Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling and Personalized Responses at Scale

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large Language Models (LLMs) have emerged as *personalized* assistants for users across a wide range of tasks. Over time, the interaction history between a user and an LLM can provide extensive information about an individual's traits and preferences. However, open questions remain on how well LLMs today can effectively leverage such history to (1) internalize the user's inherent traits and preferences, (2) track how the user profiling and preferences evolve over time, and (3) generate personalized responses accordingly in new scenarios. In this work, we introduce the PERSONAMEM benchmark. PERSONAMEM features curated user profiles with over 180 simulated user-LLM interaction histories, each containing up to 60 sessions of multi-turn conversations across 15 real-world tasks. We observe that current LLMs still struggle to deliver responses that align with users' current situations and preferences, with frontier models such as GPT-4.1, GPT-4.5, o4-mini, or Gemini-2.0 achieving only around 50% overall accuracy.

1 Introduction

An increasing number of users now rely on Large Language Models (LLMs) as *personalized* assistants in everyday tasks. While no single AI system can satisfy all users, personalization, i.e., adapting responses to individual traits, preferences from user-chatbot interaction histories, helps move beyond generic outputs toward more relevant and engaging ones.

Personalizing LLMs is challenging because models cannot easily access all information information about the user, especially user preferences evolving over time (Radlinski & Craswell, 2017; Dean & Morgenstern, 2022). For instance, For example, as illustrated in Figure 1, a user who once said *"I like pizza"* may later request gluten-free options after discovering an allergy. Current chatbots often fail to track such changes, making them feel less helpful and empathetic (Aggarwal et al., 2023; Ait Baha et al., 2023).

We address this gap by introducing the PERSONAMEM benchmark, which contains over 180 simulated user-LLM histories, up to 60 sessions and 1M tokens in context windows, built from evolving personas. These histories capture shifting traits and preferences through multi-turn interactions across 15 conversation scenarios such as food, travel, and therapy consultations.

Using PERSONAMEM, we evaluate 15 state-of-the-art models on 7 types of *in-situ* user queries, measuring whether LLMs can (1) internalize the user's inherent traits and preferences, (2) track how the user profiling and preferences evolve over time, and (3) generate personalized responses accordingly in new scenarios. We find that frontier models like GPT-4.1, GPT-4.5, o4-mini, and

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

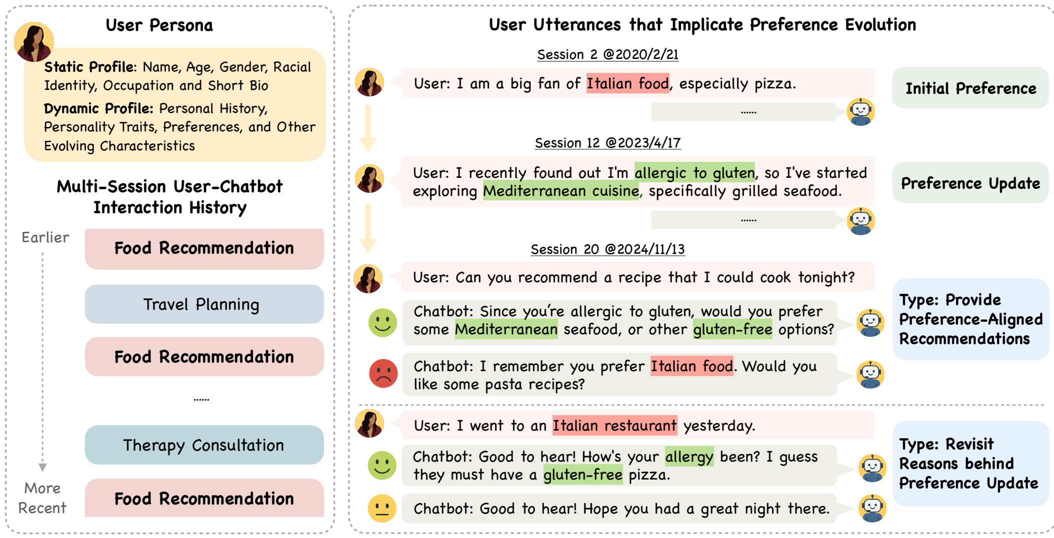

Figure 1: Overview of PERSONAMEM benchmark. Each sample is a user persona with static (e.g., demographic info.) and dynamic attributes (e.g., evolving preferences). Users engage with a chatbot in multi-session interactions across a variety of topics such as food recommendation, travel planning, and therapy consultation. As the user's preferences evolve over time, the benchmark offers annotated questions assessing whether models can track and incorporate the changes into their responses.

