Whose Boat Does it Float?
Improving Personalization in Preference Tuning via Inferred User Personas

Nishant Balepur Vishakh Padmakumar Fumeng Yang
Shi Feng Rachel Rudinger Jordan Boyd-Graber

Paper

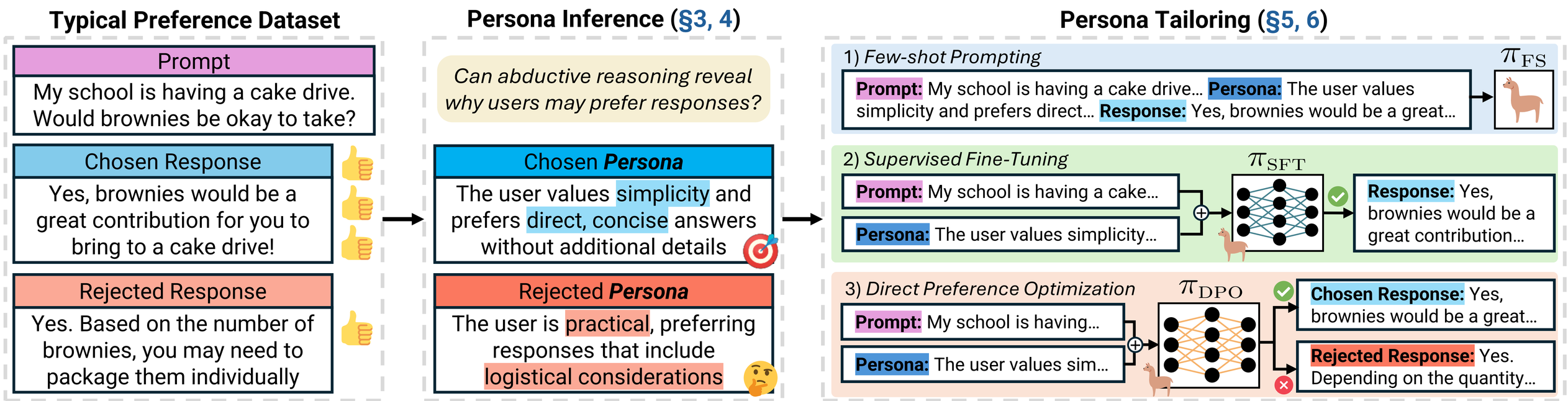
Current preference training strategies struggle to personalize to user needs…

…so we introduce a simple synthetic data strategy that generalizes to **real users!**

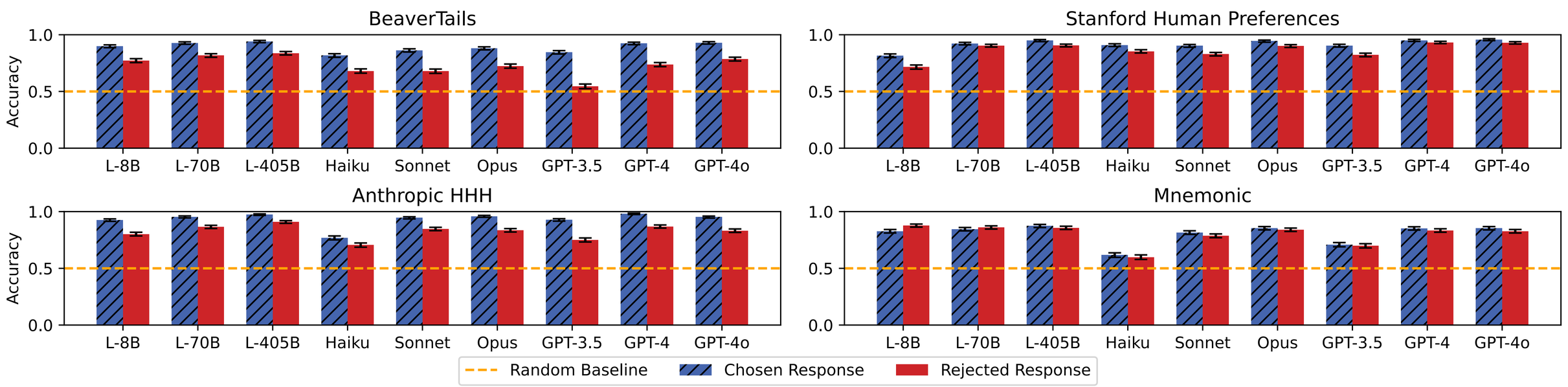
Question
I have a party tonight, can you help me find what would really liven it up? I'm sober and prefer suggestions that do not involve the use of substances

Direct Preference Optimization	Personalized DPO **(Ours)**
Sure! To liven up your party, you could… or even hire a bartender to make specialty cocktails…	Sure! … Whatever you decide, make sure it's something that everyone can enjoy and stay safe!

