

# Learning to Reason for Multi-Step Retrieval of Personal Context in Personalized Question Answering

Maryam Amirizani\*  
Information School  
University of Washington  
Seattle, WA, USA  
amaryam@uw.edu

Alireza Salemi  
Center for Intelligent Information  
Retrieval  
University of Massachusetts Amherst  
Amherst, MA, USA  
asalemi@cs.umass.edu

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

## Abstract

Personalization in Question Answering (QA) requires answers that are both accurate and aligned with users' background, preferences, and historical context. Existing state-of-the-art methods primarily rely on retrieval-augmented generation (RAG) solutions that construct personal context by retrieving relevant items from the user's profile. Existing methods use the user's query directly to retrieve personal documents and such strategies often lead to surface-level personalization. We propose PR<sup>2</sup> (Personalized Retrieval-Augmented Reasoning), a reinforcement learning framework that integrates reasoning and retrieval from personal context for personalization. PR<sup>2</sup> learns adaptive retrieval-reasoning policies, determining when to retrieve, what evidence to retrieve from user profiles, and how to incorporate it into intermediate reasoning steps. By optimizing multi-turn reasoning trajectories under a personalized reward function, the framework reinforces reasoning paths that better align with user-specific preferences and contextual signals reflected by the reward model. Extensive experiments on the LaMP-QA benchmark using three LLMs show that PR<sup>2</sup> consistently outperforms strong baselines, achieving an average relative improvement of 8.26%-12.0% in personalized QA.

## CCS Concepts

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

## Keywords

Personalization; Information Retrieval; Question Answering

### ACM Reference Format:

Maryam Amirizani, Alireza Salemi, and Hamed Zamani. 2026. Learning to Reason for Multi-Step Retrieval of Personal Context in Personalized Question Answering. 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, 7 pages. <https://doi.org/10.1145/3805712.3809944>

\*Work done as part of the CIIR Internship Program at University of Massachusetts Amherst.



This work is licensed under a Creative Commons Attribution 4.0 International License.

SIGIR '26, July 20–24, 2026, Melbourne, VIC, Australia

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2599-9/2026/07

<https://doi.org/10.1145/3805712.3809944>

## 1 Introduction

Personalization in LLMs has become increasingly important, particularly in information-seeking applications such as recommendation, search, and Question Answering (QA) [5, 7, 32, 35, 39]. Among these, personalized text generation for question answering (QA) remains underexplored. In personalized QA, models must generate responses that are tailored to users' preferences and contextual information [17, 25, 28]. For example, answering "What's the best way to lose weight?" requires accounting for personalized factors such as dietary restrictions, activity level, and medical history. The same question asked by different users may require different answers; identical responses suggest a failure to achieve personalization.

State-of-the-art personalized QA methods adopt a standard retrieval-augmented generation (RAG) paradigm, where personal context is retrieved from the user profile using the user's question as the search query and appended to the input prompt to guide response generation [12, 24, 27, 43]. We argue that effective personalization can be achieved by identifying potential personalized aspects that would impact the generated answer. Identifying these aspects requires accurate reasoning and sometimes generation of multiple queries that explore different personalization aspects. For instance, to answer the above mentioned question, it is important to identify the user's occupation to estimate the level of activity in their lifestyle and an effective model may generate the query "occupation" or later the query "age" to provide an accurate personalized answer. This demonstrates the importance of both reasoning and multi-step querying and retrieval from the user profile.

Inspired by recent work on retrieval-augmented reasoning (RAR) methods for question answering and deep research agents [13, 14, 19, 31, 34], we propose PR<sup>2</sup> (Personalized Retrieval-Augmented Reasoning), a reinforcement learning (RL) [33] framework that embeds retrieval into the reasoning process to enable adaptive and user-specific personalization. To optimize this integrated retrieval-reasoning process, PR<sup>2</sup> is trained with Group Relative Policy Optimization (GRPO) [11, 30], which models retrieval as part of the environment and optimizes full generation trajectories under personalized reward signals. During training, PR<sup>2</sup> samples both personalized and non-personalized trajectories to stabilize RL optimization and favoring user-aligned reasoning. The model supports multi-turn retrieval and reasoning, dynamically deciding when to retrieve, what information to retrieve, when to stop for generating the response, and how to incorporate it by interleaving retrieval actions with reasoning steps. This unified design enables adaptive and iterative retrieval and consumption of personal data throughout the