DeepSeek-R1 achieve only around 50% accuracy. While they handle fact recall and preference tracking reasonably well, they struggle to adapt responses to new scenarios.

2 PERSONAMEM Benchmark: Overview

Each instance in the benchmark dataset features a *user profile or persona* expanded from PersonaHub (Ge et al., 2024), which includes basic demographic information (such as name, age, gender, and occupation), as well as *dynamic* user characteristics such as user traits, preferences, and events happening in the user's life. The dynamic user characteristics change over time as different events happen to the user that will lead to changes in users' traits under each task scenario.

At different points in time of a user's profile evolution, the user engages in multi-turn conversations with LLM and seeks help or suggestions from LLM on one of the task scenarios. In each task scenario, the user would ask for the LLM's suggestions given the user's need and current situation. The conversation sessions across different tasks are interleaved by the temporal order in which the sessions happen.

To understand how well LLM chatbots can track the evolution in a user's profile from the conversation histories, we evaluate LLMs by whether they can provide the most suitable response to *in-situ* user queries, where the user issues the query to LLM in a new conversation session from the first-person perspective. Depending on the time of the *in-situ* query, the expected response from the model will differ. We cast the problem as a multiple-choice selection, where LLM needs to identify the correct response out of four choices, where the incorrect choices are based on either outdated or irrelevant information with respect to the current state of the user's profile.

Types of skills evaluated. To evaluate LLMs' ability to (1) memorize the user profile, (2) track how the user profile evolve over time, and (3) generate personalized responses accordingly in new scenarios, we design the following 7 types of *in-situ* user queries in the PERSONAMEM benchmark. We include examples for each type of user queries in Table 1.

1. **Recall user-shared facts.** We evaluate whether a personalized chatbot can recall static events, activities, or interests the user has shared in previous interactions, and incorporate the information in its responses.

2. **Suggest new ideas.** We evaluate whether a chatbot can suggest new items or activities that have not been mentioned in the interaction history, when users explicitly request so, e.g. "*suggest new restaurants I haven't ordered from before*".
3. **Acknowledge latest user preferences.** We evaluate whether a chatbot can recognize the latest preference expressed by the user in the interaction history.
4. **Track full preference evolution.** We evaluate whether a chatbot can keep track of how users' preferences shift by time.
5. **Revisit reasons behind preference updates.** We evaluate whether a chatbot can recall the reason(s) or event(s) leading to the preference change from a user.
6. **Provide preference-aligned recommendations.** We test whether a chatbot can proactively offer new recommendations that aligns with the user's current preferences.
7. **Generalize to new scenarios.** We evaluate whether a chatbot can transfer what it learns about the user from other task scenarios to a new task.

Benchmark data statistics. PERSONAMEM features 20 personas, with over 180 interaction histories. Each interaction history contains 10, 20, or 60 sessions, where we dynamically adjust the total length of the history to approximately $32k$, $128k$, and $1M$ tokens, respectively. Each session consists of 15–30 conversation turns between a user and an LLM chatbot. The user-LLM conversations span across 15 diverse topics, ranging from therapy and legal advice to recommendations on books, music, movies, and food; personal matters such as family, dating, health, and finance; and practical tasks like travel planning, online shopping, studying tips, and home decoration. In total, the benchmark features around $6k$ *in-situ* user query and LLM response pairs across the 7 query types. Detailed dataset breakdown is discussed in Appendix D. The size of our benchmark is not limited by the scalability of the synthetic data pipeline but to make the evaluation cost reasonable.

3 Experimental Results

3.1 Evaluation Settings

Given an *in-situ* user query and the user's interaction history, models must select the correct response from four options, where only one reflects the user's current state and the others contain outdated or irrelevant information. Models also receive basic demographics such as name, age, gender identity, racial identity, and occupation but not other dynamic traits or personal history. We evaluate under two settings: *discriminative*, where the model chooses among four labeled options (a–d) and explains its choice, and *generative*, where the model scores each option by log-probability with length normalization and selects the highest. The generative setting requires token-level logits, which are often unavailable for proprietary models. No LLM judges are used.

