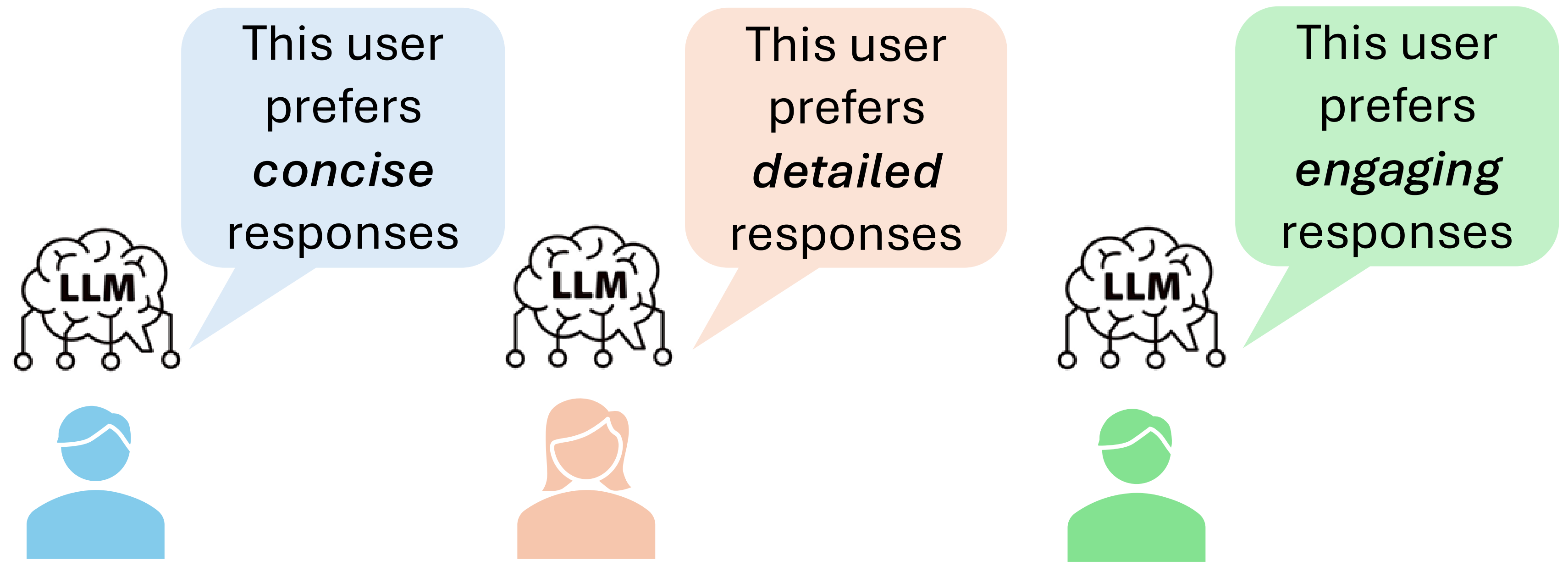
Personalized Adaptation via In-Context Preference Learning

Allison Lau[1] Younwoo (Ethan) Choi*[1] Vahid Balazadeh*[1] Keertana Chidambaram*[2]

Vasilis Syrgkanis[2] Rahul G. Krishnan[1]

