STEP-BACK PROFILING: Distilling User History for Personalized Scientific Writing

Anonymous ACL submission

Abstract

Large language models (LLM) excel at a variety of natural language processing tasks, yet they struggle to generate personalized content for individuals, particularly in real-world scenarios like scientific writing. Addressing this challenge, we introduce STEP-BACK PROFILING to personalize LLMs by distilling user history into concise profiles, including essential traits and preferences of users. Regarding our experiments, we construct a Personalized Scientific Writing (PSW) dataset to study multi-user personalization. PSW requires the models to write scientific papers given specialized author groups with diverse academic backgrounds. As for the results, we demonstrate the effectiveness of capturing user characteristics via STEP-BACK PROFILING for collaborative writing. Moreover, our approach outperforms the baselines by up to 3.6 points on the general personalization benchmark (LaMP), including 7 personalization LLM tasks. Our extensive ablation studies validate the contributions of different components in our method and provide insights into our task definition. Our dataset and code will be available upon acceptance.

1 Introduction

Recently, Large Language Models (LLMs) have made significant progress in natural language understanding and generation (Wei et al., 2022a; Zhang et al., 2023b; OpenAI, 2023; Qin et al., 2023). Concurrently, integrating LLMs with personalization paradigms has paved the way for a vast frontier in improving user-centric services and applications (Salemi et al., 2023; Chen et al., 2023; Zhiyuli et al., 2023), as they provide a deeper understanding of users' accurate demands and interests than abstract vector-based information representations. By learning to characterize and emulate user-specific language patterns, personalized LLMs can enable more engaging and

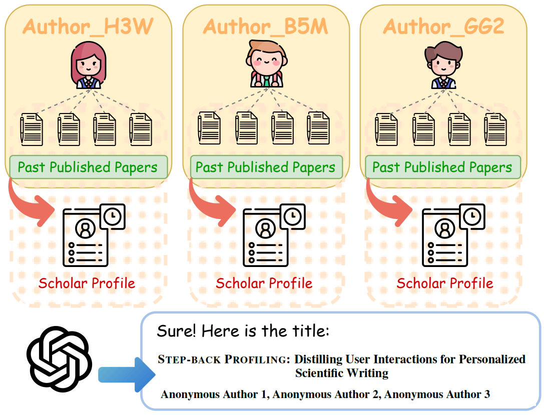

Figure 1: Overview of the STEP-BACK PROFILING.

valuable interactions in domains such as dialogue (Wang et al., 2019; Zhang et al., 2019b; Character.AI, 2022), recommendation (Zhiyuli et al., 2023; Wang et al., 2023a), role-playing (Shao et al., 2023; Jiang et al., 2023) and content creation (Cao et al., 2023; Wei et al., 2022b).

Prior work on personalizing language models (Salemi et al., 2023; Tan and Jiang, 2023; Zhang et al., 2023a; Chen et al., 2023; Zhiyuli et al., 2023) has shown promise, but primarily focused on learning user representations in a single-user context. However, many real-world applications involve multiple users collaborating on a shared task, such as team-authored scientific papers. Another practical challenge for LLM personalization is scaling to extensive user histories while respecting context length limits (Shi et al., 2023; Liu et al., 2024; Zhang et al., 2024). Directly conditioning on raw personal histories quickly becomes infeasible as user data grows. Prior methods mostly use uncompressed history for personalization (Salemi et al., 2023), which restricts the amount of user-specific information the model can utilize.

As shown in Figure 1, this work proposes a training-free LLM personalization framework that addresses these challenges through STEP-BACK

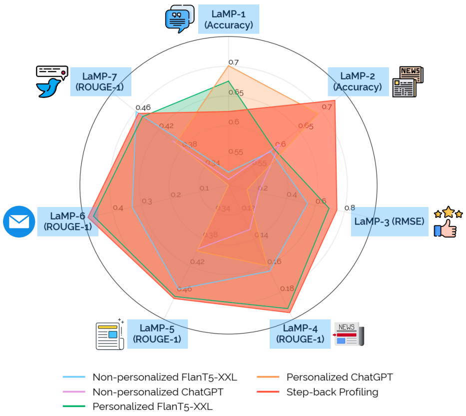

Figure 2: **STEP-BACK PROFILING performance on the LaMP benchmark.** Details of experimental setup can be found in Section 4.2.

PROFILING, we distill individual user histories into concise profile representations that capture high-level concepts and language traits. This enables efficient memory management and allows the model to focus on salient user characteristics, grounding personalized generation without excess computation or laborious data collection (Chen et al., 2023). We show that STEP-BACK PROFILING improves performance over standard personalization methods (retrieval-based) in the LaMP[1], as shown in Figure 2. Moreover, we introduce a Personalized Scientific Writing (PSW) dataset to study multi-user personalization. PSW contains research papers collaboratively written by expert teams, and each author's background publications are used to construct profiles. Modeling a group's collective expertise is crucial for this task, as different paper sections may reflect knowledge associated with particular authors. PSW thus poses a challenging and realistic testbed for multi-user personalization, requiring both abstractions of individual expertise and dynamic integration of diverse user traits throughout the collaborative writing process.

2 STEP-BACK PROFILING

2.1 Motivation

Existing methods for personalizing language models struggle to effectively utilize user histories, particularly in the presence of extraneous details that can obscure the most pertinent information for a given task (Shi et al., 2023; Liu et al., 2024). This challenge is magnified in multi-user scenarios, where models must efficiently extract and integrate knowledge from multiple users' histories. While retrieval-augmented methods, such as those employed in the LaMP benchmark (Salemi et al., 2023), have made progress in scaling to more extensive user histories, they still operate on raw user data containing relevant and irrelevant details. To address these limitations, we introduce a STEP-BACK PROFILING approach that distills a user's raw history into a concise representation focusing on 'gist' representations and preferences. Our approach aims to enable more efficient and effective personalization across diverse single and multi-user scenarios by reasoning about higher-level traits instead of verbatim user history.

