Final Year Project Interim Report

Personalization of Large Language Models for Diverse User Preferences

Student: LIU Meitong

Supervisor: Prof. LUO Ping

Department: Computer Science

Date of Submission: December 1, 2024

Abstract

Current post-training schemes of Large Language Models fail to model the heterogeneous and conflicting human preferences. This final year project explores effective and efficient methods for the personalization of Large Language Models (LLMs) to accommodate the diverse preferences of users from varying backgrounds. The project is divided into a two-stage framework: reward modeling and policy model training. In particular, the reward modeling phase aims to predict individual preferences by leveraging demographic and psychological attributes, enabling the creation of more adaptive and inclusive models. This is followed by policy model training, which employs advanced techniques such as Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO). These methods are benchmarked to assess their efficacy in generating personalized responses. Preliminary investigations have validated the hypothesis that user preferences are diverse and influenced by factors such as occupation. Further, experiments on training-free methods offer some insights on choosing user-specific identifiers. This project aspires to advance LLM personalization, ensuring equitable access to high-quality AI tools for all users while contributing to broader research on human-centric AI alignment.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
DPO	Direct Preference Optimization
LLM	Large Language Model
ML	Machine Learning
PPO	Proximal Policy Optimization
RLHF	Reinforcement Learning from Human Feedback
SFT	Supervised Finetuning

1 Introduction

1.1 Background: The Training Paradigm of Large Language Models

Recent years have witnessed a drastic development of Large Language Models (LLMs) [4]. Demonstrating remarkable capabilities in natural language processing, their deployment in numerous applications, from document refinement [15] to web agents [17], has made them not only integral to industries but also a powerful daily life tool for people from all walks of life. However, given the diverse preferences held by individuals with highly heterogeneous backgrounds, it remains an open question whether trending LLMs suffice to cater to such wide opinions [11]. To tackle this problem, it is essential to revisit the current training paradigm of LLMs.

The workflow for training a Large Language Model typically involves two stages [1]: pre-training and post-training. During the pre-training phase, the model is trained on internet-level corpora to learn linguistic patterns and perceive a wide range of concepts. Following that, the post-training stage refines the model's abilities in instruction-following and aligns its performance with human morals and preferences.

Based on the effective framework proposed by OpenAI [8], the post-training stage also contains two steps: supervised finetuning (SFT) and preference alignment. Supervised fine-tuning (SFT) involves refining the pre-trained model on curated datasets with expert-provided demonstrations, where the model learns to generate more accurate responses by mimicking high-quality human output. Preference alignment focuses on aligning the model's behavior with user preferences and ethical guidelines. This is typically done through Reinforcement Learning from Human Feedback (RLHF), where model responses favored by human annotators are encouraged, while those disliked are penalized.

While extensive studies have been conducted to improve the effectiveness of the preference alignment stage, there exist inherent flaws that hinder an LLM's ability to personalize to individual tastes. The next section analyzes such drawbacks, serving as the motivation for this project.

1.2 Motivation: Current LLMs Failing to Capture Diverse User Preferences

The motivation for this project stems from the recognition that human preferences are heterogeneous and often conflicting. Users with different backgrounds, for instance, race, gender, and personality, may favor different LLM responses. Current preference alignment methods, despite significant progress, fail to adequately capture this diversity for two primary reasons.

First, common datasets used for preference alignment rely on anonymous annotators who rank multiple responses without accounting for their demographic or contextual information. This leads to a homogeneous representation of opinions, underrepresenting voices from those in the minority [7]. Consequently, models trained on such data struggle to reflect the broad spectrum

of user preferences. Second, current training methods essentially treat each comparison sample equally. When conflicting preferences arise, imposing equal and opposite forces that navigate how the model adapts, the model misinterprets them as data noise and eventually fails to learn any preference [3].

These two issues - insufficient profiling of annotators and ineffective algorithms to distinguish and preserve conflicting preferences - highlight a gap in current methodologies for LLM personalization. Recent advancements in dataset curation have alleviated the first limitation by incorporating detailed annotator information, allowing for a richer and more diverse representation of user preferences [11, 7, 2]. However, there is yet no proper solution to the latter drawback. Given such progress and remaining concerns, this project aims to explore feasible frameworks that embed diverse user preferences in one model and enable it to personalize adaptively.

