PersonalLLM: Tailoring LLMs to Individual Preferences

Anonymous Author(s)
Affiliation
Address
email

Abstract

As LLMs become capable of complex tasks, there is growing potential for personalized interactions tailored to the subtle and idiosyncratic preferences of the user. We present a public benchmark, PersonalLLM, focusing on adapting LLMs to provide maximal benefits for a particular user. Departing from existing alignment benchmarks that implicitly assume uniform preferences, we curate open-ended prompts paired with many high-quality answers over which users would be expected to display heterogeneous latent preferences. Instead of persona prompting LLMs based on high-level attributes (e.g., user race or response length) that yields homogeneous preferences relative to humans, we develop a method that can simulate diverse preferences from a set of pre-trained reward models. Our dataset and generated personalities offer an innovative testbed for developing personalization algorithms that grapple with continual data sparsity—few relevant feedback from the particular user—by leveraging historical data from other (similar) users. We explore basic in-context learning and meta-learning baselines to illustrate the utility of PersonalLLM and highlight the need for future methodological development.

1 Introduction

The *alignment* of LLMs with human preferences has recently received much attention, with a focus on adapting model outputs to reflect universal population-level values. A typical goal is to take a pre-trained model that cannot reliably follow complex user instructions [32] and can easily be made to produce dangerous and offensive responses [25], and adapt it to the instructions of its user base [24] or train a generally helpful and harmless assistant [1]. By assuming a *uniform preference* across the population, recent successes [35, 24, 6] demonstrate the feasibility of learning and optimizing a monolithic preference ("reward model"). Alignment techniques have provided the basis for popular commercial applications like ChatGPT, as well as instruction-tuned open-source models [30].

The rapid advancement in LLM capabilities opens the door to an even more refined notion of human preference alignment: personalization. A personalized model should adapt to the preferences and needs of a particular user, and provide maximal benefits as it accumulates interactions (see Figure 1). Given the expected data sparsity in this setting, beyond a particular user's data, such personalized language systems will likely also rely on historical data from other (similar) users in order to learn how to learn from a small set of new user feedback. For instance, personalized learning experiences could be crafted by adapting educational chat assistants to the specific learning pace and style of individual students based on previous successful interactions with similar students. Customer support chatbots could offer more accurate and empathetic responses by drawing on a wealth of previous interactions, leading to quicker resolution times and higher customer satisfaction. In healthcare, personalized chatbots could provide tailored advice and support to patients based on their users with relevant medical history and communication preferences By discovering patterns across users, these

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

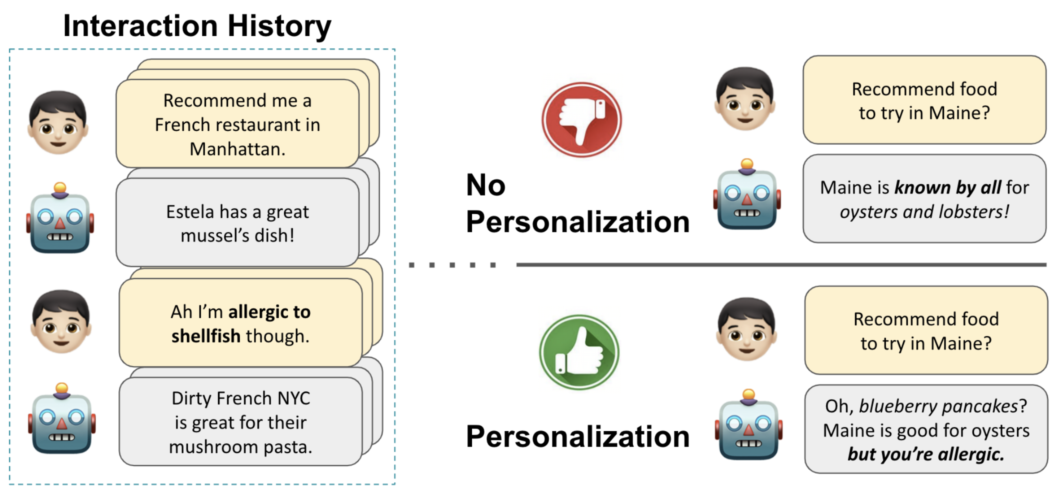

Figure 1: Standard LLMs require tedious re-prompting to learn a user's preferences in each session. PersonalLLM learns a unique user's diverse preferences to maximize long-term satisfaction.

systems will be able to efficiently optimize their responses, ultimately leading to more effective and beneficial conversational AI.

Departing from standard applications where prompts have a uniform notion of "ground truth" (e.g., question answering), the study of true personalization requires open-ended prompts where among many high-quality answers, different users exhibit heterogeneous preferences. While personal preferences may vary according to simple features like user age [5, 4] and answer length and technicality [21], they also involve more abstract dimensions of culture, politics, and language [17], as well as aspects of personality that are difficult to explain [13]. A personalized LLM should be able to adapt to subtle, idiosyncratic, and sometimes sensitive differences between user tastes as it gathers more interactions.

Inspired by the vision of a future with personalized AI, we introduce PersonalLLM, a public, open-source benchmark designed to adapt LLMs to provide maximal benefits for individual users. In order to explore complex differences in user tastes, our benchmark features a set of prompts with many high-quality LLM responses (from state-of-the-art LLMs like GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro), such that humans *are expected* to express diverse preferences over the responses. Such an approach to dataset-building stands in contrast to existing alignment datasets, where responses exhibit observable quality differences (see Figure 2). For each prompt and set of responses, our dataset also includes scores from a set of 10 reward models with heterogeneous preferences over those responses. We leverage these reward models to sample many new "users" (or personal preference models) via weighted ensembles of their preferences, and in doing so we are able to *simulate an entire user base*, which we argue to be a critical ingredient in a truly useful personalization benchmark. Through extensive analysis of the preferences of these users over our dataset, we show these simulated personal preference models to be diverse and non-trivial (e.g., with respect to length, formatting, or tone), and illustrate the difficulty of creating such an environment by comparing to the increasingly popular persona prompting baseline [4, 5, 15], which produces preferences only half as diverse as a set of PersonalLLM users across multiple metrics. Taken together, the prompts, responses, and personalities present in PersonalLLM offer an innovative tested for benchmarking personalization algorithms as they tailor interactions based on previous interactions with an individual user.

While fine-tuning and reinforcement learning approaches [29, 26] are effective for aligning to population-level preferences, personalization requires a new algorithmic toolkit, as it is not practical to gather enough data or store a separate copy of the model or even low-rank adapter weights [12] for every user. PersonalLLM offers the versatility necessary to spur development across a range of new approaches to personalization: in-context learning (ICL) [3], retrieval augmented generation (RAG) [20], ranking agents, efficient fine-tuning, and other adaptation techniques. In our experiments, we highlight a particularly salient challenge inspired by the recommendations setting: since the space of "actions/responses" is prohibitively large to be able to explore based on interactions on a single user, we want to *learn across users*. We model this as a meta-learning problem, where the goal is to leverage a wealth of prior interactions from historical users to tailor responses for a new user who do not have a significant interaction history.