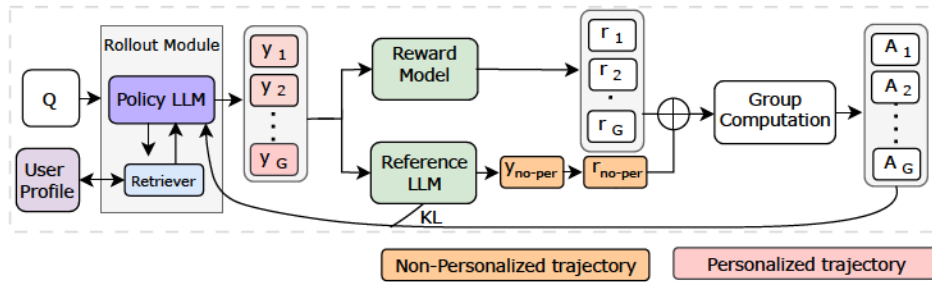


Figure 1: An overview of  $PR^2$  optimization based on GRPO.

reasoning process. Experiments on LaMP-QA [28], a recent personalized long-form question answering benchmark consisting of three diverse datasets, show that  $PR^2$  consistently outperforms strong baselines, achieving relative improvements of 8.26%-12.0% across different LLMs, highlighting its effectiveness for personalized QA. To improve reproducibility of this research, we open-source our implementation at <https://github.com/maryamamiri114/PR2>.

## 2 Related Work

**Reasoning in LLMs.** Many studies investigate methods to improve LLM reasoning across a wide range of tasks [1, 6, 42, 46] including QA [14, 24]. A common paradigm is step-by-step reasoning, often formalized as chain-of-thought (CoT) [38], which decomposes complex problems into intermediate steps. However, many methods rely on carefully designed prompts, limiting generalization. Recent work instead explores fine-tuning and RL to learn more robust reasoning behaviors [2, 20, 40, 44, 45].

**Personalized QA.** Personalized QA has largely focused on conditioning model behavior using user history at inference time [3, 15, 27, 29]. A common strategy is In-Context Learning, where past user query–response pairs are inserted into the prompt to induce personalized behavior [8, 22, 26, 37, 47, 47]. Other approaches directly utilize complete user histories for prompting [16, 21, 36], enabling personalization through structured profile representations rather than raw interaction logs. RAG extends this paradigm by retrieving semantically similar examples from user histories [10, 27, 48]. Subsequent work improves retrieval using reward signals from LLM outputs or dynamically selecting retrieval strategies based on input [24, 25]. Personalized reasoning in QA lies at the intersection of LLM reasoning and user-aware personalization. Early work relies on personal embeddings such as PPlug [21] for reasoning templates, while recent approaches use RL with contrastive rewards [43]. However, existing work primarily improves retrieval to enhance personalized answers. To address this gap, we propose  $PR^2$ , which integrates retrieved user-specific evidence directly into the intermediate reasoning process to generate personalized responses.

## 3 Methodology

We present  $PR^2$ , a framework for training LLMs for personalized QA via retrieval-augmented reasoning with GRPO optimization, enabling the model to learn how to effectively retrieve personal information from user profiles. Figure 1 shows the  $PR^2$  framework.

### 3.1 Problem Formulation

The training dataset is defined as  $D = \{(x_i, u_i, r_i, E_{x_i})\}_{i=1}^{|D|}$ , where  $x_i$  denotes the input question,  $u_i = \{d_i\}_{i=1}^{|u_i|}$  represents the associated user profile containing documents about user’s previous interactions with the system,  $r_i$  is the corresponding user-provided narrative for the question, and  $E_{x_i} = \{e_j\}_{j=1}^{|E_{x_i}|}$  is a set of personalized rubrics that can be used to evaluate generated personal responses for the user. Given  $(x_i, u_i)$  to the policy model, the objective is to generate a personalized response  $\hat{y}_i$  by leveraging the user-specific information contained in  $u_i$ . To evaluate  $\hat{y}_i$ , following [28], we use rubric aspects  $E_{x_i}$  and the user narrative  $r_i$  that the response to question  $x_i$  is expected to address. We then compute a scoring function  $\mu(x_i, \hat{y}_i, E_{x_i}, r_i)$  to assess the extent to which  $\hat{y}_i$  adequately covers these aspects (the evaluation strategy is explained in Section 4.1). We model the policy using an LLM  $\pi_\theta$  that generates responses conditioned on the question  $x_i$  and the user profile  $u_i$ , which it can access by submitting queries to a retrieval model  $\mathcal{R}$ . The policy  $\pi_\theta$  is initialized from a pretrained LLM checkpoint and optimized under the GRPO [30] reinforcement learning framework. A frozen reference model  $\pi_{\text{ref}}$  is used to regularize policy updates.