1.3 Project Objectives

This project aims to address the aforementioned challenge of personalizing Large Language Models with a two-stage process - reward modeling and policy model training. We break this down as specific objectives.

In the first stage, we target to obtain a reward model that can predict the preferences of different individuals provided with their demographic and psychological attributes. This model could help understand how LLMs adjust their output given user information and shall be an essential element guiding the latter policy training stage. The deliverables of this stage will be a pipeline report on preference reward modeling and the codebase implementing the best practices.

The second stage aims to obtain a desired policy model that generates tailored content for different users. We adopt and compare two algorithms that have been consistently proven effective in policy training, namely Reinforcement Learning from Human Feedback (RLHF) and its alternative, Direct Preference Optimization (DPO) [9]. The deliverables of this stage will incorporate a benchmark on the two methods supported by extensive experimental evidence, the codebase for policy model training, and the complete model parameters that can be deployed and evaluated.

Integrating all outcomes, the project will yield a comprehensive report on LLM personalization, providing theoretical and empirical insights into the related research fields. The holistic contributions can be summarized as follows:

- Examined how preferences among diverse individuals differ and are formed by factors such as occupation.
- Empirically benchmarked various post-training schemes for LLM personalization and identified effective directions for future work.
- Called out to the research community the importance and possibility of equal and high-quality access of LLM to every social member.

1.4 Report Outline

The rest of this report is organized into three sections. Section 2 details the methodologies employed. It begins with a review of existing works on related topics, followed by the specific methods considered, including the reward modeling techniques and the policy training algorithms. Evaluation approaches for both the reward and policy model is also mentioned.

Section 3 will present the main experimental results. It will include an evaluation of the reward model's ability to predict diverse user preferences and a performance analysis of the policy model in generating personalized responses. We will compare different training methodologies and personalization techniques, highlighting the strengths and weaknesses of each approach.

Finally, Section 4 will summarize the key findings and contributions of the project. It will reflect on the effectiveness of the proposed frameworks in embedding and personalizing diverse user preferences in LLMs, as well as identify areas for future improvement and research.

2 Methodology

2.1 Overview

This section presents the methodologies adopted in the project. Section 2.2 revisits the related endeavors on dataset curation and algorithms balancing groups' opinions. Sections 2.3 and 2.4 specify the different methods considered in our study. In particular, Section 2.3 introduces reward modeling techniques, including roleplay prompting, few-shot promptings, and conditional finetuning. Section 2.4 discusses the policy training algorithms, namely Reinforcement Learning from Human Feedback (RLHF) and its alternative Direct Preference Optimization (DPO).

2.2 Related Work

Various attempts have been made to address the first limitation mentioned in Section 1.2, which calls for curated preference datasets profiling annotator information. The OpinionsQA dataset [11] collects US citizens' opinions on controversial political issues, with each annotator's demographic information recorded in 8 attributes. Built on this, the GlobalOpinionsQA dataset [7] further incorporates more countries and more questions. Although these two datasets include annotator information, they only include questions from limited topics, mainly political, and thus cannot serve as a good general preference learning dataset for LLMs. A recently established PERSONA dataset [2] provides a better alternative. With reference to real survey data and assistance from language models, it creates 1586 synthetic personas with 33 attributes each. The personas are then roleplayed by GPT-4 to generate diverse preferences for questions from a wide range of topics. We intend to mainly use this PERSONA dataset as our training source and testbed.

Due to the lack of proper preference datasets, there exists little work that targets preserving diverse user tastes in one model. Nevertheless, some algorithms have tried to solve an easier problem where one finds a compromising solution among the preferences of different groups. Chakraborty et al. [3] adopt the typical RLHF workflow, training a separate reward model for each group and optimizing the policy model for the worst-performing one in each training round. Ramesh et al. [10] uses the alternative DPO algorithm equipped with mirror descent that also prioritizes undertrained groups. Although such methods do not target the same goal as this project, their training techniques remain valuable for reference and, indeed, inspire most of the algorithms we employ.