A new two-step pipeline for **personalization**: Persona Inference + Persona Tailoring



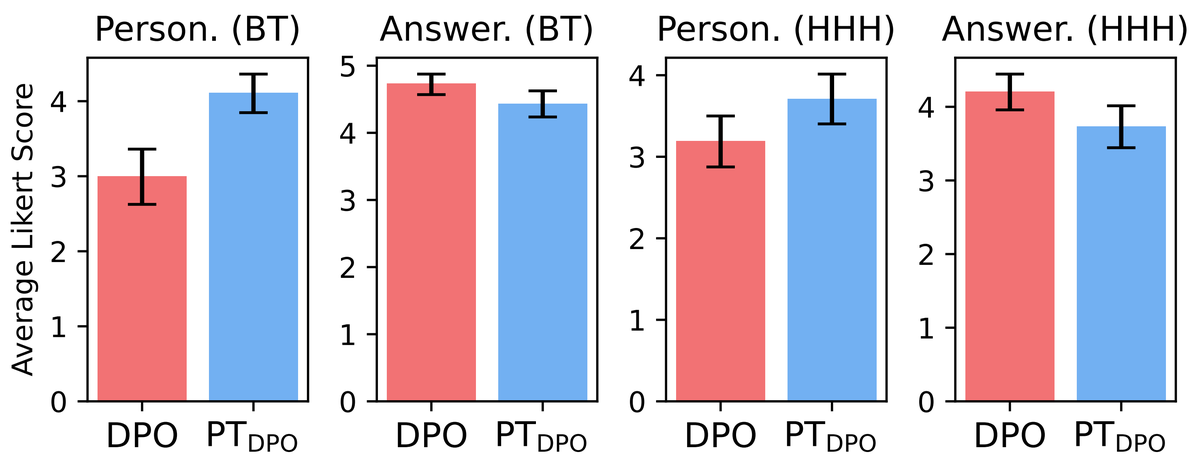
Finding 1: LLMs can infer users who would prefer chosen ***and*** rejected responses!



Finding 2: Rejected personas are hard to tailor to

Dataset	π_{base}	π_{test}	Person. W/T/L	Quality W/T/L	ΔPQ
BT Chosen	DPO+$\mathcal{P}_{retr}$	PT+$\mathcal{P}_{retr}$	**46.7**/29.3/24.0	**38.5**/30.5/31.1	+21.3
	DPO+$\mathcal{P}_{gold}$	PT+$\mathcal{P}_{gold}$	**42.3**/29.3/28.5	**34.9**/33.9/31.3	+12.5
BT Reject	DPO+$\mathcal{P}_{retr}$	PT+$\mathcal{P}_{retr}$	**45.1**/31.7/23.2	**35.1**/32.5/32.5	+17.9
	DPO+$\mathcal{P}_{gold}$	PT+$\mathcal{P}_{gold}$	**51.1**/25.9/23.0	**35.3**/32.7/32.1	+21.3
HHH Chosen	DPO+$\mathcal{P}_{retr}$	PT+$\mathcal{P}_{retr}$	**40.8**/25.4/33.8	35.0/28.0/**37.0**	+3.3
	DPO+$\mathcal{P}_{gold}$	PT+$\mathcal{P}_{gold}$	**42.0**/27.4/30.6	**39.0**/24.4/36.6	+9.4
HHH Reject	DPO+$\mathcal{P}_{retr}$	PT+$\mathcal{P}_{retr}$	**56.2**/21.0/22.8	**48.6**/24.6/26.8	+35.6
	DPO+$\mathcal{P}_{gold}$	PT+$\mathcal{P}_{gold}$	**54.1**/20.6/25.3	**44.7**/26.1/29.3	+28.6
Mnem Chosen	DPO+$\mathcal{P}_{\text{retr}}$	PT+$\mathcal{P}_{\text{retr}}$	**42.6**/31.2/26.2	**40.2**/31.6/28.2	+20.7
	DPO+$\mathcal{P}_{\text{gold}}$	PT+$\mathcal{P}_{\text{gold}}$	—	—	—
Mnem Reject	DPO+$\mathcal{P}_{\text{retr}}$	PT+$\mathcal{P}_{\text{retr}}$	**37.4**/32.6/30.0	**42.0**/27.4/30.6	+13.3
	DPO+$\mathcal{P}_{\text{gold}}$	PT+$\mathcal{P}_{\text{gold}}$	—	—	—
Average	DPO	PT_{DPO}	**45.8**/27.4/26.7	**39.3**/29.1/31.5	+18.4

Finding 3: LLMs personas generalize to real users!



Large improvements in personalization while maintaining answerability to the input query!