Project Plan: Personalization of Large Language Models for Diverse User Preferences

Student: LIU Meitong (FYP24080)

Supervisor: Prof. LUO Ping

2024-45 Computer Science Final Year Project
The University of Hong Kong

1 Project Background

Large Language Models (LLMs) have emerged as a powerful daily life tool for people from all walks of life. For an LLM to demonstrate better instruction-following abilities and align with human morals and preferences, a vital "post-training" stage is required. Based on the wellknown framework proposed by OpenAI (Ouyang et al., 2022), this stage typically contains two steps: supervised finetuning (SFT) and preference alignment. In SFT, a pre-trained model is finetuned to mimic human experts' demonstrations when answering a set of questions. In preference alignment, the model is further optimized to generate outputs more to human taste.

While extensive studies have been done to improve the effectiveness of the latter preference alignment step, there still exist inherent flaws in both data collection practices and training paradigms that hinder an LLM's ability to **personalize**. First, to collect human preferences, a group of anonymous annotators is asked to rank multiple LLM-generated responses to a user query based on their preferences. *However, human preferences are usually inhomogeneous.* Users with diverse backgrounds - for instance, races, genders, and occupations - may favor different sets of responses. Under such anonymous ratings with missing annotator information, the minority population's voices are underrepresented (Santurkar et al., 2023). Second, From the perspective of training paradigms, each pairwise comparison sample is treated equally. This becomes a problem when samples *conflict with each other*, imposing equal and opposite forces navigating how the model adapts. In this case, the model misunderstands such conflicting samples as data noise and eventually fails to learn any preference (Wang et al., 2024).

Considering the two issues we have - **lack of annotator information for inhomogeneous preference profiling**, and **lack of proper training paradigm to distinguish and preserve conflicting preferences in the final policy model**, there is yet no proper solution to the latter due to the former causing a lack of training data. For the former, we mention some existing attempts here. The OpinionsQA dataset (Santurkar et al., 2023) collects US citizens' opinions on controversial political issues, with each annotator's demographic information recorded in 8 attributes. The GlobalOpinionsQA dataset (Durmus et al., 2023) 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 provides a better alternative (Castricato et al., 2024). 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.

In this project, we focus on solving the second issue. We aim to **explore possible frameworks that embed diverse user preferences in one model and enable it to personalize adaptively**. We approach this by a two-stage process. In the first stage, in regard to the thumb rule that evaluation is usually easier than generation, we will first try to obtain a reward model that can predict the preferences of diverse users, in other words, *evaluate* different responses from a certain user's perspective. In the second stage, we will further study how to obtain a desired policy model that *generates* tailored content for different users. We will try two mainstream approaches for policy training - the typical Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), where the reward model obtained in the first stage guides Proximal Policy Optimization (PPO) (Schulman et al., 2017), and Direct Preference Optimization (DPO) that converts reinforcement learning to supervised learning (Rafailov et al., 2024). We will evaluate the learned policy model on the learned reward model, the latest GPT family model, and human annotators if possible. Alongside practical experiments, we will also try to establish mathematical models that frame this learning process and propose submodule algorithms for certain steps if necessary.

2 Project Objectives

In this project, we target three objectives:

1. Stage 1: **Reward model** training for personalized preference prediction.

   In particular, we break this down into two smaller questions:

   - What user information works best? For example, different sets of demographic and psychological attributes.
   - What kind of training practice works best? For example, roleplay prompting, few-shot dialog, or directly training a conditional reward model given user information.

2. Stage 2: **Policy model** training for personalize response generation

   We mainly explore two high-level workflows for training a policy model, utilizing the reward model obtained in the first stage - RLHF (reward model + PPO) and DPO (probably iterative with more data generated by the reward model). Other simple methods without parameter tuning, such as roleplay prompting and few-shot dialog will also be considered as baselines.

3. Mathematical modeling and algorithmic refinement

   We examine different frameworks for analyzing the preference learning process, such as conditional distribution fitting and multi-objective optimization. We also investigate whether certain steps in the overall training process can be improved by tailored algorithmic design.

3 Methodology

In this section, we briefly explain the training methodologies we intend to try and compare for each objective listed above.

3.1 Stage 1: Reward model training

Using the PERSONA dataset, we aim to obtain a model that predicts a user's preference given some information that implicitly reflects or represents his/her underlying logic. Specific training schemes include but are not limited to:

- **Roleplay promoting**: The user information is given to the LLM as a system prompt. For example, we have "system prompt: Imagine you are P; user prompt: Given the question Q, which of the following responses do you prefer? (R_1, R_2)", where P is the user information and (P, Q, R_1, R_2) are sampled from the PERSONA data. This method does not involve parameter tuning - its performance is directly the accuracy of the model's output on such example prompts.
- **Few-shot dialog**: The model is still given the user information P in the prompt and will be additionally provided with a small number of (P, Q, R_1, R_2, A) data, where A is the ground-truth label. This method also does not involve parameter tuning. Different from roleplay prompting, the model is now given a few examples to learn from. Its performance is then evaluated by the prediction accuracy on unseen (P, Q, R_1, R_2) queries from the same persona profile P.
- **Conditional reward model training**: Given the user information P in the prompt, an SFT model is finetuned to match the ground truth label when queried with (P, Q, R_1, R_2). Here, two different loss functions can be applied - the Bradley-Terry style and the pairwise preference style (Dong et al., 2024). The performance is then evaluated on a leave-out test set. This method is called the "conditional" reward model because instead of fitting the marginal preference distribution $\pi(\cdot \mid Q, R_1, R_2)$ mixing all users up, we are fitting the conditional distribution $\pi(\cdot \mid Q, R_1, R_2, P)$ given specific user information, which enables the model to preserve opposite opinions from diverse groups.

3.2 Stage 2: Policy model training

We aim to obtain a single LLM that can generate personalized responses given a user's preference profiles. In particular, we plan to try and compare two approaches:

- **Reward model-guided PPO**: This is a classic workflow of RLHF. The trained evaluator in stage 1 serves exactly as the reward model, while the policy model is trained in a typical reinforcement learning style.
- **DPO**: DPO converts RLHF into a supervised learning problem by implicitly incorporating the reward model into the loss function. It demonstrates more stability and takes up less GPU memory for not needing to load multiple models, but sometimes cannot achieve as good performance as the RL counterpart.

Finally, the trained policy model is evaluated on the trained reward model, the latest GPT model, and human annotators if available.

4 Project Schedule

Phase	Objectives	Progress
September, 2024	Literature review; First deliverables: detailed project plan and project webpage	Ongoing Down
October, 2024	Stage 1: Reward model training 1. PERSONA dataset preprocessing 2. Set up environment and codebase 3. Establish a reward model training pipeline	Pending
November to December, 2024	Stage 1: Reward model training 1. Experiment with different methods 2. Experiment with different sets of attributes as user information	Pending
January, 2024	Stage 1: Reward model training 1. Result analysis and method refinement 2. Second deliverable: interim report	Pending
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

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.

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.

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.

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.

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.

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.

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

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.