[1]https://lamp-benchmark.github.io/

2.2 Procedure

Consider a set of n users denoted by $\mathbb{U} = \{u_i\}_{i=1}^{n}$, where each user u_i has a preference history $\mathbb{H}_i = \{(x_{ij}, y_{ij})\}_{j=1}^{m}$ consisting of m input-output pairs. To effectively generate $P(y|x, \mathbb{H}_{\mathbb{U}})$ based on users' preference history, we create a set of user profiles $\mathbb{P}_{\mathbb{U}} = \{\mathbb{P}_{u_i} | u_i \in \mathbb{U}\}$ using STEP-BACK PROFILING. The complete procedure involves the following steps:

User Profile Gisting: Each user's history is condensed into a short "gist" representation using an abstraction function $\mathrm{Gist}(\cdot)$: $\mathbb{P}_{u_i} = \mathrm{Gist}(\mathbb{H}_i)$. The "gist" captures the user's high-level traits and interests.

Multi-User Profile Concatenation: Individual user profiles $\mathbb{P}_{u_1}, \mathbb{P}_{u_2}, \cdots, \mathbb{P}_{u_n}$ are concatenated to form a unified representation $\mathbb{P}_{\mathbb{U}} = [\mathbb{P}_{u_1}; \mathbb{P}_{u_2}; \cdots; \mathbb{P}_{u_n}]$, where $[\cdot;\cdot]$ is a permutation-sensitive function combining the user profiles.

Retrieval-Augmented Generation (Optional): Relevant snippets from user histories $\mathbb{H}_{\mathbb{U}}$ may be retrieved for input x using a retrieval function $\mathrm{Retrieve}(\cdot)$. We have $\mathbb{R}_{i,k} = \mathrm{Retrieve}(x, \mathbb{H}_i, k)$, where $\mathbb{R}_{i,k}$ is a set of top-k retrieved input-output snippets from user u_i's history $\mathbb{H}_i$. The top-k retrieved snippets $\mathbb{R}_k = \{\mathbb{R}_{i,k}\}_{i=1}^{N}$ can be concatenated with x to form an augmented input $\hat{x} = [x; \mathbb{R}_{1,k}; \mathbb{R}_{2,k}; \cdots; \mathbb{R}_{n,k}]$.

Personalized Output Generation: The personalized language model generates an output $y = \mathrm{Generate}(\hat{x}, \mathbb{P}_{\mathbb{U}})$ by conditioning on the augmented input $\hat{x}$ (if retrieval is used) or the original input x, along with the concatenated user profile $\mathbb{P}_{\mathbb{U}}$. The generated output y aligns with the user preferences captured by the STEP-BACK PROFILING while following the input x.

3 The Personalized Scientific Writing (PSW) Benchmark

3.1 Problem Formulation

Personalized language models aim to generate outputs that follow a given input and align with the users' styles, preferences, and expertise. In multi-author collaborative writing, each data entry in the PSW benchmark consists of four key components: (1) An input sequence x serves as the model's input; (2) A target output y that the model is expected to generate; (3) A set of user histories $\mathbb{H}_{\mathbb{U}} = \{\mathbb{H}_{u_i}\}_{i=1}^{l}$, where l is the number of collaborating authors, and each entry $\mathbb{H}_{u_i}$ contains historical input-output pairs for user u_i; (4) A set of author roles $\mathbb{C} = \{c_i\}_{i=1}^{l}$, each representing the role of the corresponding author u_i in the collaborative writing process.

A personalized language model aims to generate an output y that aligns with the conditional probability distribution $P(y|x, \mathbb{H}_{\mathbb{U}}, \mathbb{C})$. This means the model should produce an output that follows the input x and the collaborating authors' writing styles, preferences, and expertise, as captured by their user histories $\mathbb{H}_{\mathbb{U}}$ and roles $\mathbb{C}$.

3.2 Dataset Construction

The dataset encompasses one individual task, **User Profiling** (UP-0), which involves compiling a list of research interests based on their publication history. The label is extracted from Google Scholar to accurately reflect each author's expertise by searching their name. Additionally, it includes four collaborative tasks: **Research Topics Generation** (PSW-1) using OpenAI's GPT-4 to derive relevant topics from selected papers; **Research Question Generation** (PSW-2) using GPT-4 for research question extraction; **Paper Abstract Generation** (PSW-3) by retrieving abstracts through the Semantic Scholar API; **Paper Title Generation** (PSW-4), which gathers data from the Semantic Scholar API to create suitable paper titles. Details of the dataset construction can be found in Appendix B.1

3.3 GPT-based Evaluation

LLM-based evaluators, such as G-Eval, have shown high consistency with human evaluators (Liu et al., 2023; Chang et al., 2024), particularly in personalized text generation (Wang et al., 2023b). Therefore, we utilize GPT-4-turbo with chain-of-thought prompting as a judge to evaluate the generated outputs on the PSW benchmark in multiple dimensions (Zhang et al., 2019a), including consistency, fluency, relevance, and novelty. An example of our evaluation (G-Eval) prompt can be found in Appendix D.

4 Experimental Setup

4.1 Datasets and Evaluation

LaMP Dataset We follow the established LAMP benchmark Salemi et al. (2023), encompassing three classification and four text generation tasks. Specifically, these tasks are Personalized Citation Identification (LaMP-1), Personalized News Categorization (LaMP-2), Personalized Product Rating (LaMP-3), Personalized News Headline Generation (LaMP-4), Personalized Scholarly Title Generation (LaMP-5), Personalized Email Subject Generation (LaMP-6), and Personalized Tweet Paraphrasing (LaMP-7).