2.3 Reward Modelling Techniques

The reward modeling stage aims to obtain a model that predicts a user's preference given some information that implicitly reflects or represents his/her underlying logic. The aforementioned PERSONA dataset containing individual attributes and the corresponding preference samples is adopted for both training and evaluation. We investigate three lines of techniques, where the first two are training-free methods eliciting model capabilities purely through prompting, while the third involves parameter finetuning, which is what we are mainly interested in. The detailed descriptions are as follows:

Roleplay prompting. The user information is given to the LLM as the system prompt to facilitate its preference prediction. For example, given a data point $(P, Q, R1, R2)$ randomly sampled from the PERSONA dataset, where P is a persona profile, Q is a query, and $R1$ and $R2$ are two model responses to be compared, the model is then prompted by: "system prompt: Imagine you are P; user prompt: Given the question Q, which of the following responses do you prefer? $R1$ or $R2$". The performance of this method is directly evaluated by its prediction accuracy on all or a subset of PERSONA samples.

few-shot prompting. Apart from the user information P, the model is additionally provided with a small number of $(P, Q, R1, R2, A)$ examples, where A is the ground-truth label. Different from roleplay prompting, this method enables the model to infer the preference of a given client through limited demonstrations. Its performance is evaluated by the prediction accuracy on unseen $(P, Q, R1, R2)$ queries from the same persona P.

Conditional reward modeling. Given the user information P in the prompt, an SFT model is fine-tuned to match the ground truth label when queried with $(P, Q, R1, R2)$. Here, two types of loss functions are applied and compared, namely the Bradley-Terry style and the pairwise preference style [6]. The performance is evaluated on a leave-out test set. This method is "conditional" in the sense that instead of fitting the marginal distribution $\pi(\cdot \mid Q, R1, R2)$ which mixes preferences of all users together, it fits the *conditional* distribution $\pi(\cdot \mid Q, R1, R2, P)$ given specific client attribute. Such conditioning enables the model to preserve conflicting opinions from diverse groups, providing the key ingredient for LLM personalization.

2.4 Algorithms for Policy Model Training

The policy training stage aims to obtain a single large language model that can generate personalized responses given a user's preference profiles at deployment. Inspired by the aforementioned existing literature on balanced policy training, two popular pipelines are investigated, namely Reinforcement Learning from Human Feedback (RLHF) and its alternative Direct Preference Optimization (DPO). The following paragraphs elaborates on these two approaches.

RLHF. RLHF is a widely adopted method for finetuning large language models for preference alignment. Figure 1 [6] provides an overview of its workflow.

