.25	1.03	2.23	25.19	11.79	20.47	27.34	26.83	9.69	20.36	23.74
Random	8.21	1.73	7.11	9.97	26.10	10.99	21.35	30.39	30.85	11.89	25.00	28.61
Recency	8.36	1.96	7.29	10.64	28.69	12.65	23.67	32.65	30.35	11.79	24.85	28.52
BM25	8.91	1.73	7.61	10.76	31.07	13.72	25.20	35.20	31.03	11.90	25.26	29.17
BGE	9.66	2.15	8.39	11.78	30.21	13.45	24.91	34.03	32.46	12.90	26.67	30.57
ROPG	9.88	2.30	8.49	11.95	31.19	13.56	25.25	34.66	31.68	12.52	25.92	29.60
CFRAG	9.76	2.23	8.34	12.06	31.41	14.56	26.05	34.59	33.42	13.11	27.45	31.49
PrLM	**12.02**	2.59	**10.35**	**14.69**	**36.74**	**17.25**	**30.93**	**40.56**	**35.96**	**14.41**	**30.11**	**33.80**
w/o r_{personal}	11.03	**2.66**	9.43	13.28	35.07	16.40	29.40	38.82	33.98	13.58	28.55	31.97

sampled response is defined as:

$$r = r_{\text{correct}} + \alpha r_{\text{think}} + \beta r_{\text{personal}}, \tag{3}$$

where α and β are tunable hyperparameters controlling the importance of each component. The reward terms are defined as follows:

- **Format reward** r_{think}: this binary reward verifies whether the generated output includes a `<think>` segment, encouraging the model to explicitly reason before generating the final response.
- **Correctness reward** r_{correct}: this reward measures the textual similarity between the generated output y and the annotated reference using the sum of ROUGE-1, ROUGE-2, and ROUGE-L scores. Although it captures basic response quality, it only weakly reflects personalization.
- **Personalization reward** r_{personal}: this component is computed by a separately trained reward model that quantifies the alignment between the output and the user's historical preferences. Details are provided in §3.3.

We initialize the LLM using `DeepSeek-R1-Distill-1.5B`. To control training cost, we fine-tune the model on a randomly selected subset of 500 training samples.

3.3 Personalized Reward Model Training

To quantitatively evaluate the degree of personalization in LLM outputs, we train a reward model via preference-based learning inspired by Direct Preference Optimization (DPO) [16]. Unlike conventional response quality metrics, our reward model is explicitly designed to capture whether the generated output reflects user-specific preferences.

We implement the reward model as a BERT [5] encoder, taking as input the concatenated pair "[CLS] x [SEP] y [SEP]", where x is the user query and y is the generated response. The model outputs a scalar score representing the degree of personalization for the (x, y) pair. To construct training data, we collect 5,000 triplets (x, y_+, y_-), where y_+ is a personalized response generated by conditioning the LLM on both the query x and retrieved profile H, and y_- is a non-personalized response generated in a zero-shot setting (i.e., using x alone). These paired comparisons allow us to train the reward model to prefer responses that reflect personalized content. The optimization objective is a contrastive preference loss:

$$\mathcal{L} = -\log\left(\sigma(s_p - s_n)\right), \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and s_p and s_n are the model's predicted scores for the preferred and non-preferred responses, respectively:

$$s_p = \text{BERT}(x, y_+), \quad s_n = \text{BERT}(x, y_-) \tag{5}$$

We use BGE [29] as the retriever to construct the personalized inputs for generating y_+. Once trained, the reward model can be applied to score the personalization degree of any output y given a query x:

$$r_{\text{personal}} = \text{BERT}(x, y). \tag{6}$$

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset and Metric. We use the datasets from the LaMP [21], selecting LaMP-4: Personalized News Headline Generation; LaMP-5: Personalized Scholarly Title Generation; and LaMP-7: Personalized Tweet Paraphrasing. We use the temporal splitting. Statistical analyses are in Table 1. Following [26], we evaluate the outputs using ROUGE-1, ROUGE-2, ROUGE-L, and BLEU.