### 3.2 Personalized Rollout Sampling Strategy

This section describes the rollout sampling procedure in  $PR^2$ . Following Search-R1 [14],  $PR^2$  adopts the RAR paradigm, allowing the policy model  $\pi_\theta$  to interact with an external tool, here a retriever, during rollout generation. For each training instance  $(x_i, u_i)$ , the current policy samples a group of  $G$  personalized trajectories  $\{\hat{y}_i^{(g)}\}_{g=1}^G \sim \pi_\theta(\cdot | x_i, u_i) \otimes \mathcal{R}$ , where  $\otimes$  denotes the interleaving of token generation with retriever invocations (via  $R$ ) to fetch relevant user information from the profile. Thus, for each instance, we obtain a set of personalized sampled trajectories  $Y = \{y_1, \dots, y_G\}$ . The prompt template used for this personalized sampling is:

---

Your task is to generate a personalized response to the user’s question. To do this, you can perform a series of actions, including thinking in `<think>` and `</think>` tags, searching for information from the user past interactions with the system (i.e., previous asked questions and the detailed information need) by generating a non-empty search query in `<search>` and `</search>` tags, and finally providing the answer in `<answer>` and `</answer>` tags. The retrieved information from user history will be provided to you inside `<information>` and `</information>` tags. You need to first think about the question and how to generate a personalized answer for the user. In this thinking process, you should try to understand the user’s preferences and needs based on its past interactions with the system. The thinking process should be inside `<think>` and `</think>` tags. If you need to search for information about the user from its history, you can do this by generating a non-empty search query inside `<search>` and `</search>` tags. You can use this information in thinking process and answer generation. Nothing should be outside the mentioned tags except the initial question. Now, answer the following question: {question}

---

### 3.3 Personalized Reward Modeling

The reward function provides the primary supervision signal for policy optimization. In  $PR^2$ , we adopt an outcome-based reward scheme that relies exclusively on the final response rewards to guide the policy model for optimization. Specifically, as described in Section §3.1, for each query–response pair  $(x_i, \hat{y}_i)$ , an automatic evaluator from [28] assigns a discrete score based on  $\mu(x_i, \hat{y}_i, E_{x_i}, r_i)$ , measuring the extent to which the response satisfies the user-specific criteria derived from  $E_{x_i}$ . Formally, the reward is defined as  $r(x_i, \hat{y}_i) = \mu(x_i, \hat{y}_i, E_{x_i}, r_i)$ . To compute reward, following [28], we employ Qwen2.5-32B [41] as the judging model. For each query, the judge assigns a 0–2 score to each personalized aspect based on how well it is addressed. These values are rescaled to the interval  $[0, 1]$  and averaged across aspects. Additional details can be found in [28]. These rewards are then used to compute group-relative advantages that drive policy optimization.

### 3.4 Advantage Computation

In GRPO, the advantage is computed as a group-normalized reward for each output [30], and policy updates are driven by relative comparisons within a sampled response set. This formulation enables the policy to learn from differences in reward across multiple trajectories generated for the same input. However, prior work has shown that personalization does not consistently yield performance gains, and that non-personalized models can outperform personalized ones in many cases [24]. Consequently, it is desirable to encourage the policy to apply personalization by retrieving from the user profile only when doing so improves performance relative to non-personalized inference. To address this within the GRPO framework, we incorporate a non-personalized baseline when computing the advantage. This enables direct reward-based comparison between retrieval-augmented personalization and non-personal model, allowing the policy to learn when personalization truly improves response quality. As a result, the model is encouraged to favor retrieval from profile only when it yields measurable gains over non-personalization. A detailed analysis is provided in Section §5.

To obtain non-personalized responses, we generate a baseline trajectory without retrieval by conditioning only on the input,  $\hat{y}_{no-per} \sim \pi_{ref}(\cdot | x_i)$ . Such trajectories involve no retrieval and therefore no personalization. Each non-personalized baseline response  $\hat{y}_{no-per}$  is assigned a reward  $r_{no-per}$  computed as defined in Section §3.3. To compute the group-relative advantage under the personalized setting, we first compute group-level statistics over all personalized responses,  $\mu = \text{mean}(r_1, \dots, r_n)$  and  $\sigma = \text{std}(r_1, \dots, r_n)$ . For each personalized response  $\hat{y}_i \in Y$ , the group-relative advantage is defined as  $A_i = \frac{r_i - r_{no-per} - \mu}{\sigma}$ . This formulation encourages the model to produce personalized responses when they outperform the non-personalized baseline, and to fall back to non-personalized responses when personalization is detrimental, since the policy can choose to perform no retrieval from the user profile.

### 3.5 Summary

$PR^2$  is optimized using GRPO, which performs updates based on relative comparisons among grouped response samples without learning a value function [11, 30]. For each query  $x_i$ , we sample a group of responses  $Y$  as described in Section §3.2. We then apply the

GRPO objective and corresponding policy update to each sampled response  $\hat{y}_i \in Y$ , following the approach in [14]. In summary, the contributions of this approach compared to the literature include: (1) developing and evaluating the first reasoning models that retrieve from user profile for personalization, (2) optimizing a policy model for sampling personalization rollouts, (3) applying an aspect-based personalized evaluation to the generated responses as the reward function, and (4) including both personalized and non-personalized trajectories in computing the relative advantage function.