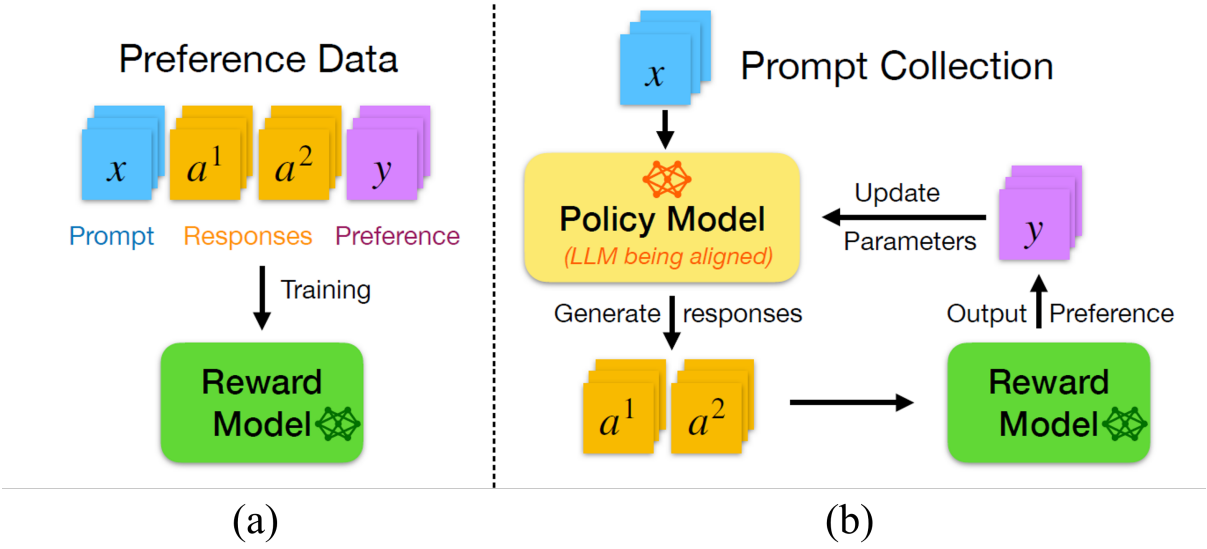


Figure 1: An illustration of the workflow of RLHF.

As demonstrated in Figure 1(a), RLHF begins with the creation of a reward model capable of predicting human preferences, which is done in the first stage of the project. Compared to existing ones, the novelty of our reward model arises from the training dataset of user-specific preferences, enabling it to distinguish diverse and conflicting opinions from different clients.

Then, as shown in Figure 1(b), the established reward model is incorporated into a Reinforcement Learning (RL) cycle, guided by typical RL algorithms such as Proximal Policy Optimization (PPO) [12]. During each iteration, the policy model generates several answers to a prompt. The reward model then ranks the responses favoring those more desired, with such comparisons turned into a real-valued score by the adopted RL algorithm such as PPO. The policy model is then iteratively updated to maximize the reward score. This process allows the model to learn and adapt, refining its behavior to meet user expectations. In our case, to preserve diverse opinions, the policy model, as well as the reward model, are prompted not only with a query but also the specific client attributes.

DPO. DPO is a commonly used alternative for RLHF in academia for its simplicity and stability [14]. It converts Reinforcement Learning into a supervised learning problem by implicitly incorporating the reward score into the loss function through mathematical derivation. Specifi-

cally, DPO uses the following loss function:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(Q,R_w,R_l)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(R_w \mid Q)}{\pi_{\text{ref}}(R_w \mid Q)} - \beta \log \frac{\pi_\theta(R_l \mid Q)}{\pi_{\text{ref}}(R_l \mid Q)} \right) \right] \quad (1)$$

where π_θ is the policy model, π_{ref} is a referencing model typically set as the SFT model, Q is the querying prompt, R_w and R_l are the preferred and un-preferred responses selected by the reward model, σ is the sigmoid function, and β is a tunable hyperparameter. By its formulation, this objective rewards the "won" response and penalizes the "lost", while regularizing the parameter of the policy model to its reference to prevent overfitting. Again, in our case, we insert the user profile P into the prompt, resulting in the following adjusted DPO loss:

$$\mathcal{L}'_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$
$$-\mathbb{E}_{(P,Q,R_w,R_l)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(R_w \mid Q, P)}{\pi_{\text{ref}}(R_w \mid Q, P)} - \beta \log \frac{\pi_\theta(R_l \mid Q, P)}{\pi_{\text{ref}}(R_l \mid Q, P)} \right) \right] \quad (2)$$

Although slightly surpassed by RLHF as reported in some settings [16], DPO demonstrates more training stability and takes up less GPU memory by lifting the necessity to load multiple models. Hence, we experiment with this algorithm as well.

Finally, the evaluation of the policy model adopts several popular methods. We first examine its scoring on the trained reward model. Then, the win rates of its responses over those other state-of-the-art general LLMs are also considered, where the comparisons are made by the latest GPT agent or real human annotators if resources allow.

2.5 Summary

This section outlines the methodologies adopted in the project, beginning with an exploration of prior efforts in dataset curation and algorithmic approaches for balancing diverse group preferences. The following sections delve into the specific techniques investigated in this study by highlighting various reward modeling methods. The policy training algorithms are then addressed, focusing on RLHF and DPO, both of which are assessed for their potential to enhance the personalization of large language models for diverse user preferences.