4.1.2 Comparison Methods. PrLM is retrieval-agnostic approach, we compared it with the following methods: **Zero-Shot**: Directly input the user's query into the LLM to obtain the output without any retrieval. This is a non-personalized baseline. **Random**: Randomly select k user profiles. **Recency**: Select the most recent k user profiles according to the time. **BM25** [19]: Retrieve the top k user profiles using BM25 based on the user's query. **BGE** [29]: Retrieve the top k user profiles using BGE based on the user's query. **ROPG** [20]: Retrieval after fine-tuning BGE with feedback from the LLM. **CFRAG** [26]: Introducing similar user retrieval, retrieving first and then re-ranking. We alse compare with **w/o** r_{personal}.

4.1.3 Implementation Details. We retrieve five documents and use greedy decoding during generation to make the results easily reproducible. BGE is bge-base-en-v1.5 [29]. We trained the PrLM using GRPO. We sampled 4 outputs each time, with a maximum completion length of 768. We employed LoRA [7] for efficient training, where the rank of LoRA was set to 16 and the alpha was also 16. The learning rate was set to 1e-6. In Eq. 3, both α and β were set to 0.1. We used DeepSpeed ZeRO-2 [17] and vLLM [9] for training and testing. For personalized reward model training, we trained the BERT-base-uncased model. We conducted our experiments on A6000 GPUs. More details can be found at https://github.com/ke-01/PrLM.

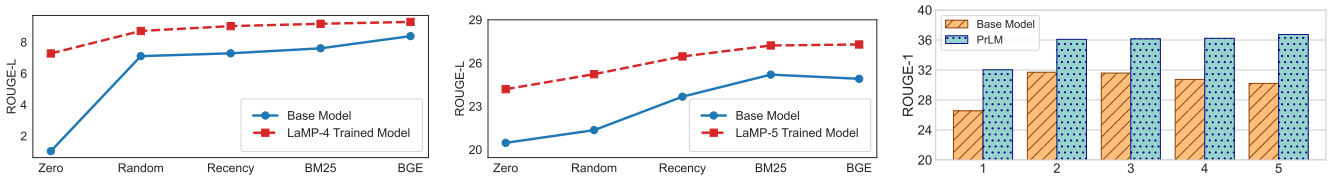
4.2 Main Results

The experimental results are shown in Table 2. We observe that:

User profiles are important. It can be seen that as long as the user's historical behavior is retrieved and added to the prompt of the LLM, the output of the model is better than the zero-shot setting. This is similar to the conclusions of previous work [26].

Table 3: The retriever robustness results on the LaMP-4, LaMP-5, and LaMP-7 datasets. Bold indicates the best results.

	LaMP-4				LaMP-5				LaMP-7			
Model	**Rouge-1**	**Rouge-2**	**Rouge-L**	**BLEU**	**Rouge-1**	**Rouge-2**	**Rouge-L**	**BLEU**	**Rouge-1**	**Rouge-2**	**Rouge-L**	**BLEU**
Zero-Shot	1.29	0.25	1.03	2.23	25.19	11.79	20.47	27.34	26.83	9.69	20.36	23.74
+PrLM	**3.21**	**0.71**	**2.75**	**4.19**	**34.36**	**17.22**	**28.80**	**37.33**	**34.45**	**13.37**	**28.53**	**33.12**
Random	8.21	1.73	7.11	9.97	26.10	10.99	21.35	30.39	30.85	11.89	25.00	28.61
+PrLM	**10.10**	**1.97**	**8.68**	**12.50**	**33.23**	**15.22**	**28.01**	**37.57**	**35.62**	**14.11**	**29.86**	**33.27**
Recency	8.36	1.96	7.29	10.64	28.69	12.65	23.67	32.65	30.35	11.79	24.85	28.52
+PrLM	**11.17**	**2.31**	**9.73**	**13.78**	**35.15**	**16.56**	**29.70**	**39.16**	**36.20**	**14.55**	**30.29**	**34.23**
BM25	8.91	1.73	7.61	10.76	31.07	13.72	25.20	35.20	31.03	11.90	25.26	29.17
+PrLM	**10.95**	**2.06**	**9.44**	**13.43**	**36.56**	**17.22**	**30.55**	**40.38**	**36.58**	**14.52**	**30.68**	**34.32**
ROPG	9.88	2.30	8.49	11.95	31.19	13.56	25.25	34.66	31.68	12.52	25.92	29.60
+PrLM	**11.92**	**2.42**	**10.18**	**14.51**	**36.24**	**16.80**	**30.21**	**39.57**	**36.38**	**14.55**	**30.45**	**34.29**
CFRAG	9.76	2.23	8.34	12.06	31.41	14.56	26.05	34.59	33.42	13.11	27.45	31.49
+PrLM	**12.50**	**2.66**	**10.68**	**14.90**	**37.22**	**17.68**	**31.20**	**40.83**	**36.54**	**14.70**	**30.62**	**34.32**