Motivated by key methodological gaps in personalizing LLMs, here we summarize our contributions:

- We release a new open-source dataset with over 10K open-ended prompts paired with 8 high-quality responses from top LLMs.
- We propose a novel method for sampling "users" (i.e., personal preference models) that, unlike existing methods, creates diverse preferences and allows for the simulation of large historical user bases.
- We illustrate new possibilities for algorithmic development in learning *across* users.

Our goal in creating the open-source PersonalLLM testbed is to facilitate work on methods to personalize the output of an LLM to the individual tastes of many diverse users. We do not claim our simulated personal preference models provide a high-fidelity depiction of human behavior, but rather offer a challenging simulation environment that provides the empirical foundation for methodological innovation in capturing the complex array of human preferences that arise in practice. As an analogy, while ImageNet [27] is noisy and synthetic—e.g., differentiating between 120 dog breeds is not a realistic vision task—it provides a challenging enough setting that methodological progress on ImageNet implies progress on real applications. Similarly, we believe PersonalLLM is a reasonable initial step toward the personalization of language-based agents, building on the common reinforcement learning paradigm of benchmarking personalization algorithms with simulated rewards [34, 14].

2 PersonalLLM

Our PersonalLLM testbed is composed of two high-level components: 1) A dataset of prompts, each paired with a set of high-quality responses among which humans would be expected to display diverse preferences. 2) A method for sampling diverse personal preference models, such that we can test methods for personalization using these "personas" as our simulated users. Next, we will describe each of them in detail. Our data and code will be publicly available and actively maintained.

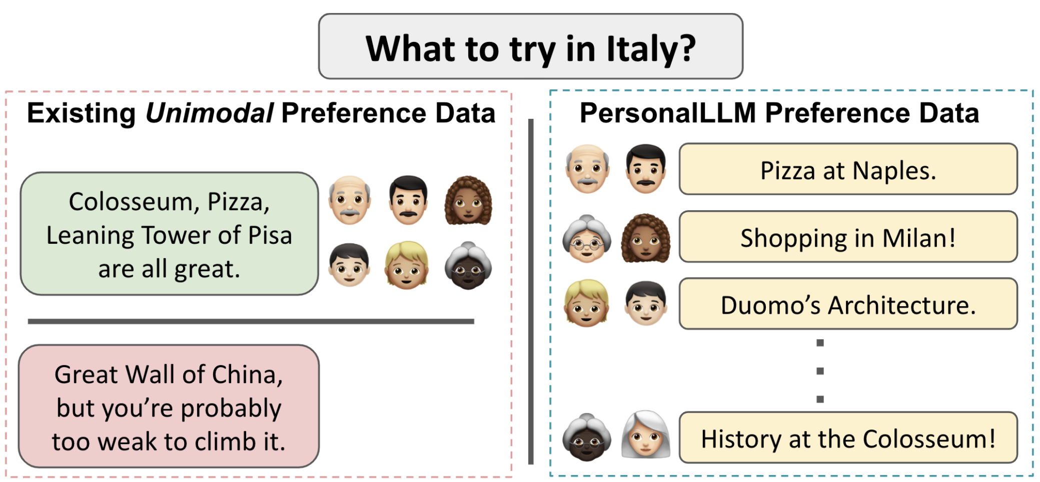

Figure 2: **Left:** Existing alignment datasets contain prompts paired with multiple responses, where the majority of people are expected to prefer one specific response (e.g., a harmless response). **Right:** Meanwhile, our dataset consists of prompts paired with many high-quality responses, each favored by different personas. Such a dataset induces diverse preferences in our personal preference models, creating a testbed to build PersonalLLMs.

2.1 Dataset

Since our goal is to study diverse preferences, our first focus was the collection of *open-ended* prompts, similar to a chat setting. As a source of these open-ended prompts, we compiled 37,919 prompts from Anthropic Helpful-online, Anthropic Helpful-base [1], Nvidia Helpsteer [31], and RewardBench [19]. From this set, prompts were filtered to those with a length of 2400 characters or fewer as most reward models are limited to 4096 context length. We then randomly drew 10,402 prompts to form our final set. Our next aim was to collect many high-quality responses for each prompt. The hope is that responses vary not in terms of undesirable contents (like misinformation or toxicity) or obvious dimensions of helpfulness or length, as is typical in RLHF datasets, but instead with respect to interesting dimensions of personal preference like political viewpoint and culture, as well as difficult to describe latent features. To achieve this, we generated eight responses for each of these 10,402 prompts using a selection of the top models from ChatArena and other important benchmarks: **GPT-4o, Claude 3 Opus, Gemini-Pro-1.5, Command-R-Plus, GPT-4-Turbo, Claude 3 Sonnet, Llama3-70B-Instruct, and Mixtral 8x22B**. We split the resulting dataset into 9,402 training examples and 1,000 test examples.

2.2 Simulating Personal Preference Models

We design our approach to creating simulated PersonalLLM users with several goals in mind. First, we aim for PersonalLLM to allow for the simulation of a large number of users, enabling study of the full personalization paradigm for applications such as search engines and recommender systems [8, 7, 33, 11] wherein a historical database of user data is leveraged to personalize new interactions. Next, when applied to our dataset, our preference models should allow for the study of alignment based on diverse and complex latent preferences, as opposed to simple attributes such as answer length or sensitive and reductive user characteristics, for example race or gender. Finally, our evaluation should not rely on GPT4, which can be cost-prohibitive and less than ideal for research purposes given model opacity and drift. While human evaluation like that of Kirk et al. [17] is a gold standard, wherein fine-grained preference feedback is gathered from a representative sample of diverse, multicultural participants, it is impractical or even impossible to get this feedback throughout the methodology development cycle, meaning that synthetic personal preference models will ultimately be necessary.

To overcome these challenges, we propose a solution based on a set of strong open-source RLHF reward models, which we find to have diverse preferences over our dataset given its differences relative to typical monolithic RLHF datasets. Since the number of existing top-quality reward models is much smaller than the number of users we would like to simulate, we propose to generate users by sampling weightings over the set of reward models, such that the reward score assigned to a (prompt, response) pair by a user is a weighted sum of the reward scores assigned by the pre-trained reward models. Technical details can be found in Section E.

3 Scope of Study

Given space limitations, the remainder of our study is deferred to the Appendix. In summary:

- In Section A, we offer extensive analysis of simulated populations of PersonalLLM users. We find them to produce heterogeneous preferences over our dataset of prompts and responses, display reasonable and diverse preferences with respect to syntactic and semantic content of prompts, and simulate a user base that better represents diverse human opinions than many popular LLMs, without resorting to explicit stereotyping.
- In Section B, we perform experiments in personalized in-context learning and meta-learning personalization across users, highlight key questions and the need for new methodology.
- In Section D, we discuss the opportunities, risks, and limitations of our work.