PSW Dataset The dataset includes one individual task, User Profiling (UP-0), and four collaborative tasks: Research Topics Generation (PSW-1), Research Question Generation (PSW-2), Paper Abstract Generation (PSW-3), and Paper Title Generation (PSW-4).

Our evaluation follows the LaMP (Salemi et al., 2023) and we employ the metrics specified in the LaMP for each task. Those include F1, Accuracy, MAE, and RMSE for classification tasks and ROUGE-1 and ROUGE-L for generation tasks. We compare several baselines, including non-personalized language models, models fine-tuned on history data without personalization, and models that use a simple concatenation of user histories for personalization with retrieval models.

4.2 Main Result

LaMP Results To ensure a fair comparison, we utilize a user-based separation from LaMP (Salemi et al., 2023). We only grant the model access to the provided user history and restrict it from accessing any other information. Additionally, we utilize *the same pre-trained retriever in LaMP baselines*, without any additional fine-tuning, to retrieve the top five examples. This approach is identical to the *Non-Personalized* setting in (Salemi et al., 2023). Finally, we compare our results with the outcomes reported in the study.

As shown in Table 1, our analysis unveils a notable performance enhancement through our method's application, significantly when leveraging the same backbone language models (*GPT-*

Dataset	Metric	Non-personalized FlanT5-XXL[†]	Non-personalized ChatGPT[†]	Personalized[§] FlanT5-XXL[†]	Personalized[§] ChatGPT[†]	Ours
LaMP-1	Accuracy	0.522	0.510	0.675	**0.701**	0.624
LaMP-2	Accuracy	0.591	0.610	0.598	0.693	**0.729**
	F1	0.463	0.455	0.477	0.455	**0.591**
LaMP-3	MAE	0.357	0.699	0.282	0.658	**0.274**
	RMSE	0.666	0.977	0.584	1.102	**0.559**
LaMP-4	ROUGE-1	0.164	0.133	0.192	0.160	**0.195**
	ROUGE-L	0.149	0.118	0.178	0.142	**0.180**
LaMP-5	ROUGE-1	0.455	0.395	0.467	0.398	**0.469**
	ROUGE-L	0.410	0.334	0.424	0.336	**0.426**
LaMP-6	ROUGE-1	0.332	-	0.466	-	**0.485**
	ROUGE-L	0.320	-	0.453	-	**0.464**
LaMP-7	ROUGE-1	**0.459**	0.396	0.448	0.391	0.455
	ROUGE-L	**0.404**	0.337	0.396	0.324	0.398

Table 1: **Performance comparison of models on the LaMP dataset.** [†]Baseline results are obtained directly from (Salemi et al., 2023). [§] Personalized means we use retrieval modules before LLMs.

3.5-turbo). It is clear that our "gist"-style information compression is much more necessary than retrieval methods as the comparisons in Table 1. In the domain of text generation tasks (LaMP-4∼7), our method achieves an average improvement of 0.048 in Rouge-1 and 0.053 in Rouge-L, corresponding to gains of 15.2% and 19.5%, respectively. Similarly, for the classification tasks (LaMP-1∼3), we observe an average +12.6% accuracy gain of and a +42.5% reduction in MAE compared to the *Non-Personalized* setting. Our method continues to exhibit better performance across most tasks, even when compared with **`FlanT5-XXL`**, with a fine-tuned retriever as *Personalized* setting. The prompt used in this experiment is detailed in Appendix E.

Datasets	Method	ROUGE-1	ROUGE-L	Consistency	Fluency	Relevance	Novelty
UP-0	`Single-Author`	0.267	0.233	4.32	2.01	3.59	/
PSW-1	`Zero-shot`	0.306	0.257	3.43	**2.65**	3.53	2.30
	`Single-Author`	0.325	0.266	3.44	2.47	3.61	2.59
	`Multi-Author`	**0.337**	**0.280**	**3.59**	2.58	**3.67**	**2.63**
PSW-2	`Zero-shot`	0.196	0.179	4.31	2.04	3.89	2.21
	`Single-Author`	0.190	0.171	4.20	2.23	3.67	2.01
	`Multi-Author`	**0.201**	**0.186**	**4.60**	**2.39**	**3.91**	**2.38**
PSW-3	`Zero-shot`	0.099	0.094	4.43	2.81	4.43	2.40
	`Single-Author`	0.131	0.124	**4.94**	**2.94**	4.70	2.40
	`Multi-Author`	**0.145**	**0.131**	4.92	**2.94**	**4.71**	**2.45**
PSW-4	`Zero-shot`	0.459	0.391	4.41	2.41	3.58	2.38
	`Single-Author`	0.472	0.409	4.59	2.49	3.78	2.60
	`Multi-Author`	**0.505**	**0.444**	**4.64**	**2.59**	**3.79**	**2.64**

Table 2: **Performance comparison of personalized models on the PSW dataset**. We report additional metrics such as **Consistency (1-5)**, **Fluency (1-3)**, **Relevance (1-5)**, and **Novelty (1-3)**.

PSW Results: We then evaluate our proposed model using the PSW dataset, focusing on user profiling (UP-0), personalized idea brainstorming (PSW-1, PSW-2), and personalized text generation (PSW-3, PSW-4) in three different settings:

1. **`Zero-shot`**: Generates outputs based on the input prompt x alone: $y = \text{Generate}(x)$.
2. **`Single-Author`**: Personalizes with single user's profile P_{u_i} and retrieved snippets R_i: $y = \text{Generate}(\hat{x}, P_{u_i})$, where $\hat{x} = [x; \mathbb{R}_i]$ and $\mathbb{R}_i = \text{Retrieve}(x, \mathbb{H}_i, 10)$.
3. **`Multi-Author`**: Personalizes with multiple users' profiles $\mathbb{P}_{\mathbb{U}}$ and retrieved snippets $\mathbb{R}$: $y = \text{Generate}(\hat{x}, \mathbb{P}_{\mathbb{U}})$, where $\hat{x} = [x; \mathbb{R}_1; \cdots; \mathbb{R}_n]$, $\mathbb{R}_i = \text{Retrieve}(x, \mathbb{H}_i, 10)$ for each user u_i.