3.2 Evaluating Language Models in Long-Context Settings

We first evaluate language models in the long-context setting, where the full user-LLM interaction history is provided as input to the models. Due to the length of the history, all models here were evaluated zero-shot, without demonstration examples of other histories and user queries. Quantitative Results can be seen in Figure 2 and Figure 3. **GPT-4.5, GPT-4.1, and Gemini-1.5 achieve the highest overall performance.** Among leading foundation models, GPT-4.5 and Gemini-1.5 outperform others in overall accuracy. However, their performance still hovers around 52% in a multiple-choice setting, highlighting substantial room for improvement. **Notably, reasoning models such as o1, o3-mini, o4-mini, and DeepSeek-R1-607B do not demonstrate competitive advantages over non-reasoning models in the personalization tasks we evaluate.**

LLMs demonstrate reasonably good performance in recalling simple user facts. For tasks involving the retrieval of static user information, such as previously mentioned items, activities, or reasons behind preference changes where the reasons themselves won't change, most LLMs have a reasonable chance of succeeding.

Incorporating the latest user preference into responses is more challenging than recalling the change in user profile. We observe that models struggle to incorporate the latest preference or

Query Type \ Model	Gemini 1.5-Flash	GPT-4.5	GPT-4.1	o1	Gemini 2.0-Flash	o4-mini	Gemini 2.0-Flash-Lite	GPT-4o	DeepSeek R1-671B	Llama 4-Maverick	o3-mini	GPT 4o-mini	Llama 3.1-405B	Claude 3.5-Haiku	Claude 3.7-Sonnet	Average	Random Guess
Revisit Reasons Behind Preference Updates	0.77	0.76	0.84	0.75	0.79	0.75	0.77	0.77	0.83	0.76	0.72	0.70	0.41	0.64	0.57	0.72	0.25
Tracking Full Preference Evolution	0.65	0.68	0.67	0.67	0.70	0.73	0.68	0.68	0.68	0.66	0.54	0.60	0.38	0.55	0.45	0.62	0.25
Recall User Shared Facts	0.54	0.61	0.65	0.50	0.50	0.42	0.49	0.41	0.43	0.37	0.47	0.55	0.38	0.29	0.25	0.46	0.25
Acknowledge Latest User Preference	0.59	0.55	0.50	0.54	0.52	0.55	0.51	0.46	0.42	0.43	0.39	0.34	0.31	0.27	0.09	0.43	0.25
Provide Preference Aligned Recommendations	0.55	0.44	0.57	0.42	0.51	0.41	0.52	0.37	0.49	0.42	0.41	0.41	0.37	0.32	0.20	0.43	0.25
Generalize Reasons to New Scenarios	0.54	0.46	0.53	0.39	0.46	0.38	0.33	0.32	0.38	0.32	0.30	0.33	0.21	0.20	0.29	0.36	0.25
Suggest New Ideas	0.15	0.27	0.19	0.25	0.15	0.17	0.16	0.24	0.16	0.20	0.11	0.10	0.20	0.06	0.28	0.18	0.25
Overall Accuracy	0.52	0.52	0.52	0.50	0.49	0.48	0.48	0.45	0.45	0.43	0.39	0.39	0.31	0.30	0.26	0.43	0.25

Figure 2: Evaluation results across different models on 7 *in-situ* query types. We observe models perform reasonably well at recalling user facts and preferences. However, models struggle at providing novel suggestions, or applying users' preferences in new scenarios.