References

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[2] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/P04-3031`.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[4] Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. Persona: A reproducible testbed for pluralistic alignment, 2024. URL `https://arxiv.org/abs/2407.17387`.

[5] Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2024. URL `https://arxiv.org/abs/2406.20094`.

[6] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.

[7] Abhinandan S Das, Mayur Datar, Ashutosh Garg, and Shyam Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, pages 271–280, 2007.

[8] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296, 2010.

[9] Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

[10] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback, 2024.

[11] Michael Färber and Adam Jatowt. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405, August 2020. ISSN 1432-1300. doi: 10.1007/s00799-020-00288-2. URL `http://dx.doi.org/10.1007/s00799-020-00288-2`.

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[13] EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. Aligning language models to user opinions, 2023. URL `https://arxiv.org/abs/2305.14929`.

[14] Eugene Ie, Chih wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. Recsim: A configurable simulation platform for recommender systems, 2019. URL `https://arxiv.org/abs/1909.04847`.

[15] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging, 2023. URL https://arxiv.org/abs/2310.11564.

[16] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023.

[17] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models, 2024. URL https://arxiv.org/abs/2404.16019.

[18] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023.

[19] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024.

[20] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.

[21] Xinyu Li, Zachary C. Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback, 2024. URL https://arxiv.org/abs/2402.05133.

[22] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base llms: Rethinking alignment via in-context learning, 2023. URL https://arxiv.org/abs/2312.01552.

[23] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022.

[24] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[25] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models, 2022.

[26] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. URL https://arxiv.org/abs/1409.0575.

[28] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect?, 2023.

[29] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[30] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[31] Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. Helpsteer: Multi-attribute helpfulness dataset for steerlm, 2023. URL `https://arxiv.org/abs/2311.09528`.

[32] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022.

[33] Jiajing Xu, Andrew Zhai, and Charles Rosenberg. Rethinking personalized ranking at pinterest: An end-to-end approach. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22. ACM, September 2022. doi: 10.1145/3523227.3547394. URL `http://dx.doi.org/10.1145/3523227.3547394`.

[34] Kesen Zhao, Shuchang Liu, Qingpeng Cai, Xiangyu Zhao, Ziru Liu, Dong Zheng, Peng Jiang, and Kun Gai. Kuaisim: A comprehensive simulator for recommender systems, 2023. URL `https://arxiv.org/abs/2309.12645`.

[35] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020.

A Analyzing PersonalLLM

Next, in order to validate our testbed, we explore the preferences exhibited by our simulated users over the PersonalLLM dataset.

A.1 Preference Diversity and Comparison to Persona Prompting

First, we examine whether populations of personal preference models sampled via the method outlined in Section 2.2 do in fact display heterogeneous preferences over the prompt/response pair in our dataset. In Figure 5 (left 3 columns), we provide experimental results for user bases of 1,000 PersonalLLM personal preference models sampled with parameters $\alpha = [0.01, 0.05, 0.1]$ and applied to the PersonalLLM test set to choose winning responses among the 8 included. The top row displays the percentage of prompts in the dataset for which the most popular winning response according to the population receives no more than 50%, 75%, and 95% of the population vote; higher values indicate more diversity in preferred responses. The middle row shows the percentage of prompts that have a given number of responses with at least one winning vote across the population; heterogeneous population preferences induce higher concentration on the right side of each plot. On bottom, we the overall win rates for each LLM across all users and prompts.

In the right column, we offer results for a persona prompting baseline. Persona prompting [4, 5, 15] is an emerging method for evaluating methods for LLM personalization, wherein an LLM, often GPT-4, is prompted to decide which response would be preferred by a person of a particular race, gender, age, profession, or other demographic category. While we could argue that such evaluation is *prima facie* discriminatory and reductive, and therefore not a desirable standard for algorithmic advancement, especially in sensitive areas, we are also interested in whether persona prompting meets the technical challenge of producing a simulation environment with a high degree of heterogeneity. For our baseline, we prompt the sfairXC/FsfairX-LLaMA3-RM-v0.1 reward model [9] to score responses with respect to 500 personas randomly sampled from PersonaHub Chan et al. [5], a recent effort at building a database of personas that are representative of a pluralistic population.

Observing results in Figure 5, for PersonalLLM personas, we can see that the top response receives a majority user vote for only about half of the prompts, while that figure is closer to 90% for the persona prompting baseline. Also, for roughly 60% of prompts, at least 5 different answers are chosen as the best by at least 1 under our set of personas; for LLM persona prompting, it is roughly 30%. Finally, our ensembled preference models have a fairly diffuse set of preferences over the response-generating LLMs, while persona prompting strongly prefers a subset of 4 models. With respect to changes across the left 3 columns, we can observe that as α increases, preferences become more uniform. However, if α is set too low, user preferences cluster very tightly around the base reward models; we observe this behavior for $\alpha = 0.01$.

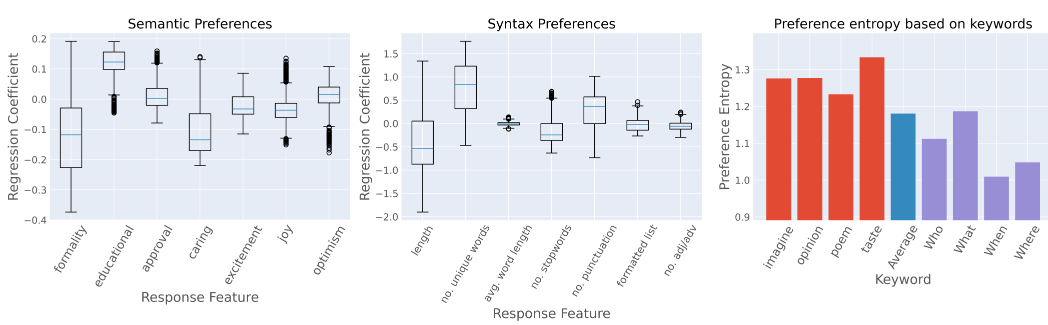

Figure 3: Analysis of simulated user preferences with respect to prompt and response contents.

A.2 Effects of Semantics and Syntax

We further analyze the effects of semantics and syntax on the preferences of a simulated user base (with $\alpha = 0.05$ and 1,000 users). We use regression analysis to understand how different features may drive the preferences of different users, including semantic response features such as the formality or educational value or the expressions of certain emotions (approval, caring, excitement, joy, optimism), as well as syntactic features like length and the use of different parts of speech and formatting. For each user, we gather their most and least preferred responses for each of the test prompts, and create a binary prediction problem to predict whether a given response is a winning or losing response. Responses are embedded using hand-crafted features (based on either syntax or semantics, which are studied separately), and a unique logistic regression model is trained *for each user*. Semantic features were captured using pretrained classifiers, while syntactic features were engineered using nltk [2]. See Appendix XX complete details.