## 4 Experiment

### 4.1 Experiment Setup

**Datasets & Evaluation.** We conduct experiments on LaMP-QA [28], a benchmark designed for personalized QA. LaMP-QA spans three domains: Art & Entertainment, Lifestyle & Personal Development, and Society & Culture.<sup>1</sup> Each instance is accompanied by a user input query and narrative, with the user’s historical questions serving as the user profile. For evaluation, following [28], we evaluate responses from a personalized perspective by measuring how well they satisfy user-specific rubrics. We use the instruction-tuned Qwen 2.5-32B [41] with temperature 0.0 as the judge LLM. For each response, the LLM scores each personalized rubric aspect on a scale from 0 to 2. The scores are then normalized to  $[0, 1]$ , and the final score is computed as the average across all aspects. This produces a scalar score measuring alignment with user preferences, enabling fine-grained evaluation of personalized answer quality. For details on the evaluation, we refer the reader to Salemi and Zamani [28].

**Baselines.** We compare  $PR^2$  against diverse baselines spanning different modeling perspectives. These include (i) a non-personalized baseline, **No Personalization**, following [28], which utilizes CoT to generate responses without incorporating user information, (ii) reasoning-oriented retrieval methods such as **Search-R1** [14], and (iii) personalization-enhanced methods for reasoning and QA, including **RAG-Personalization** following method in [28], **PPlug** [21], **HYDRA** [48], **PrLM** [43], **VAC** [29], and **PlanPers** [28]. All baselines share the same retriever, corpus, LLM, temperature (1.0), and token limit (2048).

**Training Configuration.** We conduct experiments using Qwen-2.5-3B, Qwen-2.5-7B, and Gemma3-4B Base models. For retrieval, we employ Contriever [18], a dense retriever fine-tuned on MS MARCO (msmarco-bert-base-dot-v5) [23]<sup>2</sup>, together with FAISS for efficient indexing [9]. For each query, we retrieve the top  $k = 3$  from document from user profile; ablations on  $k$  are in Section §5. Training is performed on the training split of the LaMP-QA benchmark [28]. We optimize all models for 200 steps using a fixed rollout budget of 5 per instance, following [43]. We use the vLLM<sup>3</sup> with temperature 1.0 for efficient response generation and rollout collections. We train with a policy learning rate of  $1 \times 10^{-6}$  using the TRL GRPO implementation<sup>4</sup> and a warm-up ratio of 0.285. The KL-divergence regularization coefficient  $\beta$  and the clipping parameter

<sup>1</sup>For brevity, in this paper, we refer to *Art & Entertainment* as “Art,” *Lifestyle & Personal Development* as “Lifestyle,” and *Society & Culture* as “Society.”

<sup>2</sup>Available at: <https://hf.co/sentence-transformers/msmarco-bert-base-dot-v5>

<sup>3</sup><https://github.com/vllm-project/vllm>

<sup>4</sup>Available at: [https://hf.co/docs/trl/en/grpo\\_trainer](https://hf.co/docs/trl/en/grpo_trainer)

**Table 1: Performance on the LAMP-QA test set. The superscript \* denotes statistically significant improvements compared to all baselines with Bonferroni correction ( $p$ -value < 0.01).**

Method	Qwen2.5-3B				Qwen2.5-7B				Gemma3-4B			
	Art	Lifestyle	Society	Avg.	Art	Lifestyle	Society	Avg.	Art	Lifestyle	Society	Avg.
No personalization	0.2503	0.3511	0.4073	0.3362	0.2981	0.4014	0.4352	0.3782	0.2917	0.3871	0.4462	0.3750
Search R1-GRPO	0.3411	0.4691	0.5140	0.4414	0.3668	0.4939	0.5416	0.4674	0.3527	0.4731	0.5266	0.4508
RAG-Personalization	0.2753	0.3728	0.4282	0.3588	0.3086	0.4107	0.4590	0.3928	0.2887	0.3965	0.4234	0.3695
PPlug	0.3374	0.4522	0.5077	0.4324	0.3488	0.4773	0.5301	0.4520	0.3459	0.4362	0.5141	0.4320
HYDRA-Adapter	0.2942	0.4504	0.5211	0.4219	0.3612	0.4786	0.5319	0.4572	0.3271	0.4625	0.5273	0.4389
PlanPers	0.3206	0.4619	0.5234	0.4353	0.3518	0.4818	0.5240	0.4525	0.3382	0.4731	0.5296	0.4469
VAC	0.3193	0.4832	0.5311	0.4445	0.3454	0.5288	0.5351	0.4691	0.3418	0.4769	0.5325	0.4504
PrLM	0.3228	0.4762	0.5276	0.4422	0.3849	0.4808	0.5422	0.4693	0.3266	0.4817	0.5301	0.4461
PR <sup>2</sup>	0.3751*	0.4924*	0.5761*	0.4812*	0.4265*	0.5519*	0.5983*	0.5256*	0.3855*	0.5271*	0.5732*	0.4952*

$\epsilon$  are fixed to 0.001 and 0.2, respectively, across all experiments. All experiments use NVIDIA A100 GPUs (80GB VRAM, 128GB RAM) with a maximum 2048-token input–output budget.