As shown in Table 2, our **`Multi-Author`** setting demonstrates superior performance across all tasks. In PSW-1 and PSW-2, the **`Multi-Author`** setting outperforms both **`Zero-shot`** and **`Single-Author`** settings, with an average improvement of +6.9% in ROUGE-1 and +7.1% in ROUGE-L. Similarly, for the PSW-3 and PSW-4, the **`Multi-Author`** setting achieves the highest ROUGE scores, with an average gain of +28.2% in ROUGE-1 and +26.6% in ROUGE-L, compared to the **`Zero-shot`** and **`Single-Author`** settings. Furthermore, the **`Multi-Author`** setting exhibits the highest scores for additional metrics such as Consistency, Fluency, Relevance, and Novelty across all tasks, with an average improvement of +5.1%, +6.7%, +3.8%, and +6.4%, respectively, compared to the **`Zero-shot`** and **`Single-Author`** setting. The prompt used in this experiment is detailed in Appendix F. Finally, to evaluate the contribution of each component, we perform an ablation study in Appendix A, which reports the results when: 1) Switching the order of users and 2) Removing user profiling.

5 Conclusion

In summary, we introduce a training-free technique, STEP-BACK PROFILING, for personalizing large language models by distilling user interactions using gist into concise profiles. Moreover, we extend the LaMP dataset into the Personalized Scientific Writing (PSW) dataset to evaluate multi-user scenarios in collaborative scientific writing. Our experiments show that the proposed method is effective on the LaMP and PSW datasets. In particular, both single-user and multi-user settings validate the benefits of profile-guided personalization. Finally, studying the interpretability and controllability of profile-guided models can help build user trust and allow for more fine-grained customization.

Limitation

Our proposed STEP-BACK PROFILING framework has a few limitations that warrant discussion and could be addressed in future work:

Dataset Specificity The experiments and results presented are primarily based on the Personalized Scientific Writing (PSW) dataset and the LaMP benchmark. While these datasets provide a diverse set of tasks, the performance and applicability of the STEP-BACK PROFILING framework may vary with different datasets or domains not covered by our experiments. Future work should evaluate the model on more varied datasets to ensure generalizability.

Complexity of Profiles The profile generation process involves distilling user histories into concise representations. While this method captures essential traits, it may oversimplify user preferences and neglect nuanced behaviors present in longer and more complex histories. More sophisticated profiling techniques that can retain and effectively compress these complexities are needed.

Scalability and Efficiency Although the STEP-BACK PROFILING method improves memory management, the approach still has scalability concerns, particularly with very large user histories or an increasing number of collaborators. Efficiently managing and retrieving relevant user data from extensive histories without compromising performance remains a challenge.

Dynamic Adaptation The current method creates static profiles based on available user histories at a given time. However, user preferences and styles may evolve, especially in dynamic collaborative environments. Developing a mechanism to update profiles dynamically based on real-time user interactions and feedback could further enhance the personalization capabilities.

Evaluation Metrics The evaluation relies heavily on established metrics such as ROUGE and human-aligned scoring via G-Eval, which, while comprehensive, may not capture all dimensions of personalized content quality. Developing and employing more specialized evaluation metrics for personalized content generation, particularly in scientific and collaborative writing, would provide deeper insights into the effectiveness of the methods.

Human Factors: Although tools like GPT-4 mitigate the involvement of human evaluation, it is inherently subjective. Future work should consider more robust and unbiased methods of human evaluation to validate the effectiveness of personalized outputs objectively.

Ethical and Privacy Concerns Personalizing models using user histories raises potential ethical and privacy issues. It is crucial to ensure that user data is handled securely and that privacy concerns are adequately addressed. Future research should explore more privacy-preserving techniques for personalization, such as federated learning.

Adapting STEP-BACK PROFILING to long histories spanning multiple sessions is another valuable direction. Future work can explore more advanced profiling strategies, such as hierarchical representations and dynamic profile updates based on user feedback.

Ethical Statement

Dataset Licensing We have constructed the Personalized Scientific Writing (PSW) dataset, which will be publicly released under the MIT license. This permissive license allows users to freely use, modify, and distribute the dataset. By releasing the PSW dataset under the MIT license, we aim to promote transparency, reproducibility, and wide adoption of our research within the community.

Artifact Use Consistent With Intended Use Regarding our use of existing artifacts, we have ensured that our usage is consistent with their intended purposes, as specified by their creators. For the artifacts we create, including the PSW dataset, we specify that the intended use is for research purposes. This is compatible with the original access conditions of any derivative data we utilized. Derivative data accessed for research purposes should not be used outside of research contexts.

Personally-Identifying Info We acknowledge that the PSW dataset construction involved the use of researchers' real names to accurately reflect their contributions and expertise. However, to protect individual privacy and prevent any potential personal information leakage, the publicly released version of our dataset replaces real names with unique identifiers (IDs). This anonymization step ensures that no personally identifying information is disclosed while maintaining the dataset's utility for research purposes.

We have taken these steps to safeguard the privacy and personal information of the individuals whose data contributed to our research. Additionally, we have reviewed the dataset to ensure it does not contain any offensive content.

Documentation Of Artifacts While our dataset does not involve artificial distributions, we have collected and included gender information in the metadata. This metadata, along with other relevant descriptive information about the dataset, will be made publicly available upon the paper's acceptance.

