In-Context Personalized Alignment with Feedback History under Counterfactual Evaluation

Xisen Jin [1] Zheng Li [2] Zhenwei Dai [2] Hui Liu [2] Xianfeng Tang [2] Chen Luo [2] Rahul Goutam [2] Xiang Ren [1] Qi He [2]

Abstract

Accommodating diverse preferences of users is an arising challenge in large language model (LLM) alignment. A prevalent solution is to prompt LLMs with past user feedback in earlier conversations, so that LLMs can infer and adapt generations to the user preferences. In this paper, we revisit such in-context LLM personalization paradigm under a synthetic counterfactual evaluation setup, where each candidate response can be the preferable response depending on the preferences. We examine whether model responses can be steered to diverse preferences with distinct feedback history provided in-context. Our experiments suggest that off-the-shelf LLMs struggle in understanding user preferences from in-context feedback for personalized reward modeling and response generation. We show that fine-tuning is almost necessary so that in-context feedback is leveraged, where small 7-8B LLMs improve over off-the-shelf LLMs. Lastly, we improve fine-tuned response generation models via rejection sampling of training data guided by the personalized reward model.

1. Introduction

Language model alignment tries to align generations of large language models (LLMs) with objectives such helpfulness and harmlessness (Bai et al., 2022). Recent research calls for improved utility of LLMs to individual users by personalizing the responses (*e.g.,* to the expertise level of the users) (Kirk et al., 2024; Sorensen et al., 2023; Casper et al., 2023).

A prevalent approach to LLM personalization is in-context learning from past user feedback (Zollo et al., 2024; Salemi et al., 2023; Shi et al., 2024; Don-Yehiya et al., 2024). For example, Zollo et al. (2024) show that models better predict preferred responses by a given user when the feedback history of the same user is provided. However, existing work has not examined whether LLMs adjust their generation given distinct feedback for the same instruction. Unfortunately, such evaluation is infeasible in most of the benchmarks, as they lack golden responses under multiple preferences for a single instruction.

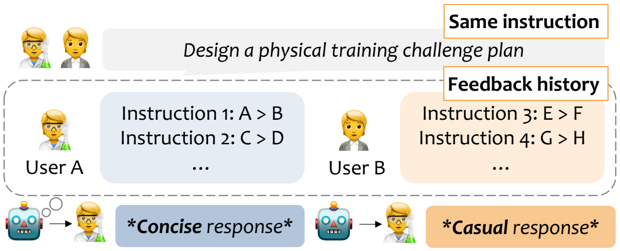

Figure 1. Our counterfactual evaluation setup of in-context personalized alignment. Given a common instruction, we evaluate whether model generates personalized responses given distinct feedback history.

In this paper, we revisit in-context personalized LLM alignment under a counterfactual evaluation setup, focusing on LLM's capability to steer their generation given distinct feedback history for the same instruction. We create training and evaluation data by transforming MultiFacet (Lee et al., 2024), a synthetic dataset of instructions and preferred responses under diverse preferences. Given an instruction, each of the candidate responses has an equal chance to be the preferred response depending on the feedback history. We evaluate (1) reward modeling, where models choose preferable responses from two candidates; and (2) response generation, both given the feedback history of a user. Figure 1 illustrates the setup.

Our experiments suggest that off-the-shelf LLMs fail to predicting preferred responses or generate personalized responses given past user feedback in-context. For example, GPT-4o only achieves 56% accuracy where the random

[*]Equal contribution [1]University of Southern California [2]Amazon Inc.. Correspondence to: Xisen Jin <xisenjin@usc.edu>.

ICML 2025 Workshop on Models of Human Feedback for AI Alignment, Vancouver, Canada. Copyright 2025 by the author(s).

guess accuracy is 50%. We show that fine-tuning is almost necessary so that LLMs can utilize in-context feedback, as fine-tuned 7B to 8B Mistral or Llama models improve over powerful off-the-shelf LLMs, classifying preferred responses at 67% accuracy and generating more personalized responses. To teach the models to leverage contexts better, we apply rejection sampling over the training data guided by the in-context personalized reward model. This strategy closes the gap between 7B LLMs and GPT-4o, evaluated with MultiFacet personalization metrics over the generated responses.