3 Current Progress

3.1 Overview

This section reports the current project status. Section 3.2 demonstrates the preliminary results that validate the overall assumption of this project: different individuals possess diverse preferences. This section also offers insights into how such heterogeneous tastes are formed, focusing on user occupations. Section 3.3 states some limitations of current methods and difficulties encountered, followed by tentative solutions for mitigation.

3.2 Preliminary Investigation

The fundamental assumption necessitating LLM personalization is that users with different backgrounds have diverse and conflicting preferences that cannot be addressed universally. Our preliminary investigation provides affirmative verification of this assumption.

3.2.1 Different Users Prefer Different Responses

To simulate clients of diverse backgrounds, we extracted 36 profiles from the PERSONA dataset with distinct social positions, ranging from elementary students to pediatricians, together with 1000 randomly chosen questions. The open-sourced large language model Llama3 8B by Meta is then roleplay prompted to select from two given model responses the one more likely preferred by the character. The results are summarized in Figure 2.

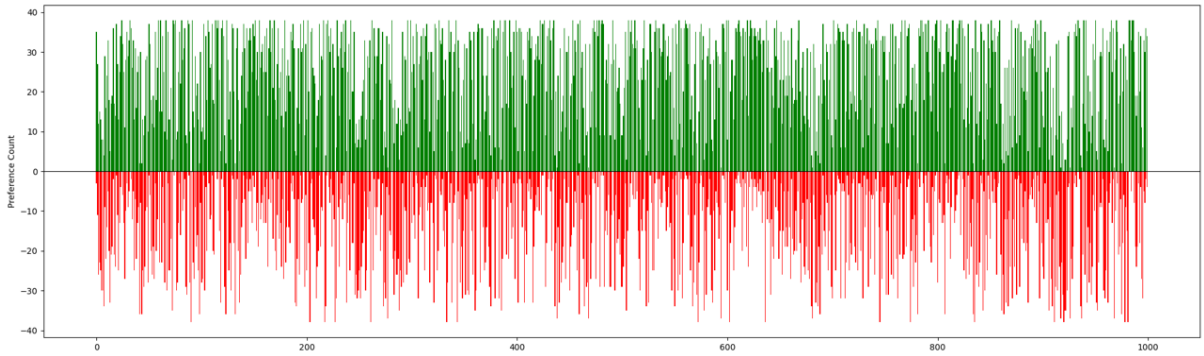


Figure 2: Number of characters preferring different responses across 1000 questions.

The green bars represent the number of characters choosing the first response, while the red bars represent that for the second response. From the comparable total counts for the two options, it is evident that human tastes are heterogeneous among individuals.

3.2.2 Preferences Are Related to User Occupations

Another interesting question to investigate is how such diverse preferences relate to social identifiers. In this subsection, the correlation between preferences and user occupations is explored.

For each user, its preferences for all questions are integrated into a one-dimensional vector. By calculating pair-wise cosine similarities of such vectors, a similarity matrix is obtained, as illustrated in Figure 3. An entry with a darker color represents higher similarity. One can observe that the rows and columns for “elementary student” are consistently light, indicating low similarity with all other professional occupations. This highly aligns with our intuition that young children process information and make decisions significantly different from adults.

Furthermore, several graphs are constructed by connecting each user with his or her mutual top five similar neighbors. In other words, two users are connected if and only if they both ranked top five in each other’s similarity list. These connected graphs, as shown in Figure 4, reveal