[1]*University of Toronto* [2]*Stanford University*

Motivation



This user prefers *concise* responses

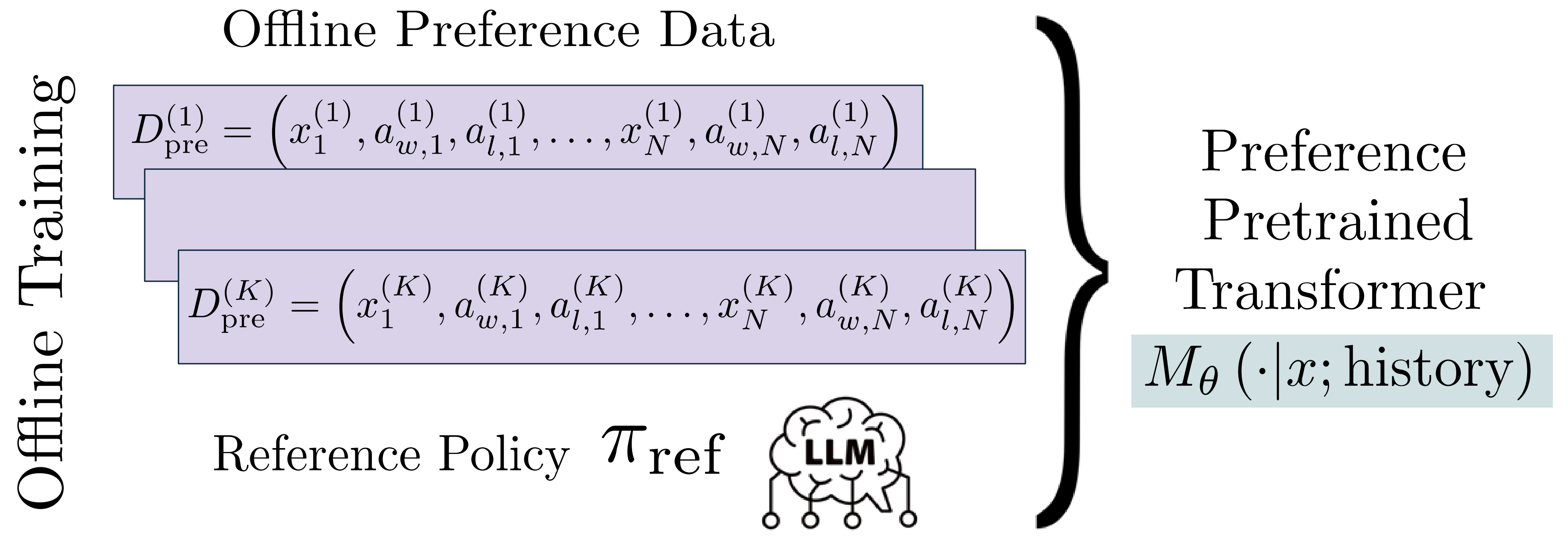
This user prefers *detailed* responses

This user prefers *engaging* responses

- To account for the diverse preferences of different populations in Reinforcement Learning from Human Feedback (RLHF)
- Limitations of existing methods (multi-objective RL and "Personalized Soups" (PS) [1])
  - Require training and maintaining multiple models which take up memory resources
  - Entirely offline which limit personalization at the user-level

Preference Pretrained Transformer (PPT)

Offline Training: Train an in-context learning model



Offline Training

Offline Preference Data

$D_{\text{pre}}^{(1)} = \left(x_1^{(1)}, a_{w,1}^{(1)}, a_{l,1}^{(1)}, \ldots, x_N^{(1)}, a_{w,N}^{(1)}, a_{l,N}^{(1)}\right)$

$D_{\text{pre}}^{(K)} = \left(x_1^{(K)}, a_{w,1}^{(K)}, a_{l,1}^{(K)}, \ldots, x_N^{(K)}, a_{w,N}^{(K)}, a_{l,N}^{(K)}\right)$