To summarize, the contributions of the paper are (1) examining in-context LLM personalization performance under counterfactual evaluation; (2) presenting approaches that better leverage in-context user feedback for personalized reward modeling and response generation.

2. Background

Personalized alignment from feedback history. We consider a setup where a user continuously interacts with the LLM, providing T instructions. The LLM returns two responses, and the user subjectively labels the preferred response. This results in a user feedback history of T preference pairs $\mathcal{H} = \{\langle x, y_+, y_- \rangle^{(t)}\}_{t=1}^{T}$ of T instructions x, preferred responses y_+, and rejected responses y_-. The goal of the model is to generate responses that align with individual user preferences.

In-context personalization with feedback history. Given a new instruction by the same user, the LLM takes the entire feedback history $\mathcal{H}$ as its input alongside the instruction. We never inform the models about ground truth user preferences; the models are expected to infer user preferences from the feedback history $\mathcal{H}$. We evaluate personalized reward modeling (identifying preferred responses between two candidate responses) and personalized response generation.

3. Counterfactual Evaluation of LLM Personalization

In this section, we introduce our created dataset, evaluation protocols, and models for training and counterfactual evaluation of personalized reward modeling and response generation.

3.1. Dataset Creation

We create training and evaluation data by transforming the MultiFacet collection (Lee et al., 2024). The MultiFacet collection is a synthetic dataset of 66k training and 307 test instructions. The candidate responses are generated by prompting GPT-4 with three plausible user preferences.

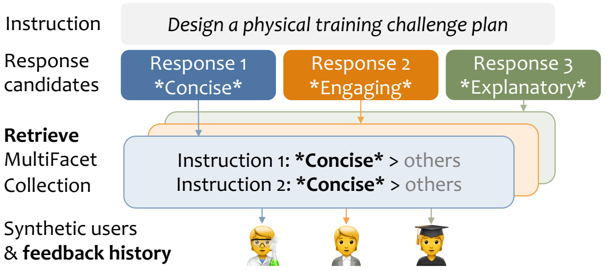
Instruction
Design a physical training challenge plan
Response candidates
Response 1 *Concise*
Response 2 *Engaging*
Response 3 *Explanatory*
Retrieve MultiFacet Collection
Instruction 1: *Concise* > others
Instruction 2: *Concise* > others
Synthetic users & feedback history

Figure 2. Transforming the MultiFacet dataset for our evaluation setup. Each instruction in MultiFacet is associated with preferable responses under three random preferences. The preferences are available in free-form text. We construct synthetic feedback history that represents a specific preference by retrieving preference pairs representing the same preferences from the dataset.

The preference specifies the preferred style, level of background knowledge, informativeness, and harmlessness of the responses. Among each of the four preference dimension (*e.g.*, style), the preference is given in free-form text (*e.g., engaging tone* or *conciseness* for style preferences). Each of the three candidate responses $y_j \in \{y_1, y_2, y_3\}$ corresponds to the reference (preferred) response under a preference. Accordingly, one of the other responses $y \neq y_j$ from $\{y_1, y_2, y_3\}$ is chosen as the rejected response, resulting in three preference pairs per instruction.

To examine in-context personalization with feedback history, we create a set of synthetic users, along with their feedback history, illustrated in Figure 2. We assume a user has consistent preference along a dimension *e.g*, conciseness for style preferences. We then retrieve the training split of MultiFacet for T=10 preference pairs that reflects the same preference, *e.g.*, examples where the concise response is preferred over the alternative. The set of retrieved preference pairs is considered as a synthetic feedback history $\mathcal{H}$ for the user. We process all training and test data, resulting in 66k training and 307 test instructions, each associated with three golden responses under three possible feedback history of T=10 preference pairs.

3.2. Evaluation Protocols

We separately evaluate personalization on four preference dimensions (style, background knowledge, informativeness, and harmlessness). The model takes the feedback history as input, but is never informed about the ground truth preferences.

For reward modeling, the model predicts a preferable response given a preference pair. We evaluate accuracy as the metrics. As each of the $\{y_1, y_2, y_3\}$ can be the preferred response under a preference, ignoring preferences would perform no better than the 50% accuracy of random guess.