References

Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. 2023. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arXiv preprint arXiv:2303.04226*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Character.AI. 2022. Character.AI. https://character.ai/.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2023. When large language models meet personalization: Perspectives of challenges and opportunities. *arXiv preprint arXiv:2307.16376*.

Suzanne Fricke. 2018. Semantic scholar. *Journal of the Medical Library Association: JMLA*, 106(1):145.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. PersonaLLM: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

OpenAI. 2023. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. *arXiv preprint arXiv:2307.16789*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. LaMP: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Zhaoxuan Tan and Meng Jiang. 2023. User modeling in the era of large language models: Current research and future directions. *arXiv preprint arXiv:2312.11518*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. 2023a. RecMind: Large language model powered agent for recommendation. *arXiv preprint arXiv:2308.14296*.

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023b. Automated evaluation of personalized text generation using large language models. *arXiv preprint arXiv:2310.11593*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Penghui Wei, Xuanhua Yang, Shaoguo Liu, Liang Wang, and Bo Zheng. 2022b. CREATER: CTR-driven advertising text generation with controlled pre-training and contrastive fine-tuning. *arXiv preprint arXiv:2205.08943*.

Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. 2023a. Memory-augmented LLM personalization with short-and long-term memory coordination. *arXiv preprint arXiv:2309.11696*.

Kaiyan Zhang, Jianyu Wang, Ermo Hua, Biqing Qi, Ning Ding, and Bowen Zhou. 2024. Cogenesis: A framework collaborating large and small language models for secure context-aware instruction following. *arXiv preprint arXiv:2403.03129*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. DialoGPT: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, et al. 2023b. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. *arXiv preprint arXiv:2311.11797*.

Aakas Zhiyuli, Yanfang Chen, Xuan Zhang, and Xun Liang. 2023. BookGPT: A general framework for book recommendation empowered by large language model. *arXiv preprint arXiv:2305.15673*.

A Ablation Studies

To assess the contribution of each component, we perform an ablation study on the PSW dataset. Table 3 and 4 report the results of two variants: 1) Switching the order of users and 2) Removing user profiling.

A.1 Impact of Author Order

Table 3 shows how changing the author order affects the performance of multi-user personalized models. We experiment with three variants:

`Original`: The original author order as provided in the dataset. **`Swap-Random`**: Randomly shuffle the order of authors. **`Swap-First`**: Move the first author to the end of the author list.

Datasets	Variants	Metrics					
		ROUGE-1	ROUGE-L	Consistency	Fluency	Relevance	Novelty
PSW-1	`Original`	**0.337**	**0.280**	**3.59**	**2.58**	3.67	**2.63**
	`Swap-Random`	0.321	0.272	3.42	2.48	**3.69**	2.45
	`Swap-First`	0.314	0.260	3.35	2.42	3.48	2.37
PSW-2	`Original`	**0.201**	**0.186**	**4.60**	**2.39**	**3.91**	2.38
	`Swap-Random`	0.193	0.178	4.53	2.30	3.85	**2.42**
	`Swap-First`	0.186	0.171	4.46	2.27	3.77	2.29
PSW-3	`Original`	**0.145**	**0.131**	**4.92**	2.94	**4.71**	**2.45**
	`Swap-Random`	0.138	0.125	4.84	2.88	4.65	2.50
	`Swap-First`	0.130	0.117	4.78	**2.98**	4.57	2.55
PSW-4	`Original`	**0.505**	**0.444**	**4.64**	**2.59**	**3.79**	2.64
	`Swap-Random`	0.492	0.431	4.57	2.55	3.72	2.70
	`Swap-First`	0.483	0.421	4.50	2.50	3.64	**2.76**

Table 3: **Impact of author order on the performance of multi-user personalized models** We report additional metrics such as **Consistency (1-5)**, **Fluency (1-3)**, **Relevance (1-5)**, and **Novelty (1-3)**.

The **`Original`** order consistently achieves the best performance across all metrics on all PSW tasks. Randomly swapping authors (**`Swap-Random`**) leads to a slight decline, while moving the first author to the end (**`Swap-First`**) results in a more significant drop. This observation highlights the importance of preserving the original author order in multi-author collaborative writing scenarios. The first author, often the lead or corresponding author, significantly influences the document's content, structure, and style. As a result, their writing style and expertise tend to be most prominently reflected in the document. Disrupting this order introduces noise and hinders the model's ability to capture the individual authors' impact and the logical progression of ideas, particularly affecting the generation tasks (PSW-3 and PSW-4), where content and style are heavily influenced by the main author's expertise and preferences.

A.2 Impact of User Profiling

Table 4 reports ablation results on the user profile component:

`Original`: User profiles constructed using STEP-BACK PROFILING. **`Removed`**: No user profiles were used, only retrieving relevant snippets. **`Random`**: Replacing target user profiles with randomly sampled user profiles.

Datasets	Profile	Metrics					
		ROUGE-1	ROUGE-L	Consistency	Fluency	Relevance	Novelty
PSW-1	`Original`	**0.337**	**0.280**	**3.59**	**2.58**	**3.67**	**2.63**
	`Removed`	0.297	0.250	3.21	2.49	3.31	2.57
	`Random`	0.328	0.272	3.55	2.56	3.62	2.68
PSW-2	`Original`	**0.201**	**0.186**	**4.60**	2.39	**3.91**	2.38
	`Removed`	0.180	0.166	4.28	2.32	3.63	2.33
	`Random`	0.195	0.182	4.57	**2.42**	3.89	**2.45**
PSW-3	`Original`	**0.145**	**0.131**	4.92	2.94	**4.71**	2.45
	`Removed`	0.128	0.115	4.70	2.87	4.50	2.41
	`Random`	0.142	0.128	**4.95**	**2.96**	4.69	**2.51**
PSW-4	`Original`	**0.505**	**0.444**	**4.64**	**2.59**	**3.79**	2.64
	`Removed`	0.475	0.419	4.38	2.53	3.58	2.56
	`Random`	0.498	0.438	4.60	2.58	3.76	**2.69**

Table 4: **Impact of the user profile on the performance of multi-user personalized models.** We report additional metrics such as **Consistency (1-5)**, **Fluency (1-3)**, **Relevance (1-5)**, and **Novelty (1-3)**.