128k tokens

Model \ Num of Sessions	Overall	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Gemini-1.5-Flash	0.52	0.74	0.56	0.53	0.54	0.50	0.47	0.49	0.51	0.53	0.50	0.49	0.42	0.54	0.48	0.44	0.53	0.65	0.57	0.48	0.48
GPT-4.5	0.52	0.74	0.53	0.57	0.56	0.54	0.52	0.50	0.44	0.52	0.46	0.46	0.41	0.52	0.53	0.36	0.48	0.68	0.65	0.48	0.42
GPT-4.1	0.52	0.87	0.56	0.60	0.53	0.52	0.56	0.49	0.53	0.43	0.44	0.47	0.46	0.43	0.44	0.44	0.54	0.64	0.57	0.49	0.44
o1	0.50	0.68	0.56	0.54	0.49	0.54	0.45	0.48	0.46	0.48	0.45	0.46	0.41	0.53	0.39	0.36	0.44	0.66	0.58	0.47	0.44
Gemini-2.0-Flash	0.49	0.73	0.52	0.55	0.48	0.45	0.48	0.48	0.51	0.50	0.44	0.42	0.41	0.52	0.46	0.42	0.46	0.61	0.51	0.52	0.42
o4-mini	0.48	0.82	0.49	0.46	0.51	0.46	0.43	0.43	0.42	0.39	0.39	0.39	0.45	0.45	0.48	0.44	0.46	0.65	0.52	0.50	0.45
Gemini-2.0-Flash-Lite	0.48	0.76	0.45	0.52	0.50	0.42	0.48	0.44	0.40	0.46	0.35	0.38	0.40	0.44	0.47	0.44	0.56	0.63	0.53	0.50	0.40
GPT-4o	0.45	0.83	0.51	0.55	0.44	0.43	0.47	0.38	0.42	0.43	0.40	0.36	0.38	0.42	0.32	0.29	0.38	0.66	0.54	0.48	0.36
DeepSeek-R1-671B	0.45	0.84	0.56	0.51	0.49	0.50	0.47	0.50	0.45	0.41	0.28	0.35	0.28	0.43	0.30	0.38	0.46	0.61	0.50	0.44	0.37
Llama-4-Maverick	0.43	0.76	0.31	0.45	0.48	0.38	0.33	0.37	0.45	0.36	0.39	0.30	0.41	0.37	0.39	0.39	0.54	0.62	0.50	0.50	0.36
o3-mini	0.39	0.80	0.48	0.44	0.45	0.36	0.39	0.39	0.36	0.37	0.27	0.31	0.38	0.35	0.32	0.26	0.41	0.56	0.39	0.35	0.33
GPT-4o-mini	0.39	0.73	0.45	0.46	0.36	0.34	0.37	0.36	0.35	0.25	0.30	0.29	0.32	0.34	0.33	0.36	0.42	0.60	0.44	0.37	0.32
Llama-3.1-405B	0.31	0.40	0.30	0.32	0.27	0.25	0.24	0.32	0.25	0.34	0.30	0.30	0.37	0.29	0.28	0.34	0.33	0.42	0.36	0.27	0.31
Claude-3.5-Haiku	0.30	0.60	0.27	0.38	0.27	0.28	0.22	0.24	0.26	0.25	0.18	0.22	0.26	0.36	0.25	0.24	0.35	0.52	0.34	0.33	0.22
Claude-3.7-Sonnet	0.26	0.76	0.27	0.31	0.26	0.20	0.28	0.21	0.20	0.10	0.15	0.17	0.12	0.22	0.20	0.19	0.29	0.47	0.28	0.27	0.19
Average	0.43	0.74	0.46	0.48	0.44	0.41	0.41	0.41	0.40	0.39	0.35	0.36	0.36	0.41	0.38	0.36	0.44	0.60	0.49	0.43	0.37
Random Guess	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25

1M tokens

Model \ Num of Sessions	Overall	1–3	4–6	7–9	10–12	13–15	16–18	19–21	22–24	25–27	28–30	31–33	34–36	37–39	40–42	43–45	46–48	49–51	52–54	55–57	58–60
Gemini-1.5-Flash	0.45	0.64	0.60	0.51	0.55	0.46	0.39	0.42	0.46	0.37	0.33	0.38	0.42	0.37	0.39	0.51	0.50	0.42	0.51	0.50	0.46
Gemini-2.0-Flash-Lite	0.38	0.41	0.51	0.54	0.47	0.36	0.32	0.32	0.34	0.35	0.27	0.44	0.43	0.32	0.27	0.40	0.40	0.35	0.42	0.47	0.38
Gemini-2.0-Flash	0.37	0.44	0.51	0.32	0.51	0.37	0.30	0.34	0.32	0.31	0.34	0.36	0.45	0.40	0.29	0.36	0.36	0.38	0.28	0.50	0.37
Llama-4-Maverick	0.28	0.38	0.32	0.42	0.33	0.23	0.25	0.27	0.34	0.21	0.23	0.14	0.32	0.19	0.19	0.28	0.29	0.26	0.31	0.34	0.30
Average	0.37	0.47	0.49	0.45	0.46	0.36	0.31	0.34	0.37	0.31	0.29	0.33	0.40	0.32	0.29	0.39	0.39	0.35	0.38	0.45	0.38
Random Guess	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25

Figure 3: Model performances by number of sessions elapsed since most recent preferences were mentioned in long context. Top: up to 20 sessions/128k tokens; Bottom: up to 60 sessions/1M tokens. Long-context retrieval is important for personalization in practice.

state of the user in responses. Surprisingly, models generally get higher performance when asked to recall how the user preferences evolve over time. We observe that asking the model to iterate through all preference updates may encourage it to think through the preference evolutions, often making the task easier.

Models fall short on generating new ideas or providing suggestions in new scenarios. As shown in Figure 2, tasks such as *"Suggest New Ideas*", *"Provide Preference-Aligned Recommendations"*, and *"Generalize Reasons to New Scenarios"* yield the lowest performance across all models, highlighting the challenge of generating personalized responses in novel contexts—particularly when identifying new facts.