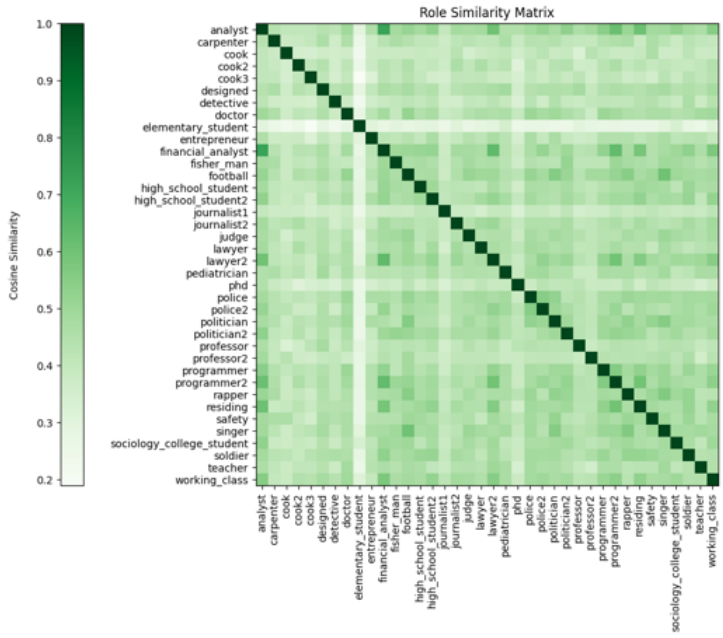


Figure 3: Similarity matrix of user preferences.

some expected structures. For instance, singers and rappers, safety positions and politicians, two policemen, and two programmers share highly similar tastes.

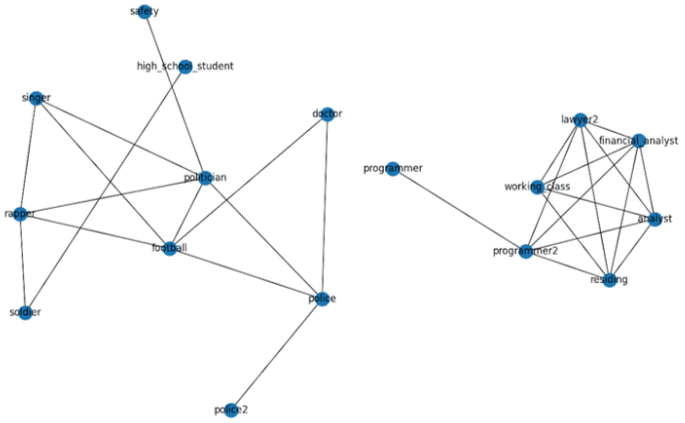


Figure 4: Connected graph formed by mutual top five similar roles.

3.3 Experiments on Few-shot Prompting

This section presents some results using few-shot prompting, one training-free method mentioned in Section 2.3.

3.3.1 Experenmental Setup

Dataset and preprocessing The initially planned PERSONA dataset is still not accessible to the public. Instead, we adopt a commonly adopted RLHF dataset, UltraFeedback [5]. Each

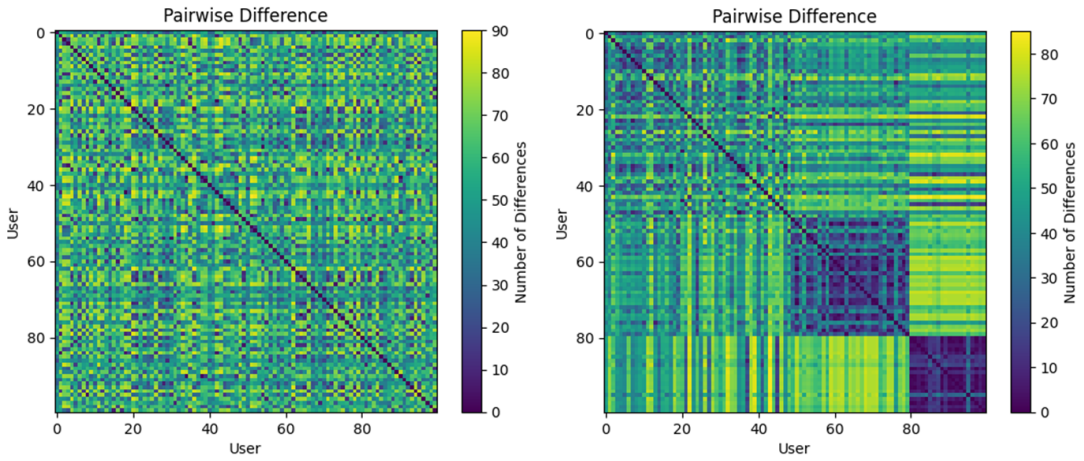
entry of UltraFeedback consists of a query, $2 \sim 4$ responses generated by language models, and their corresponding multi-objective scores rated on four dimensions, namely instruction-following, honesty, truthfulness, and helpfulness. When training a reward model, these four scores are normally aggregated as the overall score of a response, and the accepted and rejected pair can be then determined.