	Style	Knowledge	Informative	Harmless
Off-the-shelf LLMs				
LLama-8B	55.54	48.18	53.53	53.15
Claude 3 Sonnet	51.99	50.13	50.54	51.25
GPT-4o	55.57	52.14	53.97	53.41
Fine-Tuned LLMs				
LLama-8B	**67.42**	**57.11**	**58.74**	**59.17**
Mistral-7B	63.45	53.53	52.77	57.20
Upperbound Reference				
GPT-4o w/ Preference	69.10	62.13	64.15	66.78

Table 1. Accuracy (%) of reward models in predicting preferred responses given the feedback history. Methods that do not incorporate feedback history would result in a random guess accuracy of 50%.

To evaluate the whether the generated responses align with the ground truth user preferences, we follow the official LLM-as-a-judge paradigm of MultiFacet collection. A powerful LLM (GPT4o in this case) acts as an evaluator and reads a scoring rubric (included in the original dataset), and judges how a response aligns with the ground truth preferences, scoring between 1 and 5.

3.3. In-Context Personalized Models

We compare two paradigms of in-context personalized reward modeling and response generation. (1) **Off-the-shelf models**, where we directly prompt off-the-shelf LLMs such as GPT4o, Claude-3 Sonnet with instructions and feedback history; and (2) **fine-tuned models**, where smaller models are fine-tuned while also prompted with instructions and feedback history. We detail two types of reward models and generation models below.

Reward models. For off-the-shelf LLMs, we experiment with GPT4o, LLama-3.1-8B-Instruct (Dubey et al., 2024) and Mistral-0.3-7B-Instruct (Jiang et al., 2023) as generative reward models. We include detailed prompts in Appendix B. We then fine-tune Bradley-Terry reward models (Ouyang et al., 2022) with smaller models (Llama3 8B or Mistral 7B). The models are fine-tuned to output a scalar reward score for a response given an instruction and the feedback history $\mathcal{H}$.

Response Generation Models. We evaluate the same set of LLMs as our off-the-shelf generative reward models. We then perform supervised fine-tuning (SFT) of smaller LLMs over instructions and the preferred responses in the training split. Afterwards, we optionally continue training the model with direct preference optimization (DPO) (Rafailov et al., 2023) on the preferred and rejected responses. Similarly to the reward models, the input of the response generation models consists of the feedback history and the instruction at both training and inference time.

	Without $\mathcal{H}$	With $\mathcal{H}$	Δ score
Off-the-shelf LLMs			
GPT-4o	3.264	3.497	7.14%
Claude 3 Sonnet	3.072	3.152	2.60%
Mistral-7B	2.641	2.672	1.17%
Fine-tuned Mistral-7B			
SFT	2.418	2.880	19.11%
DPO	2.612	3.042	16.46%

Table 2. LLM-based evaluation of how the generated responses align with ground truth *style* preferences (scoring between 1-5). We examine whether models take feedback history $\mathcal{H}$ into account by ablating the feedback history from the model inputs.

4. Experiments

In this section, we address research questions on (1) whether off-the-shelf LLMs are sufficient for in-context personalized reward modeling and response generation under our evaluation setup, and (2) whether fine-tuned models are capable of in-context personalization. We then explore ways to improve personalized response generation by leveraging personalized reward models.

4.1. Personalized Reward Modeling under Contrastive Evaluation

Table 1 summarizes the accuracy of reward models in predicting the preferred responses from the preference pairs. The reference performance is obtained by prompting GPT4o directly with ground truth preferences, which bypasses the need to infer preferences from the feedback history. We notice that, however, none of off-the-shelf LLMs improves clear over random guess accuracy (50%) when prompted with feedback history. Over the four preference dimensions, the best performing model (GPT4o) only improves over random guess by 2.1% to 5.6%. Our results indicate that fine-tuning is necessary so that LLMs can leverage feedback history provided in the context. In the style domain, for example, fine-tuned LLama-8B and Mistral-7B achieve accuracies of 67.4% and 63.5%, improving over 55.6% of GPT-4o.