## 5 Results and Discussion

**Comparison with the Baselines.** Table 1 compares PR<sup>2</sup> with a diverse set of baselines. Across all settings, PR<sup>2</sup> consistently achieves the highest performance with statistical significance testing with Bonferroni correction [4]. For Qwen2.5-3B, PR<sup>2</sup> achieves an average score of 0.4812, outperforming the strongest baseline, VAC (0.4445), by a 8.26% relative gain. For Qwen2.5-7B, PR<sup>2</sup> achieves an average score of 0.5256, outperforming the strongest baseline, PrLM (0.4693), by a 12.0% relative gain. For Gemma3-4B, PR<sup>2</sup> reaches an average score of 0.4952, exceeding the strongest baseline, Search-R1 (0.4508) by a 9.85% relative gain. These consistent gains demonstrate the effectiveness of PR<sup>2</sup>, which trains the policy to leverage reasoning and retrieval from the user profile to access more useful personal information for each question. A case study example of PR<sup>2</sup>'s output is presented below.

**Question:** How many papers will be expected to be published as a postdoc?

**Response:** <think> To provide a personalized answer, the specific field of study and institution of the postdoc would be required. </think>  
<search> - Field of study, Name of the institution </search>  
<information> [1: 'Incompatibility between job and field of study Finding jobs for Ph.D graduate in mathematics is the main concern. Available options are namely in general industry and academia. I see, for instance, some Ph.D in math working in Bank sector! This incompatibility between field of study and job might back to the rank of math department. But I want to know about the upper bound of ranks of departments might be considered as top ranks? Additionally, which university ranking is most reliable in industry sector and in academia? 2: 'What are diplomas, postgraduate courses, specialization, training courses? I am going to apply for some faculty positions.... I have already attended several summer schools, workshops and also took online courses. Can these events be considered as one of the elements of mentioned category?' 3: 'What is the meaning of pre-doctoral program?... Can someone be a predoctoral in university X and then goes to university to start Ph.D?'] </information>  
<think> The postdoc's question about expected publications depends heavily on the field of study and the institution. In this case, the field is mathematics, and there is no information available about the institution. </think>  
<answer> As a postdoc, the number of publications expected can vary significantly depending on both the field and institutional norms. In mathematics, a common expectation may be around one to two publications per year, although this can vary widely based on the institution, the complexity of the research, and the pace of progress. It is important to align publication goals with the expectations of the postdoctoral program and the guidance of one's advisor or mentor to ensure that appropriate benchmarks are being met. </answer>

**Table 2: PR<sup>2</sup> (Qwen2.5-7B) performance with different number of retrieved documents.**

# docs	Art	Lifestyle	Society	Avg.
1	0.3866	0.4468	0.5441	0.4592
3	0.4265	0.5519	0.5983	0.5256
5	0.3918	0.4602	0.5483	0.4668

**Training Dynamics of PR<sup>2</sup>.** We analyze PR<sup>2</sup>'s GRPO training dynamics with all LLMs. As shown in Figure 2a, training reward increases steadily for all models, with improvements continuing throughout training and no signs of reward collapse or instability in 200 steps. This trend suggests that GRPO effectively optimizes PR<sup>2</sup>'s policy for personalization-aware retrieval and reasoning, and Qwen2.5-7B consistently reaches the highest reward values, indicating a stronger capacity to leverage retrieved user-specific evidence during reasoning. Figure 2b reports the average number of retrieval actions. Early in training, all models perform roughly one retrieval on average, but retrieval frequency increases over time, indicating that the LLM progressively learns to retrieve more frequently as training proceeds. Qwen2.5-7B retrieves the most, whereas Gemma3-4B and Qwen2.5-3B retrieve fewer documents, supporting the interpretation that larger models better exploit additional user-specific retrieval for personalized reasoning. Finally, Figure 2c shows a structured progression in response length, exhibiting a decrease–increase–stabilize pattern. During the first 100 steps, response length decreases while reward improves. After the first 100 steps, response length increases with reward gains, indicating the LLM more frequently invokes the retriever and incorporates retrieved content. The rising reward suggests that the model becomes more effective at leveraging retrieved content to personalize answers. Across models, Qwen2.5-7B produces the longest responses, 3B the shortest, with Gemma3-4B in between.