We follow the suggested procedure in the original work to filter out potentially contaminated entries. Additionally, to encourage disagreements among users, we only consider queries with 4 responses whose scores are not dominated by each other. Eventually, 300 questions are batched into the training set, and 95 questions constitute the test set.

Simulating diverse preferences The multi-objective ratings enable us to simulate diverse user preferences. We model each user as a preference vector $w \in \Delta_4$, where w_i represents the importance this user attaches to a certain dimension. As such, the user-specific rating for a query is the weighted sum, instead of the uniform sum, of the sub-scores. When comparing a pair of responses, different users may choose oppositely. Specifically, we generate 100 preference vectors using two strategies:

- Random: sample 99 variables from $\mathrm{Dirichlet}([1, 1, 1, 1])$, plus 1 uniform baseline
- Cluster: sample 49, 30, and 20 variables from $\mathrm{Dirichlet}([2.5, 2.5, 2.5, 2.5])$, $\mathrm{Dirichlet}([0.7, 0.3, 6.0, 3.0])$, and $\mathrm{Dirichlet}([7.5, 0.5, 1.2, 0.8])$ respectively, plus 1 uniform baseline

Here, the uniform baseline refers to $w = [0.25, 0.25, 0.25, 0.25]$, representing the uniform aggregation used in current post-training schemes. Figure 5 visualizes the user opinions, where each entry indicates the number of differences between two users in their choice of most preferred response for a certain query among the 95 test queries.



(a) Random (b) Cluster

Figure 5: Visualization of user opinions under two settings.

Question selection by information gain To facilitate the few-shot prompting method, several (query, user-specific accepted response, user-specific rejected response) should be provided in the prompt to demonstrate user preferences. To find the most informative set, we adopt a greedy selection strategy based on mutual information. In particular, the question with the highest information gain is selected in each iteration. In other words, we choose question j that maximizes

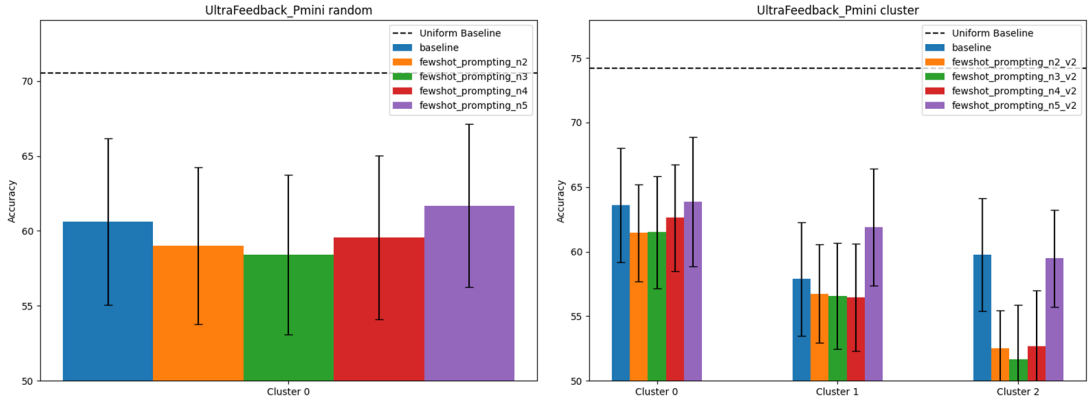
$$\begin{aligned} I(U; Q_j \mid Q_S) &= I(U; Q_{S\cup\{j\}}) \\ &= \sum_{u\in U}\sum_{a\in A} p(u,a)\log\frac{p(u,a)}{p(u)p(a)} \\ &= \sum_{u\in U} -\log\frac{\text{number of users choosing } a_u}{100} \end{aligned} \tag{3}$$

where U is the user set, A is the question set, a_u is the response user u prefer the most for query a, and S is the set of questions already chosen.