4.2. Personalized Response Generation

Table 2 summarizes LLM-as-a-judge evaluation results of generated responses as we include or ablate the feedback history $\mathcal{H}$ in the model inputs. The evaluation examines whether the generated response align with the ground truth user preferences. We also report relative improvement (Δ score) by including $\mathcal{H}$. We notice that off-the-shelf LLMs improve marginally when feedback history is provided. The fine-tuned models (SFT and DPO), in contrast, improves clearly (Δ score ¿ 15%). The results indicates that simply appending feedback history to the model input may not lead to personalized response generation; instead, fine-tuning is

	Style	Knowledge	Informativeness
Fine-tuned LLama-8B			
SFT	2.915	2.865	2.745
DPO	2.929	2.914	2.808
RS-SFT	3.040	2.962	2.824
RS-DPO	**3.287**	**3.018**	**2.828**
Fine-tuned Mistral-7B			
SFT	2.418	2.918	2.863
DPO	2.612	2.933	2.945
RS-SFT	3.082	2.966	2.918
RS-DPO	**3.337**	**3.012**	**3.037**

Table 3. LLM-based evaluation of generated responses based on the scoring rubrics (1-5) that evaluates alignment to true preferences. We do not personalize for harmlessness under ethical considerations.

crucial so that models can leverage feedback history provided in-context.

4.3. Improving Personalized Response Generation with Reward Models

Our experiments suggest the challenge of incorporating past feedback history for personalization. We explore fine-tuning techniques to further enhance LLM's capability leverage the feedback history for personalized response generation.

Specifically, we apply a rejection sampling strategy so that the capability can be better taught from the training data. We filter out training data where the reward differences are marginal between two candidate responses given by the fine-tuned personalized reward models, keeping only p=60% of the training data. We use the fine-tuned Llama-8B in-context personalized reward models, which achieves the best performance in Table 1.

Table 3 summarizes the results of SFT and DPO with rejection sampling guided by the personalized reward models, noted as RS-SFT and RS-DPO. We notice that RS-SFT and RS-DPO improves over SFT or DPO. This improvement is pronounced on Mistral 7B models and the style domain, where the score improves from 2.4 to 3.1 for RS-SFT, and 2.6 to 3.3 for RS-DPO.

5. Related Works

Existing research on LLM personalization (Zhang et al., 2024) usually assumes self-stated preferences or user profiles (Yang et al., 2024; Wang et al., 2024a; Jang et al., 2023; Zhou et al., 2023). Feedback history is advantageous, as it can accumulate naturally as users interact with the system (Shi et al., 2024), and is broadly applied for eliciting preference labels (Lin et al., 2024; Li et al., 2023), or incorporated directly as model inputs or training data (Wang et al., 2024b; Tan et al., 2024; Zhuang et al., 2024; Lee et al., 2024). In-context learning has been studied in existing works as a lightweight approach for LLM alignment (Lin et al., 2023; Zhao et al., 2024). The paradigm is especially efficient for personalized alignment with feedback history (Zollo et al., 2024; Wang et al., 2023; Wu et al., 2024) for heterogeneous user preferences.

A key aspect of training and evaluating LLM personalization is dataset curation. Shi et al. (2024); Kirk et al. (2024) collect real user-LLM conversations along with the feedback of individual users, creating an ideal testbed of personalized alignment. The datasets, however, do not include counterfactual responses under diverse preferences. Bai et al. (2024) curates preferable responses of an instruction under diverse persona, which imposes additional challenges due to the complicated associations between persona and preferences. We leave learning preferences from real users and user-model interactions as future works.

6. Conclusions

In this paper, we examined in-context personalized alignment of LLMs with feedback history under a counterfactual evaluation setup. We created our dataset by transforming MultiFacet, and evaluated performance of reward models and response generation. Our results suggest that fine-tuning is necessary so that model can utilize feedback history provided in-context to personalize. We further examine a simple strategy to improve in-context personalized response generation with rejection sampling of training data guided by the personalized reward models.

Limitations

In this paper, we focus on the algorithmic challenge in personalized LLM alignment, examining the capability of LLMs to incorporate feedback provided in the context. For this purpose, our synthetic dataset greatly simplifies the user-model interactions and user preferences. In a real-world setting, user preferences can be more fine-grained and situational to their instructions (*e.g.,* user has different levels of background knowledge in various topics) and harder to be inferred from their feedback. In addition, user preferences are not stationary and may change over the interactions. We expect future works to move forward to more challenging and realistic setups of users and their interactions.