Removing user profiles (**`Removed`**) leads to the largest performance decline, confirming the benefit of STEP-BACK PROFILING in multi-user personalization. Using random profile texts (**`Random`**) recovers some of the gaps but still underperforms the **`Original`** profiles. This demonstrates that the distilled user traits successfully capture useful information for collaborative writing, such as individual writing styles, expertise, and preferences. The performance gap between **`Original`** and **`Random`** profiles highlights the effectiveness of the STEP-BACK PROFILING technique in extracting relevant user characteristics from their background information. These findings underscore the importance of incorporating author-specific traits to enable a more personalized and contextually appropriate generation in multi-user settings.

B The PSW Dataset

Overview. The PSW dataset is constructed using data from the Semantic Scholar database (Fricke, 2018). We first selected a subset of papers from Software Engineering published after 2000, considering only papers with at least two authors to ensure the feasibility of evaluating collaborative writing scenarios. The collected papers were randomly split into training, validation, and test subsets.[2] We performed the split at the paper level to ensure that all tasks within the PSW benchmark had consistent data splits. The summary of PSW dataset statistics can be found in Table 5.

Statistic	Train	Valid	Test
# of Papers	1,744	500	500
# of Authors	6,461	1,655	1,280
Avg. Authors / Paper	4.05	3.16	3.25
Avg. History Papers / Author	63.47	75.34	92.21
Avg. Research Interests / Author	2.84	2.77	2.79
Avg. Title Length	97.03	95.54	96.16
Avg. Abstract Length	970.92	981.36	1,037.09
Avg. Research Question Length	470.57	398.22	442.31
Avg. References / Paper	60.24	54.85	58.93

Table 5: **PSW Dataset Statistics with Train / Valid / Test Splits.**

B.1 Data Construction Details

UP-0: Research Interest Generation: Before all the PSW tasks, we create a benchmark for user profiling. This involves compiling a list of research interests that accurately reflect each author's expertise and research focus based on their publication history. To acquire the necessary information, we extract the research interests of each author from Google Scholar[3] by searching their name.

PSW-1: Research Topic Generation: This task aims to generate a list of research topics that capture the collaborating authors' joint expertise and research focus, given their user profiles. The generated research topics should be relevant to the authors' past publications and help identify potential research directions for their collaborative work. We use OpenAI's *GPT-4* model to automatically extract research topics from selected papers. The extracted topics are then linked to their respective papers and author profiles.

PSW-2: Research Question Generation: This task focuses on generating a set of research questions that align with the expertise and interests of the collaborating authors and are relevant to the target paper. The generated research questions should help guide the content and structure of the collaborative writing process. We automatically use OpenAI's *GPT-4* model to extract research questions from the selected papers for this task. The extracted research questions are then linked to their papers and author profiles.

PSW-3: Paper Abstract Generation: This task involves generating a paper abstract that summarizes the key points and contributions of the collaborative research paper, given the user profiles, research interests, target paper title, and research questions. We directly retrieve the abstracts from the selected papers using the Semantic Scholar API [4]. The retrieved abstracts are then linked to their respective papers and author profiles.

PSW-4: Paper Title Generation: This task aims to generate a suitable title for the collaborative research paper, considering the user profiles, research interests, research questions, and paper abstract. The data is collected by Semantic Scholar API as well.

C Metrics Visualization on PSW Dataset

Figures 3, 4, and 5 illustrate the results discussed in Section 4.2 and Appendices A.1 and A.2, respectively.

(a) PSW-1 (b) PSW-2

(c) PSW-3 (d) PSW-4

Figure 3: **Performance metrics across three different models**: **`Zero-shot`**, **`Single-Author`**, and **`Multi-Author`**. The **`Multi-Author`** model consistently achieves the highest scores across all datasets.

[2] We only used the test split in this paper since our method doesn't require model training.

[3] https://github.com/scholarly-python-package/scholarly

[4] https://api.semanticscholar.org/

(a) PSW-1 (b) PSW-2 (c) PSW-3 (d) PSW-4

Figure 4: **Impact of author order on the performance across three different models**: **`Original`**, **`Swap-Random`**, and **`Swap-First`**. The **`Original`** model consistently achieves the highest scores across all datasets.

(a) PSW-1 (b) PSW-2 (c) PSW-3 (d) PSW-4

Figure 5: **Impact of user profiling on the performance across three different models: `Original`**, **`Removed`**, and **`Random`**. The **`Original`** model consistently achieves the highest scores across all datasets.

D Details of G-Eval

Task Description
You will be given one result generated for a science paper and several reference papers. Your task is to rate the result using the following criteria. Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria
Consistency (1-5) – the factual alignment between the result and the corresponding science paper. A factually consistent result contains only statements entailed by the source document. **Fluency (1-3)** – the quality of the result in terms of grammar, spelling, punctuation, word choice, and sentence structure. **Relevance (1-5)** – the selection of important content from the source. The result should include only important information from the source document. **Novelty (1-3)** – the uniqueness and originality of the result in terms of concept, perspective, and creativity.
Evaluation Task
Now, you are working on evaluating this prediction: {Prediction Text} Here are some ground truth results for comparison: [result$_1$, result$_2$, . . .].
Instruction
Please evaluate the prediction using the above criteria.

Table 6: Prompt template for evaluating the G-Eval metric.