Reference Policy π_{ref}

Preference Pretrained Transformer $M_\theta\left(\cdot|x;\text{history}\right)$

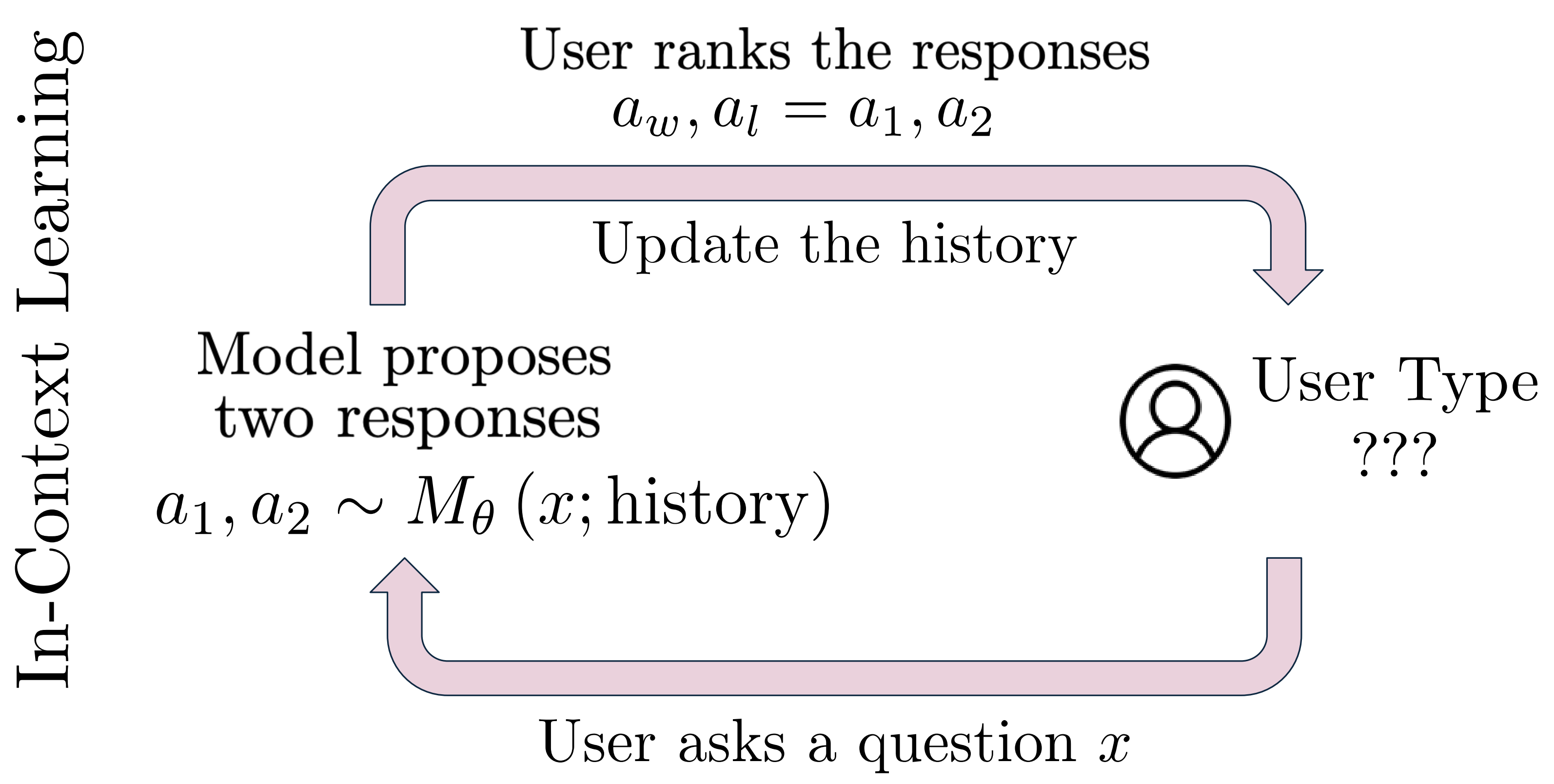
Minimize the history-dependent DPO loss

$$\mathcal{L}(\theta) = -\frac{1}{K}\sum_{g=1}^{K}\mathbb{E}_{\{(x,a_w,a_l),H\}\sim D_{\text{pre}}^{(g)}}\left[\log\sigma\left(\beta\log\frac{M_\theta(x;H)(a_w)}{\pi_{\text{ref}}(a_w|x)} - \beta\log\frac{M_\theta(x;H)(a_l)}{\pi_{\text{ref}}(a_l|x)}\right)\right]$$

In contrast to the standard DPO loss function:

$$\mathcal{L}_{\text{standard}}(\theta) = -\mathbb{E}_{(x,a_w,a_l)\sim D}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(a_w|x)}{\pi_{\text{ref}}(a_w|x)} - \beta\log\frac{\pi_\theta(a_l|x)}{\pi_{\text{ref}}(a_l|x)}\right)\right]$$

Online Deployment: User with unknown preference interacts with model



In-Context Learning

User ranks the responses $a_w, a_l = a_1, a_2$

Update the history

Model proposes two responses $a_1, a_2 \sim M_\theta\left(x;\text{history}\right)$

User Type ???