**Effect of Non-personalized Baseline in Advantage Computation.** Table 3 examines the role of including non-personalized baseline in training. Removing the baseline reduces performance: Qwen2.5-3B drops from 0.4812 to 0.4387, Qwen2.5-7B from 0.5256

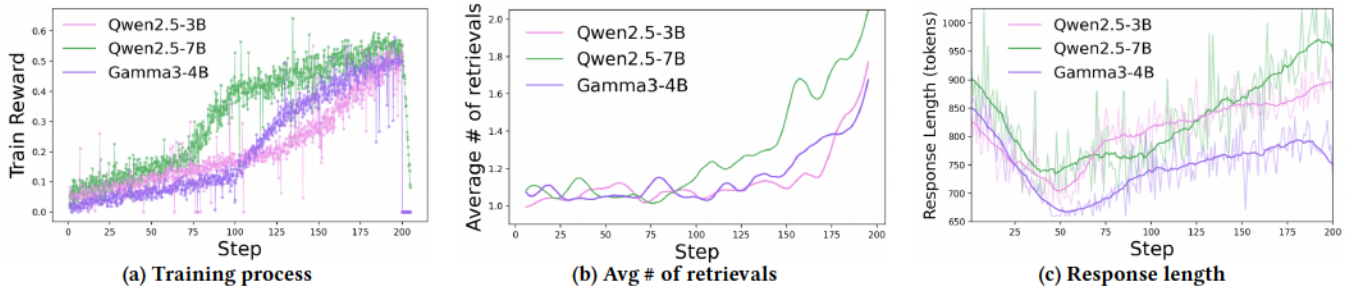


Figure 2: Model behavior during training on the LaMP-QA benchmark. The plots depict (a) the reward convergence during training, (b) the average number of retrieval requests (search queries), and (c) the average length of generated responses.

Table 3: Ablation of advantage function in PR<sup>2</sup>.

Model	Method	Art	Lifestyle	Society	Avg.
Qwen2.5-3B	w. non-personalized rollout	0.3751	0.4924	0.5761	0.4812
	w.o non-personalized rollout	0.3345	0.4476	0.5342	0.4387
Qwen2.5-7B	w. non-personalized rollout	0.4265	0.5519	0.5983	0.5256
	w.o non-personalized rollout	0.3706	0.4862	0.5641	0.4736
Gemma3-4B	w. non-personalized rollout	0.3855	0.5271	0.5732	0.4952
	w.o non-personalized rollout	0.3472	0.4793	0.5117	0.4461

to 0.4736, and Gemma3-4B from 0.4952 to 0.4461. These results show that incorporating non-personalized baseline alongside personalized sampling consistently improves performance, showing that they are a critical component of PR<sup>2</sup>'s learning process. By contrasting personalized and non-personalized trajectories under the same reward objective, the model learns when retrieval meaningfully enhances personalization rather than overusing retrieval or relying on noisy personalized signals. This contrast allows PR<sup>2</sup> to better attribute improvements to the integration of retrieved evidence rather than to generic generation gains. Without it, the policy is less capable of distinguishing when retrieval from profile improves generation, resulting in weaker learned behavior.

**Effect of the Number of Retrieved User-Profile Documents .** We analyze how the number of retrieved user-profile documents affects the performance of PR<sup>2</sup>. Following [43], our main setting uses top- $k = 3$ , and we additionally evaluate top- $k = 1$  and top- $k = 5$ . As shown in Table 2 for Qwen2.5-7B, top- $k = 3$  yields the best overall performance (0.5256), outperforming both top- $k = 1$  (0.4592) and top- $k = 5$  (0.4668). This suggests a trade-off between retrieval sufficiency and noise: retrieving too few documents may limit recall of relevant user-profile information and thus constrain the model's ability to capture personalized context during reasoning, while retrieving too many documents may introduce weakly relevant or redundant information that reduces precision and distracts the reasoning process.

## 6 Conclusion

This paper introduces PR<sup>2</sup>, an RL framework that integrates retrieval directly into the reasoning process for personalized QA. Unlike prior studies that focus primarily on retrieval without explicitly

modeling the reasoning process, PR<sup>2</sup> jointly optimizes retrieval and reasoning decisions under personalized reward signals. By training with GRPO and contrasting personalized and non-personalized rollouts in advantage calculation, the framework learns when retrieval is necessary and how retrieved user-specific evidence should shape reasoning steps. Experiments on LaMP-QA show consistent gains across multiple LLMs, demonstrating that aligning retrieval and reasoning with user-specific objectives improves personalized QA.