E Prompts for LaMP Tasks

E.1 Personalized Citation Identification (LaMP-1)

User Profile
Assuming you care a lot about these areas: **Keywords:** [keyword$_1$, keyword$_2$, keyword$_3$, . . .] **Topics:** [topics$_1$, topics$_2$, topics$_3$, . . .]
User History
I give you some titles of papers that you've written. Please imitate your reasons and recommend a paper citation for me. Each example consists of an abstract, the corresponding title, and a description of the writing style and keywords for that title.
Example 1
Title: {Title Text} **Abstract:** {Abstract Text} **Reason:** {Reason} **Citation:** [citation$_1$, citation$_2$, . . .]
Example 2
Title: {Title Text} **Abstract:** {Abstract Text} **Reason:** {Reason} **Citation:** [citation$_1$, citation$_2$, . . .] . . .
Example k
Title: {Title Text} **Abstract:** {Abstract Text} **Reason:** {Reason} **Citation:** [citation$_1$, citation$_2$, . . .]
Classification Task
Now you have written this title: **Title:** {Title Text}
Instruction
Please separately analyze the potential relevant connection of **Reference 1** and **Reference 2** to this title. You are citing from one of them. Please decide which one it would be: **Reference 1:** {option$_1$} **Reference 2:** {option$_2$} Just answer with [1] or [2] without explanation.

Table 7: Prompt template for the Personalized Citation Identification (LaMP-1) task.

E.2 Personalized News Categorization (LaMP-2)

User Profile
Assuming you care a lot about these areas: **Keywords:** [keyword$_1$, keyword$_2$, keyword$_3$, . . .] **Topics:** [topics$_1$, topics$_2$, topics$_3$, . . .]
User History
I give you some titles and articles that you've written with category. Please imitate your reasons for giving this category. Each example consists of an abstract, the corresponding title, and a category of it.
Example 1
Article: {Article Text} **Title:** {Title Text} **Reason:** {Reason} **Category:** [category$_1$, category$_2$, . . .]
Example 2
Article: {Article Text} **Title:** {Title Text} **Reason:** {Reason} **Category:** [category$_1$, category$_2$, . . .] . . .
Example k
Article: {Article Text} **Title:** {Title Text} **Reason:** {Reason} **Category:** [category$_1$, category$_2$, . . .]
Classification Task
Now you have written this article with the title: **Article:** {Article Text} **Title:** {Title Text}
Instruction
Which category does this article relate to among the following categories? **Category 1:** {option$_1$} **Category 2:** {option$_2$} . . . **Category K:** {option$_N$} Just answer with the category name without further explanation.

Table 8: Prompt template for the Personalized News Categorization (LaMP-2) task.

E.3 Personalized Product Rating (LaMP-3)

User Profile

Assuming you have written product reviews with the following characteristics:
Most Common Rating: {$\text{score}_{\text{most}}$}
Rating Patterns: [pattern_1, pattern_2, . . .]

User History

I provide you with some product reviews you've written, along with their corresponding ratings. Please imitate your reasoning for assigning these ratings. Each example consists of a product review and its rating.

Example 1

Product Review: {Review Text}
Rating: {Rating}

Example 2

Product Review: {Review Text}
Rating: {Rating}
. . .

Example k

Product Review: {Review Text}
Rating: {Rating}

Rating Task

Now you have written this new product review:
Product Review: {Review Text}
Based on the review, please analyze its sentiment and how much you like the product.

Instruction

Follow your previous rating habits and these instructions:

- If you feel satisfied with this product or have concerns but it's good overall, it should be rated 5.
- If you feel good about this product but notice some issues, it should be rated as 4.
- If you feel OK but have concerns, it should be rated as 3.
- If you feel unsatisfied with this product but it's acceptable for some reason, it should be rated as 2.
- If you feel completely disappointed or upset, it should be rated 1.

Your most common rating is {$\text{score}_{\text{most}}$}. You must follow this rating pattern faithfully and answer with the rating without further explanation.

Table 9: Prompt template for the Personalized Product Review Rating (LaMP-3) task.

E.4 Personalized News Headline Generation (LaMP-4)

User Profile

Assuming you have written headlines with the following characteristics:
Writing Style: [style_1, style_2, . . .]
Content Patterns: [patterns_1, patterns_2, . . .]

User History

I will provide you with some news articles along with the headlines you've written for them. Please imitate your writing style and content patterns when generating a new headline. Each example consists of a news article and its corresponding headline.

Example 1

Article: {Article Text}
Headline: {Headline}

Example 2

Article: {Article Text}
Headline: {Headline}
. . .

Example k

Article: {Article Text}
Headline: {Headline}

Generation Task

Now that you have been given this news article:
Article: {Article Text}

Instruction

Please write a headline following your previous writing styles and habits. If you have written headlines with similar content, you could reuse those headlines and mimic their content.

Table 10: Prompt template for the Personalized News Headline Generation (LaMP-4) task.

E.5 Personalized Scholarly Title Generation (LaMP-5)

User Profile
Assuming you have written scholarly titles with the following characteristics: **Writing Style:** [style$_1$, style$_2$, . . .] **Title Patterns:** [pattern$_1$, pattern$_2$, . . .]
User History
I will provide you with some research paper abstracts along with the titles you've written for them. Please imitate your writing style and title patterns when generating a new title. Each example consists of a paper abstract and its corresponding title.
Example 1
Abstract: {Abstract Text} **Title:** {Title}
Example 2
Abstract: {Abstract Text} **Title:** {Title} . . .
Example k
Abstract: {Abstract Text} **Title:** {Title}
Generation Task
Now that you have been given this paper abstract: **Abstract:** {Abstract Text}
Instruction
Please write a title following your previous style and habits, keeping it clear, accurate, and concise.

Table 11: Prompt template for the Personalized Scholarly Title Generation (LaMP-5) task.