(a) LaMP-4-trained LLM applied to LaMP-5. (b) LaMP-5-trained LLM applied to LaMP-4.. (c) Different numbers of user profiles in LaMP-5.

Figure 3: The Base model is the original Deepseek-R1-Distill-Qwen1.5B. (a) and (b): Cross domain Results. The X-axis represents the retrieval strategy. (c) The X-axis represents the number of historical user behaviors retrieved.

The user's historical behavior helps the LLM to better capture the user's preferences, thereby providing more personalized outputs.

PrLM achieves the best performance. We can see that in all datasets, compared with all the baselines, PrLM achieved the best results. This indicates the effectiveness of PrLM. The personalized reward model helps guide the LLM to better capture the user's true preferences, thereby generating more personalized outputs.

Both the retriever and the LLM are important. We can see that the retriever is very important. For example, we can see that in LaMP-5, the effect of BGE retrieval is not as good as that of BM25. In addition, PrLM and w/o $r_{personal}$ trained the LLM, making the LLM perform better than the original LLM. This shows that training a better LLM is also very important for personalized RAG.

4.3 Retriever Robustness

In this section, we apply the LLM trained with documents retrieved by the BGE strategy to different retrieval strategies. The results are shown in Table 3. It can be seen that adding PrLM achieves improvements across all retrieval strategies. This indicates that PrLM has high robustness to different retrieval strategies. PrLM enables the LLM to better utilize the user profiles.

4.4 Cross-domain Robustness

In this section, we investigate how PrLM performs on cross-domain tasks. Specifically, we apply the model trained on the LaMP-4 to the LaMP-5 and vice versa. We conduct experiments under different retrieval strategies. The results are shown in Figure 3(a) and 3(b). It can be seen that when the model trained on the LaMP-4 is applied to LaMP-5, and vice versa, PrLM achieves better performance. This indicates that PrLM has good cross-domain capabilities. This further demonstrates the robustness of PrLM.

4.5 Robustness of User Profiles Numbers

In this section, we investigate the model's performance when retrieving different numbers of user profiles, with the results shown in Figure 3(c). It can be observed that the performance of Deepseek-R1-Distill-Qwen-1.5B initially increases and then decreases as the number of retrieved user profiles grows. This is because a larger volume of user behaviors introduces more noise, making it difficult for the model to distinguish which parts of the user behaviors are relevant. In contrast, PrLM's performance gradually improves with the increase in the number of documents and does not decline. This indicates that PrLM exhibits better robustness and generalizability, effectively utilizing the user profiles

5 Conclusion

In this work, we propose PrLM, a reinforcement learning framework for personalized retrieval-augmented generation that enables large language models to perform explicit reasoning over retrieved user profiles. Specifically, we adopt a group-based reinforcement learning strategy to guide the model's reasoning process to address the lack of annotated reasoning supervision. Additionally, we introduce a contrastive reward modeling approach to assess the personalization quality of LLM-generated responses. Experimental results on three personalized generation benchmarks demonstrate that PrLM consistently improves response quality and enhances robustness across varying retrieval settings.

Acknowledgments

This work was funded by the National Key R&D Program of China (2023YFA1008704), the National Natural Science Foundation of China (62472426). Supported by fund for building world-class universities (disciplines) of Renmin University of China.

GenAI Usage Disclosure

I know that the ACM's Authorship Policy requires full disclosure of all use of generative AI tools in all stages of the research (including the code and data) and the writing. No GenAI tools were used in any stage of the research, nor in the writing.

References

[1] Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697* (2025).

[2] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925* (2024).

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

[4] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567* (2025).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[6] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[8] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 7969–7992.

[9] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*. 611–626.

[10] Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023. Teach LLMs to Personalize–An Approach inspired by Writing Education. *arXiv preprint arXiv:2308.07968* (2023).