4 Conclusion

In this paper, we introduce the PERSONAMEM benchmark, featuring scalable and persona-oriented multi-session user-LLM interaction histories, as well as fine-grained *in-situ* user query types designed to evaluate LLM capabilities in memorizing, tracking, and incorporating users' dynamic profiles into personalized responses. Through comprehensive assessments of 15 state-of-the-art LLM models, we highlight current challenges in enabling LLMs to deliver truly personalized conversations with users, especially in novel scenarios and long contexts. We hope that our benchmark opens new avenues for future exploration and advancement in personalized LLM chatbot development.

References

Saleh Afzoon, Usman Naseem, Amin Beheshti, and Zahra Jamali. Persobench: Benchmarking personalized response generation in large language models. *arXiv preprint arXiv:2410.03198*, 2024.

Abhishek Aggarwal, Cheuk Chi Tam, Dezhi Wu, Xiaoming Li, and Shan Qiao. Artificial intelligence–based chatbots for promoting health behavioral changes: systematic review. *Journal of medical Internet research*, 25:e40789, 2023.

Tarek Ait Baha, Mohamed El Hajji, Youssef Es-Saady, and Hammou Fadili. The power of personalization: A systematic review of personality-adaptive chatbots. *SN Computer Science*, 4(5):661, 2023.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3119–3137, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.172. URL `https://aclanthology.org/2024.acl-long.172/`.

David Castillo-Bolado, Joseph Davidson, Finlay Gray, and Marek Rosa. Beyond prompts: Dynamic conversational benchmarking of large language models. *arXiv preprint arXiv:2409.20222*, 2024.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. Persona: A reproducible testbed for pluralistic alignment. *arXiv preprint arXiv:2407.17387*, 2024.

Sarah Dean and Jamie Morgenstern. Preference dynamics under personalized recommendations. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 795–816, 2022.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. Can llm be a personalized judge? *arXiv preprint arXiv:2406.11657*, 2024.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024.

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

Jiho Kim, Woosog Chay, Hyeonji Hwang, Daeun Kyung, Hyunseung Chung, Eunbyeol Cho, Yohan Jo, and Edward Choi. Dialsim: A real-time simulator for evaluating long-term dialogue understanding of conversational agents. *arXiv preprint arXiv:2406.13144*, 2024.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*, 2024.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. In search of needles in a 10m haystack: Recurrent memory finds what llms miss. *arXiv preprint arXiv:2402.10790*, 2024.

Xintong Li, Jalend Bantupalli, Ria Dharmani, Yuwei Zhang, and Jingbo Shang. Toward multi-session personalized conversation: A large-scale dataset and hierarchical tree framework for implicit reasoning. *arXiv preprint arXiv:2503.07018*, 2025.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*, 2024.

Elliot Nelson, Georgios Kollias, Payel Das, Subhajit Chaudhury, and Soham Dan. Needle in the haystack for memory based large language models. *arXiv preprint arXiv:2407.01437*, 2024.

Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pp. 117–126, 2017.

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

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pp. 29971–30004. PMLR, 2023.

Juntao Tan, Liangwei Yang, Zuxin Liu, Zhiwei Liu, Rithesh Murthy, Tulika Manoj Awalgaonkar, Jianguo Zhang, Weiran Yao, Ming Zhu, Shirley Kokane, et al. Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data. *arXiv preprint arXiv:2502.20616*, 2025.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*, 2024.

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

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

J Xu. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*, 2021.

Xiaoyue Xu, Qinyuan Ye, and Xiang Ren. Stress-testing long-context language models with lifelong icl and task haystack. *arXiv preprint arXiv:2407.16695*, 2024.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2639–2650, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.207. URL https://aclanthology.org/2022.findings-acl.207/.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. ∞ bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15262–15277, 2024.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. In *The thirteenth international conference on learning representations*, 2025.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. Personalllm: Tailoring llms to individual preferences. *arXiv preprint arXiv:2409.20296*, 2024.

A Limitations and future work

A.1 Broader context in user privacy concerns

Privacy is a critical aspect of LLM personalization in the real world. In our setting, we personalize responses based on only preferences and activities shared by the user in previous user-chatbot interactions, and the model uses this information for its own responses without external sharing. To avoid potential privacy risks associated with real user data, we intentionally propose a synthetic data curation pipeline in this work. This synthetic approach allows researchers in the community to safely explore personalization methods. One possible direction for future work could be designing question-answer pairs that specifically involve sensitive user information.