Besides, we did not touch on personalization approaches that learn explicit user representations or user-specific modules (Li et al., 2024; Park et al., 2024; Ning et al., 2024). Future works may evaluate these approaches under a counterfactual evaluation setup.

Ethical Considerations

We highlight the boundary of personalization in real-world scenarios. Personalized alignment aims to improve the helpfulness of LLMs for individual users; in practice, extra care should be taken to prevent models from favoring users without being helpful by hacking the rewards (Denison et al., 2024; Everitt & Hutter, 2019). Besides, model developers should actively prevent models from being hacked by the users to generate harmful contents.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Dassarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022. URL https://api.semanticscholar.org/CorpusID:248118878.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, D., Drain, S., Fort, Blodgett, S. L., Barocas, S., Daumé, H., Castricato, L., nathan lile, Anand, S., Chakraborty, S., Qiu, J., Yuan, H., Kop-681, A., Huang, F., Manocha, D., Singh, A., Mengdi, B., Wang, Maxmin-rlhf, Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, W., Shi, J., Xie, S., Li, R., Yang, Zhu, T., Chen, A., Li, N., Chen, C. L., Hu, S., Wu, S., Ren, Z., Fu, Y., Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., Schiefer, A., Askell, A., Bakhtin, Carol, Chen, Z., Hatfield-Dodds, D., Hernandez, Fränken, J.-P., Kwok, S., Ye, P., Gandhi, K., Arumugam, D., Moore, J., Tamkin, A., Gerstenberg, T., Zelikman, E., Rafailov, R., Gandhi, K., 2024, N. D. G., Kirk, H. R., Whitefield, A., Röttger, P., Bean, A. M., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., Chittepu, Y., Park, R., Sikchi, H. S., Hejna, J., Knox, B., Finn, C., Ramesh, S. S., Hu, Y., Chaimalas, I., Rettenberger, L., Reischl, M., Schutera, M., Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., Hashimoto, T., Shaikh, O., Lam, M., Shao, Y., Siththaranjan, A., Laidlaw, C., Sorensen, T., Fisher, J. R., Gordon, M., Mireshghallah, N., Rytting, M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., Stiennon, N., Long, O., Wu, J., Ziegler, R., Lowe, C., Voss, A., Radford, Amodei, D., Learn-860, C. ., Sun, C., Yang, R. K., Reddy, G., Chan, H. P., and Zhai, C. Persona: A reproducible testbed for pluralistic alignment. In *International Conference on Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:271602732.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P. J., Wang, T., Marks, S., Ségerie, C.-R., Carroll, M., Peng, A., Christoffersen, P. J. K., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., di Langosco, L. L., Hase, P., Biyik, E., Dragan, A. D., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback. *ArXiv*, abs/2307.15217, 2023. URL https://api.semanticscholar.org/CorpusID:260316010.

Denison, C. E., MacDiarmid, M. S., Barez, F., Duvenaud, D. K., Kravec, S., Marks, S., Schiefer, N., Soklaski, R., Tamkin, A., Kaplan, J., Shlegeris, B., Bowman, S. R., Perez, E., and Hubinger, E. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *ArXiv*, abs/2406.10162, 2024. URL https://api.semanticscholar.org/CorpusID:270521305.

Don-Yehiya, S., Choshen, L., and Abend, O. Learning from naturally occurring feedback. *ArXiv*, abs/2407.10944, 2024. URL https://api.semanticscholar.org/CorpusID:271212637.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024. URL https://api.semanticscholar.org/CorpusID:271571434.

Everitt, T. and Hutter, M. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *ArXiv*, abs/1908.04734, 2019. URL https://api.semanticscholar.org/CorpusID:199552156.

Jang, J., Kim, S., Lin, B. Y., Wang, Y., Hessel, J., Zettlemoyer, L. S., Hajishirzi, H., Choi, Y., and Ammanabrolu, P. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *ArXiv*, abs/2310.11564, 2023. URL https://api.semanticscholar.org/CorpusID:264289231.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b. *ArXiv*, abs/2310.06825, 2023. URL https://api.semanticscholar.org/CorpusID:263830494.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A. M., Margatina, K., Mosquera, R., Ciro, J. M., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S. A. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=DFr5hteojx.