## Acknowledgment

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #2402873, in part by the Office of Naval Research contract #N000142412612, and with support from by Google.org. 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] Pranjal Aggarwal and Sean Welleck. 2025. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning. In *Second Conference on Language Modeling*. <https://openreview.net/forum?id=4jdlxXBNve>
- [2] Maryam Amirizani, Elias Martin, Afra Mashhadi, and Chirag Shah. 2025. Response Generation through Social Reasoning in Large Language Models with Direct Diverse Preferences Optimization. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (Padua, Italy) (ICTIR '25)*. Association for Computing Machinery, New York, NY, USA, 315–325. <https://doi.org/10.1145/3731120.3744598>
- [3] Maryam Amirizani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. 2024. Can LLMs Reason Like Humans? Assessing Theory of Mind Reasoning in LLMs for Open-Ended Questions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (Boise, ID, USA) (CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 34–44. <https://doi.org/10.1145/3627673.3679832>
- [4] Richard A Armstrong. 2014. When to use the B onferoni correction. *Ophthalmic and physiological optics* 34, 5 (2014), 502–508.

- [5] Millennium Bismay, Xiangjue Dong, and James Caverlee. 2025. ReasoningRec: Bridging Personalized Recommendations and Human-Interpretable Explanations through LLM Reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 8132–8148. <https://doi.org/10.18653/v1/2025.findings-naacl.454>
- [6] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. Towards Medical Complex Reasoning with LLMs through Medical Verifiable Problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 14552–14573. <https://doi.org/10.18653/v1/2025.findings-acl.751>
- [7] Eunjung Cho, Alexander Hoyle, and Yoan Hermstrüwer. 2025. Modeling Motivated Reasoning in Law: Evaluating Strategic Role Conditioning in LLM Summarization. In *Proceedings of the Natural Language Processing Workshop 2025*, Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preotiuc-Pietro, and Gerasimos Spanakis (Eds.). Association for Computational Linguistics, Suzhou, China, 68–112. <https://doi.org/10.18653/v1/2025.nllp-1.7>
- [8] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering ChatGPT’s Capabilities in Recommender Systems. In *Proceedings of the 17th ACM Conference on Recommender Systems (Singapore, Singapore) (RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1126–1132. <https://doi.org/10.1145/3604915.3610646>
- [9] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data* (2025).
- [10] Wengqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Barcelona, Spain) (KDD '24)*. Association for Computing Machinery, New York, NY, USA, 6491–6501. <https://doi.org/10.1145/3637528.3671470>
- [11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirog Zhu, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).
- [12] Shayekh Bin Islam, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024. Open-RAG: Enhanced Retrieval Augmented Reasoning with Open-Source Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 14231–14244. <https://doi.org/10.18653/v1/2024.findings-emnlp.831>
- [13] Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2025. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. In *The Thirtieth International Conference on Learning Representations*. <https://openreview.net/forum?id=oU3tpa8f8m>
- [14] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. In *Second Conference on Language Modeling*. <https://openreview.net/forum?id=Rwhi91ideu>
- [15] Jieyong Kim, Maryam Amirizani, Soojin Yoon, and Dongha Lee. 2025. IPQA: A Benchmark for Core Intent Identification in Personalized Question Answering. *arXiv preprint arXiv:2510.23536* (2025).
- [16] Jaehyung Kim and Yiming Yang. 2025. Few-shot Personalization of LLMs with Mis-aligned Responses. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 11943–11974. <https://doi.org/10.18653/v1/2025.naacl-long.598>
- [17] Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016* (2024).
- [18] Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 10932–10940. <https://doi.org/10.18653/v1/2023.findings-acl.695>
- [19] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic Search-Enhanced Large Reasoning Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 5420–5438. <https://doi.org/10.18653/v1/2025.emnlp-main.276>
- [20] Jiacheng Lin, Tian Wang, and Kun Qian. 2025. Rec-R1: Bridging Generative Large Language Models and User-Centric Recommendation Systems via Reinforcement Learning. *Transactions on Machine Learning Research* (2025). <https://openreview.net/forum?id=YBRU9MV2vE>
- [21] Jiongan 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)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 9373–9385. <https://doi.org/10.18653/v1/2025.acl-long.461>
- [22] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (Merida, Mexico) (WSDM '24)*. Association for Computing Machinery, New York, NY, USA, 452–461. <https://doi.org/10.1145/3616855.3635845>
- [23] 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. (2016).
- [24] 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. <https://doi.org/10.1145/3626772.3657783>
- [25] Alireza Salemi, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Zhuowan Li, Spurthi Amba Hombaiah, Weize Kong, Tao Chen, Hamed Zamani, and Michael Bendersky. 2025. Pathways of Thoughts: Multi-Directional Thinking for Long-form Personalized Question Answering. *arXiv preprint arXiv:2509.19094* (2025).
- [26] 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. <https://doi.org/10.18653/v1/2024.acl-long.399>
- [27] Alireza Salemi and Hamed Zamani. 2024. Evaluating Retrieval Quality in Retrieval-Augmented Generation. 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, 2395–2400. <https://doi.org/10.1145/3626772.3657957>
- [28] 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*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 1139–1159. <https://doi.org/10.18653/v1/2025.emnlp-main.60>
- [29] Alireza Salemi and Hamed Zamani. 2025. Learning from natural language feedback for personalized question answering. *arXiv preprint arXiv:2508.10695* (2025).
- [30] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).
- [31] Heydar Soudani, Hamed Zamani, and Faegheh Hasibi. 2025. Uncertainty Quantification for Retrieval-Augmented Reasoning. *arXiv preprint arXiv:2510.11483* (2025).
- [32] Chenka Sun, Ke Yang, Revanth Gangi Reddy, Yi Fung, Hou Pong Chan, Kevin Small, Chengxiang Zhai, and Heng Ji. 2025. Persona-DB: Efficient Large Language Model Personalization for Response Prediction with Collaborative Data Refinement. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 281–296. <https://aclanthology.org/2025.coling-main.20/>
- [33] Richard S Sutton, Andrew G Barto, et al. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.
- [34] Hieu Tran, Zonghai Yao, Zhichao Yang, Junda Wang, Yifan Zhang, Shuo Han, Feiyun Ouyang, and Hong Yu. 2025. RARE: Retrieval-Augmented Reasoning Enhancement for Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 18305–18330. <https://doi.org/10.18653/v1/2025.acl-long.896>
- [35] Alicia Tsai, Adam Kraft, Long Jin, Chenwei Cai, Anahita Hosseini, Taibai Xu, Zemin Zhang, Lichan Hong, Ed H. Chi, and Xinyang Yi. 2024. Leveraging LLM Reasoning Enhances Personalized Recommender Systems. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok,