A.2 More advanced retrieval methods

Our current exploration of retrieval-augmented methods, such as RAG and Mem0, is intended as a proof of concept, as the primary focus of this work is on the design and release of the personalization benchmark. We are excited to encourage more exploration on state-of-the-art long-context, memory, and retrieval-augmented generation methods in future work, especially those that preserve and understand the evolution of user personas and reasons behind preference updates, as well as enhancing user personalization in new or unseen scenarios.

A.3 Potential artifacts in the synthetic data generation process

To reduce artifacts that might make the benchmark artificially easier, we've taken several steps. For example, we removed question-answer pairs where the correct answer was unintentionally obvious, such as being noticeably longer or sharing identical key words with the questions. We also filtered out queries that an LLM can answer correctly more than once in three attempts, without seeing any actual conversation context. Besides, we have included checks in our human evaluations to confirm that the correct answers can indeed be derived from the provided context.

A.4 Potential gaps between evaluations on open-ended generations and multiple choices

In purely open-ended generative settings, personalization can lead to many possible correct answers, depending on how the user persona is used and which related user preference is used. Meanwhile, open-ended evaluations are computationally expensive due to the need for LLM-as-a-Judge for each question-answer pair. As a result, we evaluate generative tasks by computing the joint log-likelihood of each candidate option, without explicitly presenting all four options in the prompt. This approach yields similar patterns with those observed in standard discriminative evaluations in our experiment, while offering a more reliable basis for benchmarking performance compared to fully open-ended ones.

B 7 Types of Questions in PersonaMem

Query type	Examples of *in-situ* user queries and chatbot responses
[1] Recall user-shared facts	*"User: I shared my playlist with my friends and they loved it. ...... (later) User: What are some creative ways to share music? — Chatbot: Curating personalized playlists can be fun."*
[2] Suggest new ideas	*"User: ...... (later) User: Suggest new restaurant I haven't ordered before."*
[3] Acknowledge latest user preferences	*"User: I am a big fan of Italian food. ...... (later) User: I prefer Mediterranean cuisine now. ... (later) User: I went to an Italian restaurant yesterday — Chatbot: Nice to hear! Though I thought you prefer Mediterranean food now."*
[4] Track full preference evolution	*"User: ...... (later) User: I am thinking about attending a cooking class. — Chatbot: I understand your perspective on cooking classes has evolved. Initially, you expressed a dislike, but later you mentioned your interest in them."*
[5] Revisit reasons behind preference updates	*"User: I am a big fan of Italian food. ...... (later) User: I recently found I'm allergic to gluten, so I've shifted to other dishes. ... (later) User: I went to an Italian restaurant yesterday — Chatbot: How's your allergy been? I guess they must have a gluten-free pizza."*
[6] Provide preference aligned recommendations	*User: I prefer Mediterranean seafood now. ...... (later) User: How can I make impactful lifestyle changes? — Chatbot: How about diving into the Mediterranean diet plan?"*
[7] Generalize to new scenarios	*"User: I prefer my yoga class because of its fixed weekly schedule. ...... (later) User: I recently haven't had time to cook. — Chatbot: Since you seem to love routine would you be interested in weekly scheduled bento-box deliveries?"*

Table 1: Examples of the 7 types of *in-situ* user queries and expected chatbot response in the PersonaMem benchmark. We highlight words that signal the user's facts or preferences.

C Related Work

C.1 Evaluating Long-Context Memory Capabilities of LLMs

Needle-in-the-haystack tests, which task models to locate specific facts within a given long context, are a common method for this evaluation. Prior benchmarks perform tasks from direct information retrieval (Kuratov et al., 2024; Nelson et al., 2024) to question answering and summarization (Xu et al., 2024; Bai et al., 2024; Zhang et al., 2024). A more real-world setting for such evaluation is through dialogue conversations. Earlier benchmarks curated human-human (Xu, 2021) or human-AI interactions Xu et al. (2022), with sessions up to 10K tokens. More recent works have used LLMs to generate much longer sessions of 100k+ tokens long (Maharana et al., 2024; Kim et al., 2024; Castillo-Bolado et al., 2024). More recently, Wu et al. (2024) present LONGMEMEVAL, a dialogue benchmark which also considers contexts up to 1M, and uses persona-driven sessions. The major differences are that sessions from PERSONAMEM consider a broader range of topics than just task-oriented ones; and that the evaluation of PERSONAMEM focuses on fine-grained personalization concerns, rather than more general memory abilities.