Models and machines We adopt the open-sourced Gemma2-2b model [13]. All experiments are conducted on one machine with one A100 GPU.

3.3.2 Results and Insights

In the following experiments, the base model is directly prompted with several preference demonstrations from a particular user, a query, and two responses. The model then answers which response is more likely to be preferred by the current user. Figure 6 illustrates the results using different numbers of few-shot demonstrations, where the metric is the prediction accuracy on the test set.



(a) Random (b) Cluster

Figure 6: Reward model prediction accuracy under two settings.

We first note that all results are much below the dotted line representing the uniform preference baseline, which again verifies that current post-training schemes may not be able to cater to di-

verse clients. Additionally, when the number of few-shot demonstrations is small, performance is actually hindered. In the clustered setting, the improvement among groups is also different, indicating that the selected demonstration set may still be somehow biased. However, it is infeasible to continually increase the number of examples given a limited context length. Instead, one may consider other embedding methods to encode this kind of information. This will also be critical when dealing with changing user preferences.

4 Conclusion and Future Plans

Human preferences are highly heterogeneous and conflicting and are not well modeled by current post-training schemes. This project aims to examine effective and efficient algorithms that enable a single LLM to adaptively output personalized responses to different users, aiming for equal access to resources for all social members. Current progress includes a literature review and preliminary investigation. Next, we will start examining methods involving fine-tuning and identifying more interesting subtopics related to personalization. The detailed tentative schedule of the project is as follows.

Table 1: Project schedule

Phase	Objectives	Progress
September, 2024	Literature review; First deliverables: detailed project plan and project webpage	Done Done
October to November, 2024	Preliminary investigation	Done
December, 2024	Stage 1: Reward model training 1. Set up environment and codebase 2. Experiment with different methods 3. Experiment with different sets of attributes as user information	Done
January, 2024	Stage 1: Reward model training 1. Result analysis and method refinement 2. Second deliverable: interim report	In progress
February, 2024	Stage 2: Policy model training Experiment with the two preference alignment approaches	Pending
March, 2024	Stage 2: Policy model training 1. Experiment with other potential methods 2. Result analysis 3. Materials integration	Pending
April to May, 2024	Final Stage 1. Draft final report 2. Prepare poster and presentation 3. Project exhibition	Pending

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] L. Castricato, N. Lile, R. Rafailov, J.-P. Fränken, and C. Finn. Persona: A reproducible testbed for pluralistic alignment. *arXiv preprint arXiv:2407.17387*, 2024.

[3] S. Chakraborty, J. Qiu, H. Yuan, A. Koppel, F. Huang, D. Manocha, A. S. Bedi, and M. Wang. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

[4] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[5] G. Cui, L. Yuan, N. Ding, G. Yao, W. Zhu, Y. Ni, G. Xie, Z. Liu, and M. Sun. Ultrafeedback: Boosting language models with high-quality feedback. 2023.

[6] H. Dong, W. Xiong, B. Pang, H. Wang, H. Zhao, Y. Zhou, N. Jiang, D. Sahoo, C. Xiong, and T. Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

[7] E. Durmus, K. Nguyen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

[8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[9] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] S. S. Ramesh, Y. Hu, I. Chaimalas, V. Mehta, P. G. Sessa, H. B. Ammar, and I. Bogunovic. Group robust preference optimization in reward-free rlhf. *arXiv preprint arXiv:2405.20304*, 2024.

[11] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[13] G. Team. Gemma. 2024. doi: 10.34740/KAGGLE/M/3301. URL https://www.kaggle.com/m/3301.

[14] H. Wang, Y. Lin, W. Xiong, R. Yang, S. Diao, S. Qiu, H. Zhao, and T. Zhang. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*, 2024.

[15] Y. Wang. Large language models evaluate machine translation via polishing. In *Proceedings of the 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence*, pages 158–163, 2023.

[16] S. Xu, W. Fu, J. Gao, W. Ye, W. Liu, Z. Mei, G. Wang, C. Yu, and Y. Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024.

[17] S. Yao, H. Chen, J. Yang, and K. Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.