In Figure 3 (left and middle), for each feature we show a box plot with the resultant regression coefficient for each feature across users. A positive coefficient suggests a feature associated with winning responses, while a negative coefficients suggests a feature's role in losing response. A tight box indicates homogeneous preferences, while greater spread represents heterogeneity. Here, we can see a reasonable mix of heterogeneity and homogeneity across user preferences for different features. Semantically, users tend to prefer responses with educational value and dislike highly formal responses, although the size of these preferences varies. Encouragingly, syntactic preferences do not seem to be driven by uniform preferences for simple features like length or the presence of formatting list bullets or lists.

In Figure 3 (right), we compare the entropy in the population preferences over the responses to a given prompt based on keywords, comparing words we would expect to inspire heterogeneity (e.g., imagine, opinion, poem) to prompts beginning with who, when, and where, which evoke more objective answers. We can see that the presence of these subjective cues leads to a more diverse set of preferences than those seeking simple entity or date responses. Such diversity among the prompts creates a setting where an algorithm *must not only learn how to personalize, but also when to personalize*.

A.3 Comparison to Human Preferences

Finally, to understand how our simulated personal preference models over relate to human preferences over text responses, we surveyed a population of our simulated personal preference models on a set of questions with responses where a large and diverse set of humans have given their preferences in the past, the OpinionQA dataset, emulating the work of [28]. OpinionQA is an appropriate validation set for our personas given that its broad coverage of topics (e.g., science, economics, politics, romance, and many other topics) aligns with the open-domain nature of our prompt set. Following this previous work, we calculate the representativeness score of the opinion distribution given by our simulated preference models using the Wasserstein distance of the synthetic population preferences from that of real human populations. To have a high representativeness score, our simulated user population would have to display heterogeneous preferences over question/response sets where humans do so, and produce homogeneous (and matching) preferences in cases where humans do the same.

Our population of simulated users produces a score of 0.839 with respect to the overall population of the US, higher than any LLM in the original study and near as representative of the overall population as some real, large demographic group. Further, in Table 1 we can see that our simulated users produce opinions that better represent a wide range of important (and sometimes protected) groups according to demographic attributes such as race, political leaning, religion, marital status, and more. In fact, this is the case for 59 of 60 demographic groups in their study (see Appendix Section F).

A.4 Summary of Analysis

Taken together, these results show that our simulated user reward models: 1) produce heterogeneous preferences over our dataset of prompts and responses, considerably more so than persona prompting an LLM, 2) display reasonable and diverse preferences with respect to syntactic and semantic content of prompts, and 3) simulate a user base that better represents diverse human opinions than many popular LLMs, without resorting to explicit stereotyping.

B Personalization Experiments

The personalization setting is often plagued by a lack of data, as most users will have a relatively sparse interaction history, and many fewer datapoints than is required to effectively fine-tune an LLM. Two first-order problems emerge from such an environment: 1) how to best leverage small amounts of

	AI21 Labs			OpenAI	PersonalLLM
Demographic	j1-jumbo	j1-grande-v2	ada	text-davinci-003	**Ours**
Asian	0.814	0.806	0.819	0.708	**0.839**
Black	0.820	0.812	0.823	0.702	**0.833**
Hispanic	0.820	0.810	0.824	0.706	**0.839**
White	0.807	0.794	0.817	0.699	**0.832**
Conservative	0.796	0.780	0.810	0.684	**0.817**
Liberal	0.792	0.788	0.799	0.721	**0.833**
Democrat	0.800	0.795	0.804	0.719	**0.834**
Republican	0.791	0.776	0.805	0.680	**0.812**
Muslim	0.794	0.788	0.792	0.697	**0.816**
Roman Catholic	0.816	0.806	0.823	0.702	**0.835**
Less than $30,000	0.828	0.813	0.833	0.693	**0.838**
$100,000 or more	0.797	0.790	0.807	0.708	**0.831**
18-29	0.818	0.808	0.828	0.700	**0.840**
65+	0.792	0.779	0.800	0.699	**0.818**
Divorced	0.809	0.796	0.817	0.696	**0.830**
Married	0.810	0.799	0.819	0.699	**0.832**

Table 1: Representativeness scores in relation to real human opinions from important demographic groups for different LLMs, as well as our PersonalLLM population.

user-specific data for personalized adaptation and 2) how to lookup similar users based on language feedback.

In order to illustrate how researchers might approach these problems, we perform experiments in two modal settings for LLM personalization research. First, we explore a scenario where we have access to a short but relevant interaction history for the user, and we aim to efficiently leverage that interaction history through ICL. Then, we explore a more complex setting that fully leverages the advantages of PersonalLLM, where the current user possibly has no relevant interaction history, and we must instead retrieve relevant interactions from similar users in a database. Overall, our results validate the solid empirical foundations of PersonalLLM while highlighting salient algorithmic questions and the fact that there is much room for improvement in terms of personalization performance.

All experiments simulate a chatbot using in-context learning to personalize responses for a test set of new users. Our test set simulates 1,000 personal preference models (or "users") drawn with $\alpha = 0.05$ (as in the analysis in Section A), and each user is associated with one test prompt from the PersonalLLM test split. For a new user with an associated test prompt, the goal is to use ICL to produce a response to maximize the reward (and win rate vs. GPT4o) given by the user's personal preference model (i.e., weighted ensemble of reward models). Our underlying chatbot is Llama3-8B-Instruct. Further details for each individual experiment are given below.

B.1 Personalized In-Context Learning

While ICL for broad alignment has been studied to some extent [22], the problem may be different when the underlying preference model is idiosyncratic and may cut against pretraining and RLHF dataset biases. In our initial set of experiments, we focus on a setting wherein we have a small set of useful data for the sake of personalizing the response to a given query, i.e., feedback gathered from the same user on similar prompts. By doing so, we can study key questions related to personalized inference with ICL, which may form the basis for more complex systems involving, e.g., looking up similar users.

B.1.1 Experiment Details

For each of our 1,000 test users, each with their own test prompt, we build a short but relevant interaction history by retrieving 5 other prompts based on embedding similarity. We build a winning/losing response pair for each prompt based on each user's most and least preferred answers from the 8 models in our dataset. In order to establish baseline results on key questions in personalization, we include several baselines for how these interaction samples are leveraged in-context during inference:

- **Winning and Losing:** Both the winning and losing responses are included.
- **Winning only:** Only the winning response is included.
- **Losing only:** Only the losing response is included.
- **Losing only (Mislabeled):** Only the losing response is included, and it is mislabeled as a winning response.

Inference is performed using 1, 3, and 5 such examples (see Appendix I for exact templates), and evaluated by scoring with each user's (weighted-ensembled) preference model. We also compare to a zero-shot baseline, with no personalization.