User asks a question x

→ As the interaction goes on, the model identifies the preference of the user

Experiments and Results

Proof-of-Concept Experimental Setup:

- **Context :** N_c vectors sampled from $[0,1]^3$
- **Responses:** $a', a'' \sim \text{Uniform}(\{0,1,2,3\})$
- **Preferences:** 3 subpopulations, each subpopulation follows the reward model

$$r_z(a,x) = f_\phi(x)^\top\theta_z(a) + \mathcal{N}(0,\sigma)$$

$f_\phi(x)$: context encoding θ_z: reward matrix for group z

Results

- PPT consistently outperforms PS
- PPT becomes increasingly accurate in predicting the user's preferred actions as the number of turns grow
- PPT can learn in-context effectively without the need for retraining or complex model selection procedures

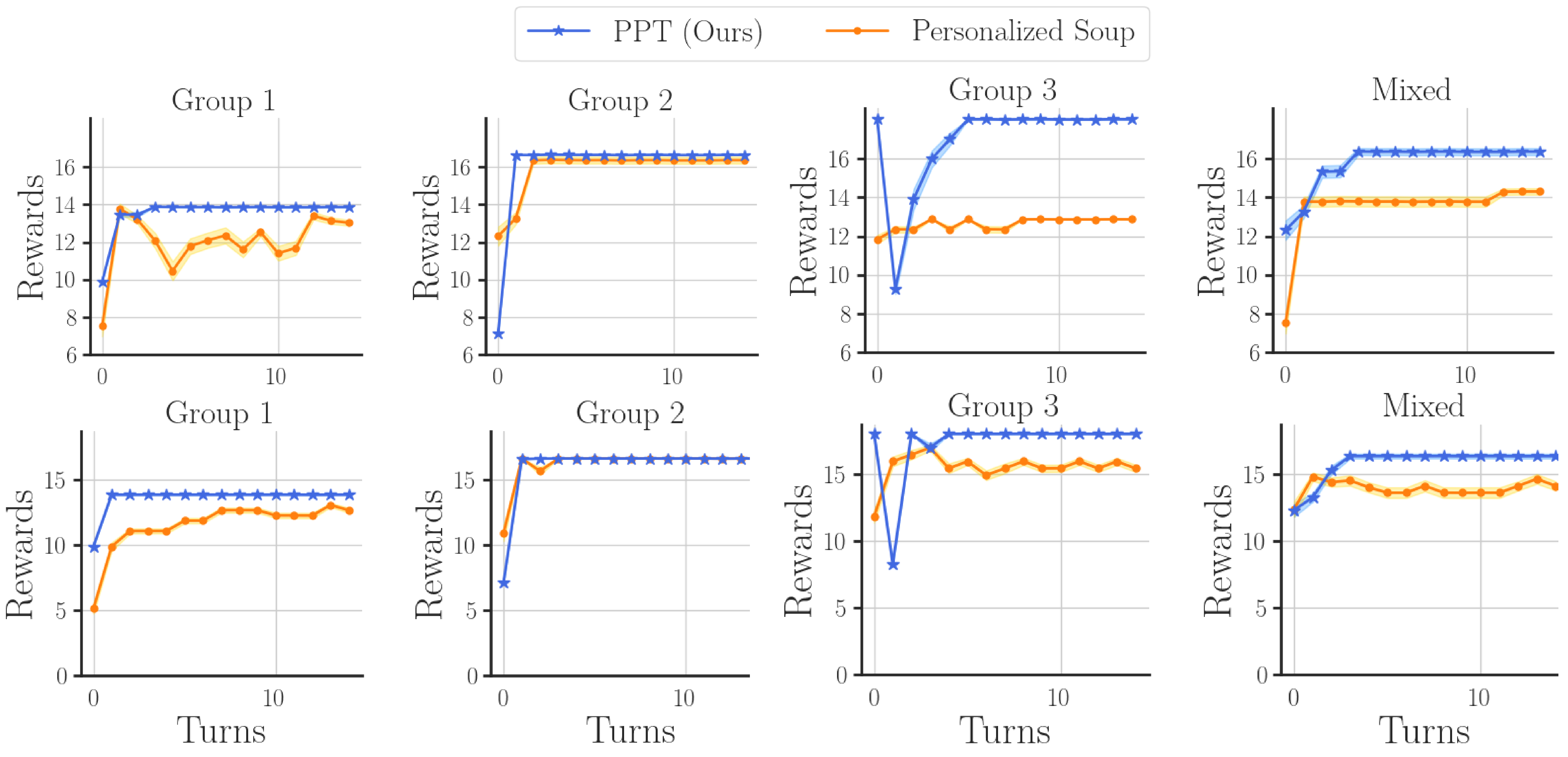


Figure 1. Comparison of rewards between PPT (ours) and Personalized Soups (PS) over 15 interaction turns for different user groups

Top row: Rewards vs Turns ($N_c = 500$)

Bottom row : Rewards vs Turns ($N_c = 1000$)

[1] 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