E.6 Personalized Email Subject Generation (LaMP-6)

User Profile
Assuming you care a lot about these areas: **Keywords:** [keyword$_1$, keyword$_2$, keyword$_3$, . . .] **Topics:** [topics$_1$, topics$_2$, topics$_3$, . . .]
User History
Let's say there are some emails you've written. Please mimic the style of these examples. Each example consists of email content, the corresponding subject, and a description of the writing style for that title.
Example 1
Content: {Email Content} **Writing Style:** {Style} **Subject:** {Email Subject}
Example 2
Content: {Email Content} **Writing Style:** {Style} **Subject:** {Email Subject} . . .
Example k
Content: {Email Content} **Writing Style:** {Style} **Subject**: {Email Subject}
Generation Task
Now that you have been given this email content: **Content:** {Email Content}
Instruction
Write a title following your previous style and habits. Just answer with the subject without further explanation.

Table 12: Prompt template for the Personalized Email Subject Generation (LaMP-6) task.

E.7 Personalized Tweet Paraphrasing (LaMP-7)

User Profile
Assuming you have written tweets with the following characteristics:
Writing Style: [style_1, style_2, ...]
Tone: [tone_1, tone_2, ...]
Length: [length_1, length_2, ...]
User History
I will provide you with some original tweets along with the paraphrased versions you've written for them. When paraphrasing a new tweet, please imitate your writing style, tone, and typical length. Each example consists of an original tweet and its paraphrased version.
Example 1
Original Tweet: {Tweet Text}
Paraphrased Tweet: {Paraphrased Text}
Example 2
Original Tweet: {Tweet Text}
Paraphrased Tweet: {Paraphrased Text}
...
Example k
Original Tweet: {Tweet Text}
Paraphrased Tweet: {Paraphrased Text}
Generation Task
Now that you have been given this tweet:
Original Tweet: {Tweet Text}
Instruction
Please paraphrase it with the following instructions: • You must use tweet styles and tones. • You must keep it faithful to the given tweet with similar keywords and length.

Table 13: Prompt template for the Personalized Tweet Paraphrasing (LaMP-7) task.

F Prompts for PSW Tasks

F.1 Research Interests Generation (UP-0)

User History
I will provide you with some research papers you've authored. Please summarize your top research interests based on these papers. Each paper consists of a title and abstract.
Paper 1
Title: {Title Text}
Abstract: {Abstract Text}
Paper 2
Title: {Title Text}
Abstract: {Abstract Text}
...
Paper k
Title: {Title Text}
Abstract: {Abstract Text}
Instruction
Please summarize your top three research interests based on the provided papers in the following format: **Research Interests:** [interest_1, interest_2, interest_3, ...]

Table 14: Prompt template for the Research Interests Generation (UP-0) task.

F.2 Personalized Research Paper Title Generation (PSW-1)

User Profile
Assuming you are an expert researcher with the following research interests:
Research Interests: [interest$_1$, interest$_2$, interest$_3$, . . .]
User History
Here are some titles and abstracts from papers you have authored:
Paper 1
Title: {Title}
Abstract: {Abstract}
Paper 2
Title: {Title}
Abstract: {Abstract}
. . .
Paper k
Title: {Title}
Abstract: {Abstract}
Brainstorm Task
Here are some related papers for reference, each with a title:
Reference 1: {Title}
Reference 2: {Title}
. . .
Reference N: {Title}
Instruction
Considering your research interests, previous works, and reference papers, please brainstorm the most promising title for your new research paper.

Table 15: Prompt template for the Personalized Research Paper Title Generation (PSW-1) task.

F.3 Research Question Generation (PSW-2)

User Profile
Assuming you are an expert researcher with the following research interests:
Research Interests: [interest$_1$, interest$_2$, interest$_3$, . . .]
User History
Here are some titles and abstracts from papers you have authored:
Paper 1
Title: {Title}
Abstract: {Abstract}
Paper 2
Title: {Title}
Abstract: {Abstract}
...
Paper k
Title: {Title}
Abstract: {Abstract}
Brainstorm Task
Now you are working on a new paper with the following title:
Title: {Title}
Instruction
Considering the title and research background, please propose the top 3 research questions you aim to address in this new paper.

Table 16: Prompt template for the Research Question Generation (PSW-2) task.

F.4 Paper Abstract Generation (PSW-3)

User Profile
Assuming you are an expert researcher with the following research interests:
Research Interests: [interest$_1$, interest$_2$, interest$_3$, ...]
User History
Here are some titles and abstracts from papers you have authored:
Paper 1
Title: {Title}
Abstract: {Abstract}
Paper 2
Title: {Title}
Abstract: {Abstract}
...
Paper k
Title: {Title}
Abstract: {Abstract}
Generation Task
Now you are working on a new paper with the following title:
Title: {Title}
And you are focusing on solving the following research questions: [question$_1$, question$_2$, $\ldots$]
Instruction
Considering the title, research questions, and your writing style in previous abstracts, please write an abstract for this new paper.

Table 17: Prompt template for the Paper Abstract Generation (PSW-3) task.

F.5 Paper Title Generation (PSW-4)

User Profile
Assuming you are an expert researcher with the following research interests:
Research Interests: [interest$_1$, interest$_2$, interest$_3$, ...]
User History
Here are some titles and abstracts from papers you have authored:
Paper 1
Title: {Title}
Abstract: {Abstract}
Paper 2
Title: {Title}
Abstract: {Abstract}
...
Paper k
Title: {Title}
Abstract: {Abstract}
Generation Task
Now, you are working on a new paper with the following abstract:
Abstract: {Abstract}
And you are focusing on solving the following research questions: [question$_1$, question$_2$, $\ldots$]
Instruction
Considering the abstract and your title writing style in previous papers, please generate a title for this new paper. The title should be clear and concise and reflect the main topic of the abstract as well as your research questions.

Table 18: Prompt template for the Paper Title Generation (PSW-4) task.