B.1.2 Results

Results are shown in Figure 4. We can see that the best performance comes from ICL with only winning examples. This underlines the outstanding issue of training LLMs to not only mimic winning responses in-context, but also leverage the contrast between winning and losing responses, especially when the differences may not described in the model's training data. Any amount of examples, even incorrectly labeled, are helpful relative to zero-shot; this may be unsurprising, as all 8 models in our dataset are stronger than our 8B parameter chat model. One interesting result lies in the comparison between Losing Only and Losing Only (Mislabeled). While the mislabeled examples may help performance versus a zero-shot baseline (once again because they are from a stronger underlying LLM), Llama-8B-Instruct gains more from having these relatively strong losing responses labeled as losing. Overall, our findings reflect that a model trained for broad alignment does have some of the necessary capabilities to do idiosyncractic personalization using only in-context examples, but that much work is left in order to fully leverage this language feedback.

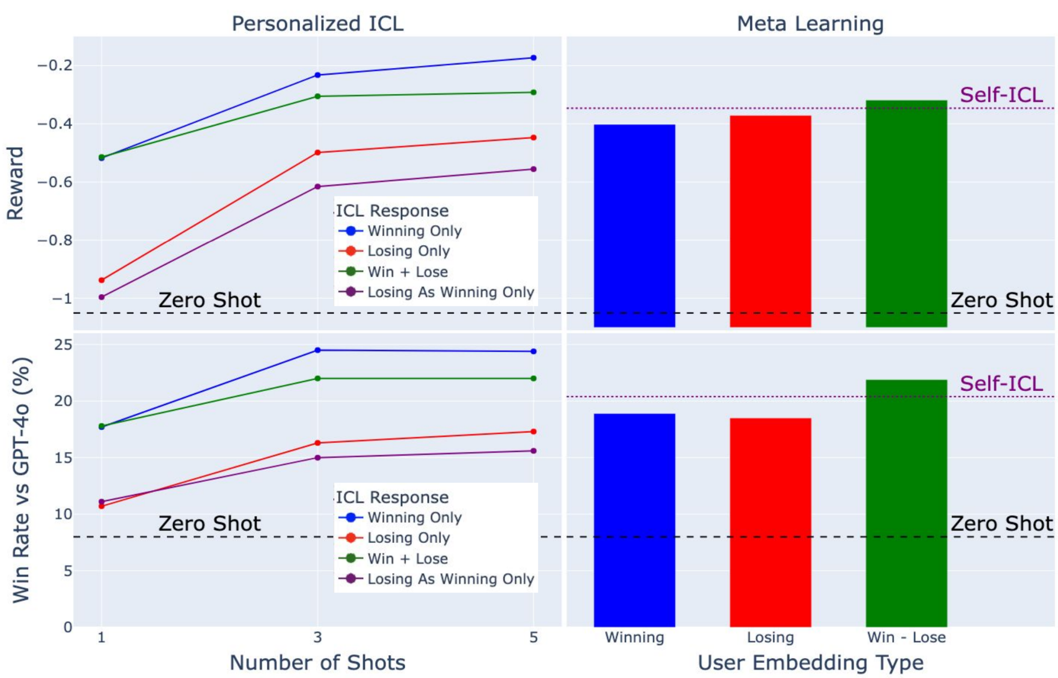

Figure 4: Results across different personalization algorithms. **(Left)** Test users are accompanied by a relevant interaction history with pairwise preference feedback, and we explore the LLM's ability to exploit this information in context. **(Right)** Test users have interaction histories that are not relevant to their test prompt, and we probe methods for embedding users based on language feedback to retrieve useful examples for ICL.

B.2 Learning Across Users

Having established some empirical foundations for in-context personalization with PersonalLLM, we next highlight a particularly significant challenge prevalent in practice that has been under-explored in the LLM community: the cold-start problem. When a new user with limited prior interaction data arrives, or a user inquires about a new topic, prior user interactions alone cannot inform a satisfactory response. We model this challenge as a meta-learning problem, where the goal is to utilize a rich reservoir of prior interactions with a diverse set of users. We are motivated by real-world scenarios where we have access to a proprietary database containing extensive interaction histories from previous users. When a new user arrives, our goal is to utilize this rich, heterogeneous dataset to provide the best possible response to the new user's query despite having only limited initial interactions with them that may not be relevant to the current query. This setting resembles typical recommendation systems, but "actions" are now defined over the space of natural language outputs instead of a fixed set of items.

B.2.1 Experiment Details

For each of our 1,000 test users, we build a short but, in contrast to our first experiment, *possibly irrelevant* interaction history by retrieving 5 random prompts. Winning/losing response pairs (i.e., preference feedback) are selected as before. In order to supplement these interaction histories, we sample a historical database of 10,000 users (also with $\alpha = 0.05$), each with a set of 50 prompt, winning response, losing response triplets from the train set, where the prompts are selected randomly and the winning and losing responses are selected as the historical user's highest and lowest scoring among the 8.

We compare 3 methods for embedding users for lookup:

- **Winning minus Losing:** Average direction in embedding space between winning and losing responses for each prompt.
- **Winning only:** Average direction in embedding space for winning responses.
- **Losing only:** Average direction in embedding space for losing responses.

For each test user, we build a set of candidate prompt/feedback data by retrieving the 20 most similar historical users based on these embeddings, and then of the pool created by those users' interaction histories, retrieving $k = [1, 3, 5]$ examples for in-context learning based on prompt embedding similarity to the user's test prompt. We compare to a **Self-ICL** baseline, where the test user's possibly irrelevant prompt/feedback history is used for ICL. Evaluation is done as before.

B.2.2 Results

Our results are shown in Figure 4. We find that using the strongest user embedding method, which most fully exploits the available pairwise preference feedback, meta-learning can beat the self-ICL baseline. This positive result for meta-learning highlights the opportunity created by leveraging historical user data, and the feasibility of embedding users based on a small amount of language feedback. However, the gain from our relatively naive method is small, illustrating the need for methodological innovation in building such systems.

C Related Work

Preference Datasets Recent developments in large language models (LLMs) emphasize the importance of *aligning* LLMs based on *preference feedback* rather than merely pre-training on large corpora of language in a self-supervised manner. Consequently, there has been a surge in the creation of open-source datasets [1, 23, 18, 10, 19] designed to support research on alignment methodologies. A significant limitation in the existing datasets is that they mainly enable fine-tuning to a single high-level notion of alignment that is uniform across the population, such as instruction-following in RLHF [24] and helpfulness and harmlessness [1].