- Thailand, 13176–13188. <https://doi.org/10.18653/v1/2024.findings-acl.780>
- [36] Jianling Wang, Haokai Lu, Yifan Liu, He Ma, Yueqi Wang, Yang Gu, Shuzhou Zhang, Ningren Han, Shuchao Bi, Lexi Baugher, Ed H. Chi, and Minmin Chen. 2024. LLMs for User Interest Exploration in Large-scale Recommendation Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems* (Bari, Italy) (*RecSys '24*). Association for Computing Machinery, New York, NY, USA, 872–877. <https://doi.org/10.1145/3640457.3688161>
- [37] Lei Wang and Ee-Peng Lim. 2023. Zero-shot next-item recommendation using large pretrained language models. *arXiv preprint arXiv:2304.03153* (2023).
- [38] Jason Wei, Xuezi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). [https://openreview.net/forum?id=\\_VjQlMeSB\\_J](https://openreview.net/forum?id=_VjQlMeSB_J)
- [39] Jiaxing Wu, Lin Ning, Luyang Liu, Harrison Lee, Neo Wu, Chao Wang, Sushant Prakash, Shawn O'Banion, Bradley Green, and Jun Xie. 2025. Rlpf: Reinforcement learning from prediction feedback for user summarization with llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 25488–25496.
- [40] 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).
- [41] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jiahong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).
- [42] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. LIMO: Less is More for Reasoning. In *Second Conference on Language Modeling*. <https://openreview.net/forum?id=TzTZ0RY4Zk>
- [43] 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* (Seoul, Republic of Korea) (*CIKM '25*). Association for Computing Machinery, New York, NY, USA, 5484–5488. <https://doi.org/10.1145/3746252.3760851>
- [44] 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. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 1586–1598. <https://doi.org/10.18653/v1/2025.findings-emnlp.84>
- [45] Kepu Zhang, Haoyue Yang, Xu Tang, Weijie Yu, and Jun Xu. 2025. Beyond Guilt: Legal Judgment Prediction with Trichotomous Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 1815–1826. <https://doi.org/10.18653/v1/2025.findings-emnlp.95>
- [46] Kepu Zhang, Weijie Yu, Zhongxiang Sun, and Jun Xu. 2025. SyLeR: A Framework for Explicit Syllagistic Legal Reasoning in Large Language Models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management* (Seoul, Republic of Korea) (*CIKM '25*). Association for Computing Machinery, New York, NY, USA, 4117–4127. <https://doi.org/10.1145/3746252.3761120>
- [47] Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do LLMs Recognize Your Preferences? Evaluating Personalized Preference Following in LLMs. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=QWunLKbBGF>
- [48] Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. Hydra: Model factorization framework for black-box llm personalization. *Advances in Neural Information Processing Systems* 37 (2024), 100783–100815.