Lee, S., Park, S. H., Kim, S., and Seo, M. Aligning to thousands of preferences via system message generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=recsheQ7e8.

Li, B. Z., Tamkin, A., Goodman, N. D., and Andreas, J. Eliciting human preferences with language models. *ArXiv*, abs/2310.11589, 2023. URL https://api.semanticscholar.org/CorpusID:264288747.

Li, X., Lipton, Z. C., and Leqi, L. Personalized language modeling from personalized human feedback. *ArXiv*, abs/2402.05133, 2024. URL https://api.semanticscholar.org/CorpusID:267547503.

Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., and Choi, Y. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*, 2023.

Lin, Y.-C., Neville, J., Stokes, J. W., Yang, L., Safavi, T., Wan, M., Counts, S., Suri, S., Andersen, R., Xu, X., Gupta, D., Jauhar, S. K., Song, X., Buscher, G., Tiwary, S., Hecht, B., and Teevan, J. Interpretable user satisfaction estimation for conversational systems with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 11100–11115, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.598. URL https://aclanthology.org/2024.acl-long.598/.

Ning, L., Liu, L., Wu, J., Wu, N., Berlowitz, D., Prakash, S., Green, B., O'Banion, S., and Xie, J. User-llm: Efficient llm contextualization with user embeddings. *ArXiv*, abs/2402.13598, 2024. URL https://api.semanticscholar.org/CorpusID:267770190.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Park, C., Liu, M., Kong, D., Zhang, K., and Ozdaglar, A. E. Rlhf from heterogeneous feedback via personalization and preference aggregation. *ArXiv*, abs/2405.00254, 2024. URL https://api.semanticscholar.org/CorpusID:269484177.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Salemi, A., Mysore, S., Bendersky, M., and Zamani, H. Lamp: When large language models meet personalization. *ArXiv*, abs/2304.11406, 2023. URL https://api.semanticscholar.org/CorpusID:258298303.

Shi, T., Wang, Z., Yang, L., Lin, Y.-C., He, Z., Wan, M., Zhou, P., Jauhar, S. K., Xu, X., Song, X., and Neville, J. Wildfeedback: Aligning llms with in-situ user interactions and feedback. *ArXiv*, abs/2408.15549, 2024. URL https://api.semanticscholar.org/CorpusID:271974624.

Sorensen, T., Jiang, L., Hwang, J. D., Levine, S., Pyatkin, V., West, P., Dziri, N., Lu, X., Rao, K., Bhagavatula, C., Sap, M., Tasioulas, J., and Choi, Y. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *AAAI Conference on Artificial Intelligence*, 2023. URL https://api.semanticscholar.org/CorpusID:261531157.

Tan, Z., Liu, Z., and Jiang, M. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://api.semanticscholar.org/CorpusID:270560467.

Wang, D., Yang, K., Zhu, H., Yang, X., Cohen, A., Li, L., and Tian, Y. Learning personalized alignment for evaluating open-ended text generation. In *Conference on Empirical Methods in Natural Language Processing*,

2023. URL https://api.semanticscholar.org/CorpusID:263671864.

Wang, H., Lin, Y., Xiong, W., Yang, R., Diao, S., Qiu, S., Zhao, H., and Zhang, T. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Annual Meeting of the Association for Computational Linguistics*, 2024a. URL https://api.semanticscholar.org/CorpusID:268063628.

Wang, X., Wang, Z., Liu, J., Chen, Y., Yuan, L., Peng, H., and Ji, H. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=jp3gWrMuIZ.

Wu, S., Fung, M., Qian, C., Kim, J., Hakkani-Tur, D., and Ji, H. Aligning llms with individual preferences via interaction. *ArXiv*, abs/2410.03642, 2024. URL https://api.semanticscholar.org/CorpusID:273162924.

Yang, R., Pan, X., Luo, F., Qiu, S., Zhong, H., Yu, D., and Chen, J. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *Forty-first International Conference on Machine Learning*, 2024. URL https://api.semanticscholar.org/CorpusID:267682397.