Personalization Personalization has been extensively researched across different fields, with previous datasets primarily focusing on applications such as search engines and recommender systems [8, 7, 33, 11]. Recently, given the success of population-level alignment, researchers have begun to develop testbeds and methodology wherein the goal is to achieve a more granular level of personalized alignment for LLMs [4, 15, 17, 21]. Much of this work has focused on alignment for real or synthetic personas based on high-level attributes like race or occupation [4, 5], or high-level notions of alignment with respect to response qualities like length, technicality, and style. For example, Jang et al. [15] decomposes personal preferences along a handful of easily observable dimensions and performs personalized generation by merging models trained with different preference data based on these dimensions. Evaluation is often done by prompting GPT4 to select the preferred response based on preferences stated in its prompt [15, 4]. In an effort to highlight the need for broad participation and representation in LLM alignment, the PRISM dataset collects user-profiles and personalized preference feedback from over 1,000 diverse human participants.

D Discussion

We present PersonalLLM, a dataset and benchmark meant to spur the development of algorithms for LLM personalization, a critical and under-explored area with significant potential for enhancing interaction quality. We discuss the potential of the empirical foundation we develop and highlight potential risks and limitations.

Meta-Learning for Personalization We hope to encourage more work in the meta-learning setting, as exemplified by our experiments. This setting mirrors many real-world use cases where an organization has a large proprietary dataset from historical users but a very limited interaction history with this particular user. Prior work on cold-start problems has focused on the task of recommending discrete content items from a media (or other) library. Extending and developing these techniques for LLMs is an exciting direction for future research.

Risks and Limitations We must consider the risks and limitations associated both with the release of our original benchmark dataset, as well as the larger goal of LLM personalization.

With respect to PersonalLLM, we note all prompts and responses have not been manually inspected for quality or safety by a human, although prompts are sourced from existing, reputable datasets, and responses are generated from state-of-the-art language models that have (presumably in the case of black box models) undergone safety alignment. Our benchmark is also limited with respect to the realism of the personas created by weighting reward models, as there exists much analysis left undone as to the preferences being displayed.

On a broader note, the goal of LLM personalization brings particular risks. One common concern is the creation of filter bubbles, where the model's outputs become increasingly tailored to the user's past existing preferences, potentially reinforcing political beliefs and biases, isolating the user from opposing viewpoints, and narrowing the diversity of information presented. Another potential issue is stereotyping, where the model may perpetuate or even amplify biases based on the user's demographic information or behavior patterns. Feedback loops may also emerge, where the model behavior affects human behavior and vice versa, leading to negative personal and unknown societal consequences. Personification risks arise, as over time the user may develop a pseudo-personal relationship with the user, potentially fostering over-reliance on the LLM for advice or companionship. Finally, if used by malicious actors, personalized LLMs can be used to manipulate and extort individuals by exploiting personal levers. Given these and many other predictable (and unpredictable) potential risks, it is important that any efforts at LLM personalization are accompanied by research in robust transparency mechanisms and safeguards for personalization algorithms. Developing an empirical foundation for such efforts is another promising avenue for future work.

Future Directions Given that LLMs have only recently reached a level of capabilities meriting their widespread adoption for industrial and personal use, the study of LLM personalization is necessarily in its earliest stages of development. It follows that there are many important and exciting avenues for future research, with respect to datasets, methodology, fairness, safety, and other aspects of responsible and reliable machine learning deployment. Since PersonalLLM is the first dataset to enable the study of complex personalized preferences expressed over many high-quality responses (to our knowledge) by a large, diverse user base, the benchmark can be extended in many ways. For example, one might imagine a distribution shift scenario, where over time, personal preferences shift, and the personalization algorithm must balance stability and plasticity. Also, we hope that our testbed drives the development of even more realistic personalization datasets and evaluation methods that more closely mirror the online and non-i.i.d. nature of the conversational setting and more closely capture the true nuance and diversity of human personal preferences. Finally, continued work in personalization algorithms must be accompanied by a proportional amount of work in personalization safety, fairness, and reliability. Future research may consider different aspects of the deployment pipeline (e.g., model architecture, data collection) and interaction model (e.g., UI/UX) with these concerns in mind.

E Details on Simulating Personal Preference Models

For an input prompt $x \in \mathcal{X}$, an LLM produces output response $y \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the set of all-natural language. Then, a preference model $\mathrm{R} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ assigns a reward score to the response given to the prompt, with higher scores indicating better responses. Next, consider a set of B base reward models, denoted as RM_b, $b = 1, \ldots, B$, and a set of k B-dimensional weightings, which represent a set of personal preference models. Then, the preference model corresponding to user i is defined by an weighted average of these B base $\mathrm{RM}_1, \mathrm{RM}_2, \ldots, \mathrm{RM}_B$, with weights $w_1, w_2, \ldots, w_B$:

$$\mathrm{R}^i(x, y) = \sum_{b=1}^{B} w_b^i \cdot \mathrm{RM}_b(x, y) \tag{1}$$

For our base reward models $\{\mathrm{RM}_b\}_{b=1}^{B}$, we select 10 reward models with strong performance on RewardBench, an open source bnechmark for evaluating reward models. These reward models are built on top of popular base models such as Llama3, Mistral, and Gemma (see Appendix G). We evaluate each (prompt, response) pair in the train and test set with each model so that for any personality created in this manner, each (prompt, response) pair in the dataset can be scored via a simple weighting.

There are many valid ways to sample the B-dimensional weighting vectors. As a simple starting point, we propose to sample preference models from a Dirichlet distribution with a uniform concentration parameter across all classes ($w \sim \mathrm{Dirichlet}(\alpha)$). As α becomes very small, the preference models converge towards the 10 base reward models; as it becomes large, preferences become unimodal. Such a parameter allows us to simulate user bases with different underlying preference structures (see Section A for more details).

F Additional Simulated User Analysis

Tables 2 and 3 include representativeness scores across all 60 demographic groups in the OpinionQA study.

G Additional Dataset Details

G.1 Dataset

We plan to open source a dataset with 10,402 rows of prompts, each with 8 diverse responses and accompanying scores from 10 reward models.

G.2 8 Models Responses

The 8 responses from each model were sampled with a temperature of 1.0, and a maximum length of 512 from OpenRouter. We chose a maximum of 512 token length because some reward models have limited context length.

G.3 Reward Models

The 10 reward models we collected are from RewardBench.