C.2 Towards Personalization in Large Language Models

As users have a diversity of preferences, both at a demographic-level (Santurkar et al., 2023) and at an individual-level (Zollo et al., 2024). *Personas* are short biographies of individuals, that capture both levels, and can be generated en masse by LLMs (Ge et al., 2024). Researchers have used personas to evaluate how LLMs can adapt to users and environments (Castricato et al., 2024; Tseng et al., 2024). Reliable evaluation of personalization is also key. Many of the aforementioned benchmarks through formulation as NLP tasks, and another line of work uses LLMs to automatically judge texts along different axes of personalization (Dong et al., 2024; Wang et al., 2023). The approach taken by PERSONAMEM follows the former, as we report performance on question-answering. Importantly though, the personalization evaluation is by design of the questions and answers, each of which is grounded in specific temporal events, and is generated to adhere to a specific question type.

Turning to the dialogue setting, earlier works like LAMP and PERSONALLM consider personalization within a single turn or session (Salemi et al., 2023; Jiang et al., 2023; Kirk et al., 2024). More recently, IMPLEXCONV (Li et al., 2025) focuses on modeling implicit reasoning within personalized conversations. PERSONABENCH (Tan et al., 2025) simulates social interactions among diverse users through numerous but shorter sessions and access to synthetic private user data. PERSOBENCH (Afzoon et al., 2024) leverages existing persona-aware datasets to evaluate language quality, persona coverage, and consistency. LONGLAMP (Kumar et al., 2024) focuses on generating long-form texts other than more interactive responses within long context. Zhao et al. (2025) introduce PREFEVAL, which evaluates LLMs' preference-following abilities for 20 topics in persona-oriented dialogues of 100k+ tokens. PERSONAMEM, besides the flexible setting of generating numerous $1M$-token contexts efficiently, places greater emphasis on personas as simulated humans in user-model interactions, featuring multiple fine-grained personalization tasks where profiles and preferences evolve through temporally grounded events.

D Detailed Breakdown of the PersonaMem Statistics

Below is a more detailed breakdown of the dataset.

D.1 Different Query Types

- Recall_user_shared_facts: 5.8%
- Acknowledge_latest_user_preferences: 30.09%
- Track_full_preference_evolution: 10.97%
- Revisit_reasons_behind_preference_updates: 9.28%
- Provide_preference_aligned_recommendations: 11.58%
- Suggest_new_ideas: 22.92%
- Generalize_to_new_scenarios: 9.35%

D.2 Different Conversation Topics

- Book Recommendation: 6.3%
- Dating Consultation: 7.2%
- Family Relations: 5.3%
- Financial Consultation: 7.3%
- Food Recommendation: 8.4%
- Home Decoration: 5.6%
- Legal Consultation: 10.4%
- Medical Consultation: 7.2%
- Movie Recommendation: 5.8%
- Music Recommendation: 1.6%
- Online Shopping: 7.2%
- Sports Recommendation: 7.2%
- Study Consultation: 5.8%
- Therapy: 9.1%
- Travel Planning: 5.7%

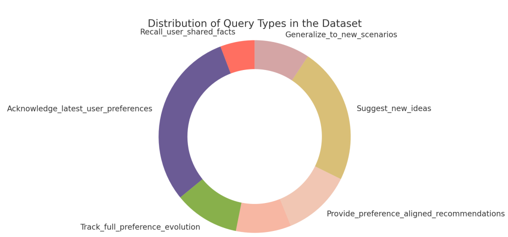

Figure 4: Distribution of Query Types in the Dataset

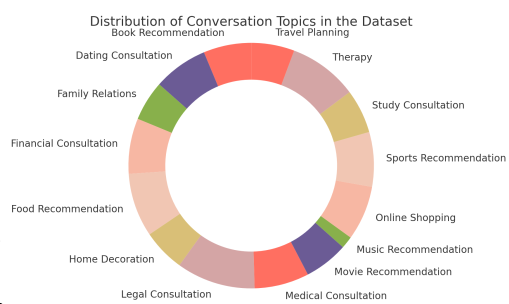

Figure 5: Distribution of Conversation Topics in the Dataset

D.3 Distance from the User Query to the Reference Information in the Context (PersonaMem_128k)

- 0-2 sessions: 5.6%
- 3-6 sessions: 20.1%
- 7-10 sessions: 17.6%
- 11-14 sessions: 17.9%
- 15-18 sessions: 23.6%
- 19-20 sessions: 15.2%

D.4 Distance from the User Query to the Reference Information in the Context (PersonaMem_128k) in Tokens