[11] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419* (2025).

[12] Jiacheng Lin, Tian Wang, and Kun Qian. 2025. Rec-R1: Bridging Generative Large Language Models and User-Centric Recommendation Systems via Reinforcement Learning. *arXiv preprint arXiv:2503.24289* (2025).

[13] Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2023. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. *arXiv preprint arXiv:2311.09180* (2023).

[14] Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. 2025. Latent Inter-User Difference Modeling for LLM Personalization. *arXiv preprint arXiv:2507.20849* (2025).

[15] Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. 2025. Measuring what makes you unique: Difference-aware user modeling for enhancing llm personalization. *arXiv preprint arXiv:2503.02450* (2025).

[16] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* 36 (2023), 53728–53741.

[17] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–16.

[18] Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081* (2023).

[19] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[20] 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*. 752–762.

[21] 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)*. 7370–7392.

[22] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

[23] Chenglei Shen, Zhongxiang Sun, Teng Shi, Xiao Zhang, and Jun Xu. 2025. StyliTruth: Unlocking Stylized yet Truthful LLM Generation via Disentangled Steering. *arXiv preprint arXiv:2508.04530* (2025).

[24] Teng Shi, Weicong Qin, Weijie Yu, Xiao Zhang, Ming He, Jianping Fan, and Jun Xu. 2025. Bridging Search and Recommendation through Latent Cross Reasoning. *arXiv preprint arXiv:2508.04152* (2025).

[25] Teng Shi, Zihua Si, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Dewei Leng, Yanan Niu, and Yang Song. 2024. UniSAR: Modeling User Transition Behaviors between Search and Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1029–1039.

[26] Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Han Li. 2025. Retrieval Augmented Generation with Collaborative Filtering for Personalized Text Generation. *arXiv preprint arXiv:2504.05731* (2025).

[27] Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Enyun Yu. 2025. Unified Generative Search and Recommendation. *arXiv preprint arXiv:2504.05730* (2025).

[28] Teng Shi, Weijie Yu, Xiao Zhang, Ming He, Jianping Fan, and Jun Xu. 2025. Benefit from Rich: Tackling Search Interaction Sparsity in Search Enhanced Recommendation. *arXiv preprint arXiv:2508.04145* (2025).

[29] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597 [cs.CL]

[30] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768* (2025).

[31] Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. 2025. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv preprint arXiv:2501.09686* (2025).

[32] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. LIMO: Less is More for Reasoning. *arXiv preprint arXiv:2502.03387* (2025).

[33] Kepu Zhang, Teng Shi, Sunhao Dai, Xiao Zhang, Yinfeng Li, Jing Lu, Xiaoxue Zang, Yang Song, and Jun Xu. 2024. SAQRec: Aligning Recommender Systems to User Satisfaction via Questionnaire Feedback. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 3165–3175.

[34] Kepu Zhang, Zhongxiang Sun, Weijie Yu, Xiaoxue Zang, Kai Zheng, Yang Song, Han Li, and Jun Xu. 2025. QE-RAG: A Robust Retrieval-Augmented Generation Benchmark for Query Entry Errors. *arXiv preprint arXiv:2504.04062* (2025).

[35] Kepu Zhang, Guofu Xie, Weijie Yu, Mingyue Xu, Xu Tang, Yaxin Li, and Jun Xu. 2025. Legal Mathematical Reasoning with LLMs: Procedural Alignment through Two-Stage Reinforcement Learning. *arXiv preprint arXiv:2504.02590* (2025).

[36] Kepu Zhang, Haoyue Yang, Xu Tang, Weijie Yu, and Jun Xu. 2024. Beyond guilt: Legal judgment prediction with trichotomous reasoning. *arXiv preprint arXiv:2412.14588* (2024).

[37] Kepu Zhang, Weijie Yu, Sunhao Dai, and Jun Xu. 2024. Citalaw: Enhancing llm with citations in legal domain. *arXiv preprint arXiv:2412.14556* (2024).

[38] Kepu Zhang, Weijie Yu, Zhongxiang Sun, and Jun Xu. 2025. Syler: A framework for explicit syllogistic legal reasoning in large language models. *arXiv preprint arXiv:2504.04042* (2025).

[39] Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. [n. d.]. HYDRA: Model Factorization Framework for Black-Box LLM Personalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.