- weqweasdas/RM-Gemma-2B [9]
- sfairXC/FsfairX-LLaMA3-RM-v0.1 [9]
- OpenAssistant/reward-model-deberta-v3-large-v2
- PKU-Alignment/beaver-7b-v1.0-cost [16]

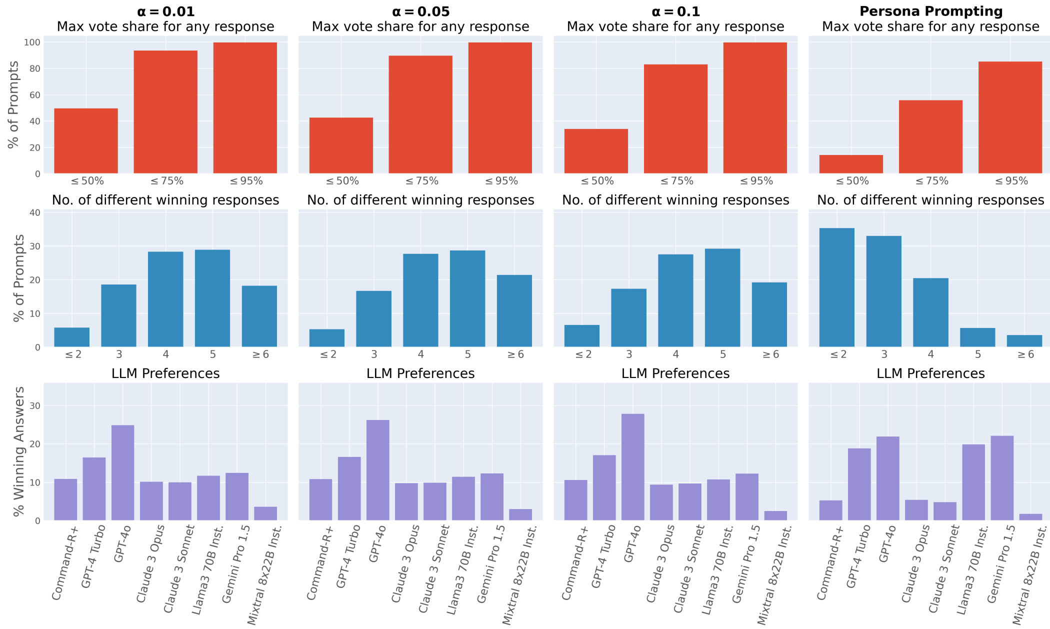

Figure 5: Probing the heterogeneous preferences across PersonalLLM prompt/responses given different settings of α, and comparing to a persona prompting baseline. **Top**: For a population of simulated users, the percentage of each population's vote share given to the most common winning response for each prompt. **Middle**: A histogram showing the number of resonses that recieve at least one vote from a simulated population for each prompt. **Bottom**: Average win rates across the population for the 8 LLMs in our dataset.

- hendrydong/Mistral-RM-for-RAFT-GSHF-v0 [9]
- OpenAssistant/oasst-rm-2-pythia-6.9b-epoch-1
- OpenAssistant/oasst-rm-2.1-pythia-1.4b-epoch-2.5
- weqweasdas/RM-Mistral-7B [9]
- Ray2333/reward-model-Mistral-7B-instruct-Unified-Feedback
- weqweasdas/RM-Gemma-7B [9]

All the reward models are obtained from Huggingface on RewardBench's leaderboard and are instantiated as per RewardBench's codebase, where reward models are submitted and edited by the contributors themselves. https://huggingface.co/spaces/allenai/reward-bench [19]

G.4 Additional Persona Analysis Details

All features are scored using pre-trained models from Huggingface.

- Formality is scored using: s-nlp/roberta-base-formality-ranker
- Educational value is scored using: HuggingFaceFW/fineweb-edu-classifier
- Emotion is scored using: SamLowe/roberta-base-go_emotions

H Additional Experiment Details

For our meta-learning approach (**Meta-Learning**), we consider a database of previous interactions between users and the language model. Specifically, for a particular user, we have M interactions, each consisting of:

1. A prompt given to the language model.

Demographic	AI21 Labs j1-jumbo	j1-grande-v2	ada	OpenAI text-davinci-003	PersonalLLM **Ours**
Northeast	0.811	0.802	0.819	0.704	0.838
Midwest	0.810	0.797	0.820	0.701	0.833
South	0.818	0.805	0.827	0.696	0.835
West	0.813	0.802	0.821	0.704	0.839
18-29	0.818	0.808	0.828	0.700	0.840
30-49	0.814	0.804	0.823	0.702	0.837
50-64	0.809	0.797	0.818	0.696	0.830
65+	0.792	0.779	0.800	0.699	0.818
Male	0.814	0.802	0.826	0.697	0.837
Female	0.810	0.800	0.816	0.702	0.833
Less than high school	0.828	0.812	0.835	0.685	0.832
High school graduate	0.816	0.799	0.826	0.691	0.832
Some college, no degree	0.814	0.804	0.823	0.701	0.836
Associate's degree	0.811	0.800	0.821	0.700	0.834
College graduate/some postgrad	0.802	0.794	0.810	0.710	0.833
Postgraduate	0.794	0.789	0.800	0.717	0.831
Yes	0.814	0.802	0.823	0.700	0.836
No	0.816	0.812	0.818	0.706	0.833
Married	0.810	0.799	0.819	0.699	0.832
Divorced	0.809	0.796	0.817	0.696	0.830
Separated	0.814	0.801	0.818	0.694	0.830
Widowed	0.800	0.785	0.807	0.694	0.819
Never been married	0.819	0.808	0.828	0.700	0.841
Protestant	0.810	0.797	0.820	0.694	0.828
Roman Catholic	0.816	0.806	0.823	0.702	0.835
Mormon	0.789	0.777	0.802	0.696	0.819
Orthodox	0.773	0.762	0.781	0.693	0.803
Jewish	0.792	0.785	0.800	0.707	0.824
Muslim	0.794	0.788	0.792	0.697	0.816
Buddhist	0.782	0.777	0.783	0.709	0.821
Hindu	0.796	0.794	0.789	0.707	0.816
Atheist	0.774	0.771	0.784	0.714	0.822
Agnostic	0.785	0.781	0.794	0.717	0.828
Other	0.794	0.790	0.801	0.703	0.824
Nothing in particular	0.815	0.802	0.824	0.700	0.839
More than once a week	0.807	0.793	0.816	0.690	0.824
Once a week	0.811	0.798	0.819	0.696	0.829
Once or twice a month	0.818	0.807	0.825	0.699	0.833
A few times a year	0.817	0.809	0.824	0.705	0.837
Seldom	0.811	0.800	0.821	0.703	0.835
Never	0.806	0.795	0.816	0.701	0.836

Table 2: Representativeness scores in relation to real human opinions from important demographic groups for different LLMs, as well as our PersonalLLM population.

2. A response generated by one of the eight different language models (treated as eight different arms in bandit literature).
3. Feedback provided by the user, representing true values from the user's reward function (rather than binary ratings).

Here, M is a random variable uniformly distributed over the integers in the interval $[25, 50)$.