Zhang, Z., Rossi, R., Kveton, B., Shao, Y., Yang, D., Zamani, H., Dernoncourt, F., Barrow, J., Yu, T., Kim, S., Zhang, R., Gu, J., Derr, T., Chen, H., Wu, J.-Y., Chen, X., Wang, Z., Mitra, S., Lipka, N., Ahmed, N. K., and Wang, Y. Personalization of large language models: A survey. *ArXiv*, abs/2411.00027, 2024. URL https://api.semanticscholar.org/CorpusID:273798244.

Zhao, H., Andriushchenko, M., Croce, F., and Flammarion, N. Is in-context learning sufficient for instruction following in llms? *ArXiv*, abs/2405.19874, 2024. URL https://api.semanticscholar.org/CorpusID:270123077.

Zhou, Z., Liu, J., Shao, J., Yue, X., Yang, C., Ouyang, W., and Qiao, Y. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:264175263.

Zhuang, Y., Sun, H., Yu, Y., Qiang, R., Wang, Q., Zhang, C., and Dai, B. HYDRA: Model factorization framework for black-box LLM personalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=CKgNgKmHYp.

Zollo, T. P., Siah, A. W. T., Ye, N., Li, A., and Namkoong, H. Personalllm: Tailoring llms to individual preferences. *ArXiv*, abs/2409.20296, 2024. URL https://api.semanticscholar.org/CorpusID:272987425.

A. Training, Evaluation, and Dataset Details

Fine-Tuning Configurations. The maximum input length of the models are set as 12,800 tokens, which accommodates approximates 10 preference pairs as feedback history.

For reward models, we fine-tune Llama3-8B-Instruct or Mistral-0.3-7B-Instruct with a learning rate of 5e-6 and an effective batch size of 256 on 4 Quadro RTX A6000 GPUs. We train the model for 3 epochs over a 1/3 subset of the training dataset.

For response generation models, we perform SFT of Llama3-8B and Mistral-0.3-7B models with a learning rate of 5e-6 and an effective batch size of 32 on 4 Quadro RTX A6000 GPUs. We train the model for 1 epoch over the 1/3 subset of the training dataset. We perform DPO with a learning rate of 5e-6 and an effective batch size of 64 on 4 A100 gpus on the rest 2/3 of the datasets.

Generation Configurations. We use gpt-4o-2024-05-13 for reward modeling, response generation, and LLM-as-a-judge evaluation. We use default generation hyperparameters. For Claude, LLama3, and Mistral generations, we set the temperature hyperparameter as 0.9. Reported performance is obtained in a single run.

Dataset Details. We create our dataset by transforming MultiFacet Collections (Lee et al., 2024), which is released under the Creative Commons Attribution 4.0 license. According to the original data documentation and our best knowledge, the dataset does not contain harmful instructions.

B. Prompts

B.1. In-context personalized reward modeling

```
###Task Description:
In each of the examples below, you will see
     pairs of responses generated by AI
    assistants given an instruction, where
    one of them is more preferred by the
    user. All preferences are provided by
    the same user, with their own personal
    preference and bias. Your goal is to
    infer the true preference of the user,
    and predict which of the responses for
    a new instruction will be more
    preferred by the user.

# START OF EXAMPLES

{feedback_history}
# END OF EXAMPLES

###The instruction to evaluate:
{instruction}

###Response A:
{response_a}

###End of Response A

###Response B:
{response_b}

###End of Response B


Now, predict which of the reponse is more
    preferred by the user for a new
    instruction. Remember the user has
    their personal preferences. Do not
    allow the length of the responses to
    influence your evaluation. Remember
    that after providing your explanation,
    output your final verdict by strictly
    following this format: "[[A]]" if
    response A is more likely preferred,
    "[[B]]" if response B is more likely
    preferred. Let's think step by step.
```

B.2. In-context personalized response generation

```
# Goal

Your goal is to generate responses that
    adhere to preferences of the user. Here
     are some examples of responses that
    are liked and disliked by the user.
    Infer the preference of the user, and
    respond to the new instruction from the
     user.

# START OF EXAMPLES

{feedback_history}
# END OF EXAMPLES

Now, here is the new instruction from the
    user.

# Instruction

{instruction}

<|assistant|>
```

B.3. Feedback History

The feedback history consists of 10 preference pair formatted as below.

```
## Example {example_id}:

### Instruction

{instruction}

### Response liked by the user

{preferred_response}
```

```
### Response NOT liked by the user

{rejected_response}
```