- 0-9.18k tokens: 5.7%
- 9.18k-22.3k tokens: 14.8%
- 22.3k-35.4k tokens: 11.3%
- 35.4k-48.5k tokens: 7.4%
- 48.5k-61.6k tokens: 8.2%
- 61.6k-74.7k tokens: 8.1%
- 74.7k-87.8k tokens: 8.6%
- 87.8k-101k tokens: 11.6%
- 101k-114k tokens: 17.1%
- 114k-128k tokens: 7.3%

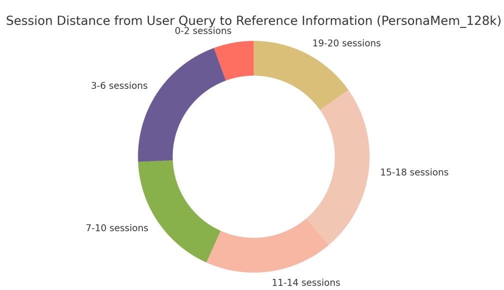

Figure 6: Session Distance from User Query to Reference Information

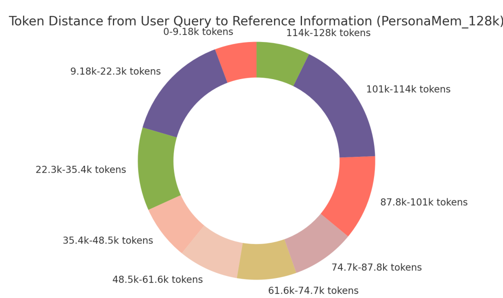

Figure 7: Token Distance from User Query to Reference Information

D.5 For PersonaMem_1M

D.5.1 Distance from the User Query to the Reference Information in the Context (PersonaMem_1M) in Terms of Sessions

- 0-7 sessions: 5.6%
- 8-13 sessions: 6.1%
- 14-19 sessions: 10.1%
- 20-25 sessions: 11.4%
- 26-31 sessions: 8.3%
- 32-37 sessions: 8.9%
- 38-43 sessions: 9.6%
- 44-49 sessions: 9.9%
- 50-55 sessions: 11.7%
- 56-60 sessions: 18.3%

D.5.2 Distance from the User Query to the Reference Information in the Context (PersonaMem_1M) in Tokens

- 0-101k tokens: 6.1%
- 101k-195k tokens: 5.5%
- 195k-288k tokens: 10.3%
- 288k-381k tokens: 10.2%
- 381k-474k tokens: 12.8%
- 474k-568k tokens: 8.3%
- 568k-661k tokens: 9.1%
- 661k-754k tokens: 9.6%
- 754k-847k tokens: 11.4%
- 847k-1M tokens: 16.7%

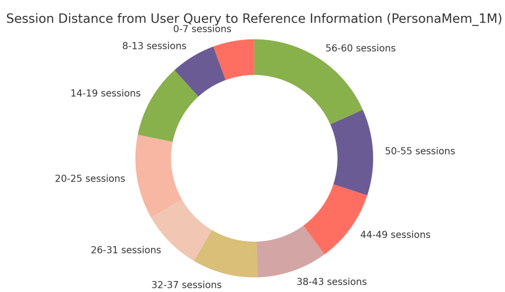

Figure 8: Session Distance from User Query to Reference Information

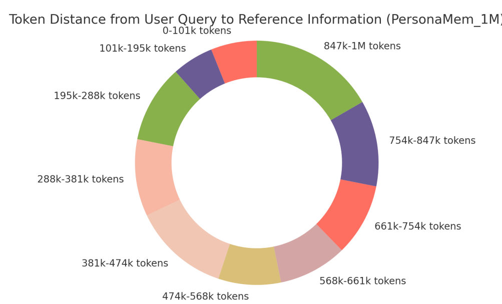

Figure 9: Token Distance from User Query to Reference Information

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we highlight the main scope of personalization and our contributions as establishing a state-of-the-art personalization benchmark, including showing main experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mentioned limitations in the appendix section A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not have theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have described evaluation settings and released all data and code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes we have open-sourced all the data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: We released a benchmark for evaluation only without training splits.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Evaluating state-of-the-art LLMs on large-scale benchmark is costly, preventing us from reporting error bars. Nonetheless, the benchmark dataset is large enough to ensure generalizability.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We rely solely on public LLM API calls, without requiring additional compute resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We mentioned broader impacts in the appendix section A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We have imposed several safety filters during the data generation process.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited our seeding data, i.e., PersonaHub (Ge et al., 2024) in our manuscript.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We provide synthetic data only. The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.