Now, consider a new user u with a new prompt p. For this new user, we have limited interactions—m interactions, where m is a random variable uniformly distributed over the integers in the interval $[1, 5]$. Our goal is to use the previous user dataset and the interactions with the new user to generate a high-quality response for prompt p. We achieve this by finding the most similar and useful (prompt,

Demographic	AI21 Labs j1-jumbo	j1-grande-v2	ada	OpenAI text-davinci-003	PersonalLLM **Ours**
Republican	0.791	0.776	0.805	0.680	0.812
Democrat	0.800	0.795	0.804	0.719	0.834
Independent	0.812	0.801	0.821	0.701	0.838
Other	0.820	0.804	0.832	0.693	0.839
Less than $30,000	0.828	0.813	0.833	0.693	0.838
$30,000-$50,000	0.814	0.802	0.822	0.698	0.834
$50,000-$75,000	0.807	0.796	0.816	0.703	0.833
$75,000-$100,000	0.800	0.791	0.811	0.705	0.829
$100,000 or more	0.797	0.790	0.807	0.708	0.831
Very conservative	0.797	0.778	0.811	0.662	0.811
Conservative	0.796	0.780	0.810	0.684	0.817
Moderate	0.814	0.804	0.822	0.706	0.838
Liberal	0.792	0.788	0.799	0.721	0.833
Very liberal	0.785	0.782	0.791	0.712	0.825
White	0.807	0.794	0.817	0.699	0.832
Black	0.820	0.812	0.823	0.702	0.833
Asian	0.814	0.806	0.819	0.708	0.839
Hispanic	0.820	0.810	0.824	0.706	0.839
Other	0.801	0.783	0.807	0.681	0.818

Table 3: Representativeness scores in relation to real human opinions from important demographic groups for different LLMs, as well as our PersonalLLM population.

response, rating) tuples in the dataset and appending them, along with the new user's interactions (prompt, response, rating), to the context for the language model to generate the response.

To enable efficient search and retrieval, we concatenate each (prompt, response, rating) tuple and feed it into the OpenAI API to generate an embedding of size 256. Assuming we have N users, the embedding table has a shape of $(N, 49)$, where some entries are null because M is not always 49. We replace the null entries with zero vectors and create a mask to identify these null entries. This transforms the embedding table into a tensor of shape $(N, 49, 256)$.

For each of the m (prompt, response, rating) tuples of the new user, we compute the cosine similarity with this tensor table, apply the zero mask, and obtain a similarity score table of shape $(N, 49)$. We then extract the top k entries with the highest similarity scores.

This process ensures that we can effectively utilize historical interactions to enhance the response quality for new users, leveraging similarities in past prompts, responses, and user feedback.

H.1 Hardware

We used two nodes of 8x A100 GPUs each. The evaluation pipeline is tested to run on 1 A100 GPU with 80GB of VRAM.

I Example Dataset

I.1 Sample Evaluation Preference Dataset

person_weight : [0.99999855, 2.16500320e-29,..., 1.0112404759e-90]
prompt_1 : What is the best way to search for a job?
response_1_a : There are several effective ways to search for a job...
response_1_b : There's no single "best" way to find a job, as the most effective approach depends ...
chosen_1 : b
⋮
prompt_5 : The fifth prompt given to the person.
response_5_a : The first response option for prompt 5.
response_5_b : The second response option for prompt 5.
chosen_5 : The chosen response for prompt 5.
user_history_length : 5
test_prompt : What card games can suggest playing with my kids? They are 8 and 10.
best_response : Here are some card games suitable for your children's ages (8 and 10): 1. Uno...
best_response_model : 1. **Go Fish**: - **Objective**: Collect pairs of cards. - ...
best_response_reward : 2.3231
gpt4o_response : The response generated by GPT-4
gpt4o_reward : -0.1232
person_id : 1

I.2 Sample Evaluation Reward Dataset

person_weight : [0.99999855, 2.16500320e-29,..., 1.0112404759e-90]
prompt_1 : What is the best way to search for a job?
response_1 : There are several effective ways to search for a job...
reward_1 : -0.1232
⋮
prompt_4 : The fifth prompt given to the person.
response_4 : The first response option for prompt 5.
reward_4 : The reward for prompt, response 5.
user_history_length : 4
test_prompt : What card games can suggest playing with my kids? They are 8 and 10.
best_response : Here are some card games suitable for your children's ages (8 and 10): 1. Uno...
best_response_model : 1. **Go Fish**: - **Objective**: Collect pairs of cards. - ...
best_response_reward : 2.3231
gpt4o_response : The response generated by GPT-4
gpt4o_reward : -0.1232
person_id : 1

J Baselines Implementation

J.1 Result Analysis

Our baseline methods are demonstrably simple, aiming to showcase the utility and realism of our dataset, as well as its capacity to generate rewards for testing personalization algorithms. We have explored two families of such algorithms.

We know that the output response is influenced by both the prompt and the method used to select previous interactions as context samples. An example is how ChatGPT utilizes Memory, which are summarized versions of conversations that are remembered and passed in as context in future conversations. Our baseline results are not groundbreaking due to the random selection of previous interactions. We encourage future methodological research to improve upon our Best-of-8 baseline, ideally using a small model.

J.2 Non Meta Learning

For non meta learning, we limit ourselves to using context from the same row. E.G., for one shot, we draw one past conversation from the previous interaction and pass that as context to the prompt.

Example for three shots.

```
prompt = "Below are some examples of the user’s past conversation
history with a chosen response per prompt."
history = []
shots = 3
for I in range(shots):
    past_prompt = row["prompt_" + str(I + 1)]
    chosen_response = row["chosen_" + str(I + 1)]
    history.append(
        "User: "
        + past_prompt
        + "\nAssistant: "
        + chosen_response
        + "\n\n"
    )
# Check if the total length of the history exceeds the maximum token limit
while len(''.join(history)) > 6000:
    # If it does, remove the earliest history
    history.pop(0)
prompt += ''.join(history)
prompt += "Use the contexts above to generate a good response for
the user prompt below."
```

J.3 Meta Learning

Below is an example of Embedding search meta-learning.

```
# Initialize the Full Prompt with instructions and a heading for current user's histories
full_prompt = "Below are some examples of the user's past conversation history"
full_prompt += "###Current User Histories###\n\n"

# Loop through each user interaction
for each interaction in user_history:
    full_prompt += '---Current User Interaction---\n\n'
    full_prompt += 'User:\n' + past_prompt + '\n\n'
    full_prompt += 'Assistant:\n' + past_response + '\n\n\n'

# Extract similar pairs from the training data
similar_pairs = extract_similar_pairs(training_data, current_interaction)

# Randomly sample the similar pairs
sampled_pairs = random_sample(similar_pairs, required_samples)

# Append similar users' interaction histories
full_prompt += "###Most Similar Users' Histories From Database###\n\n"
for each pair in sampled_pairs:
    full_prompt += '---Similar User Interaction---\n\n'
    full_prompt += 'User:\n' + similar_prompt + '\nAssistant:\n' + similar_response + '\n\n'

# Finalize the prompt with instructions for generating a response
full_prompt += "Use the above histories to generate a response for the following prompt"
full_prompt += 'User:\n' + test_prompt + '\n\nYour Response:'

# Return the full prompt
return full_prompt
```