Customizable NLP @ACL 2026

Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment

Ramya Korlakai Vinayak

Assistant Professor

Department of Electrical & Computer Engineering
Affiliated with Dept. of CS and Dept. of Statistics

ramya@ece.wisc.edu

Joint work w/ Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, Geelon So, Gokcan Tatli, Greg Canal, Blake Mason, Rob Nowak

Outline

- Motivation: Why should we care about heterogeneity in human preferences?
- Reward modeling for preference learning
- Personalizable reward modeling
- Learning metric and multiple user preferences simultaneously
- Sample efficiency by leveraging commonalities:
  Personalizable reward modeling for pluralistic alignment (PAL)
- Experimental results
- Future directions

Generated by ChatGPT

References

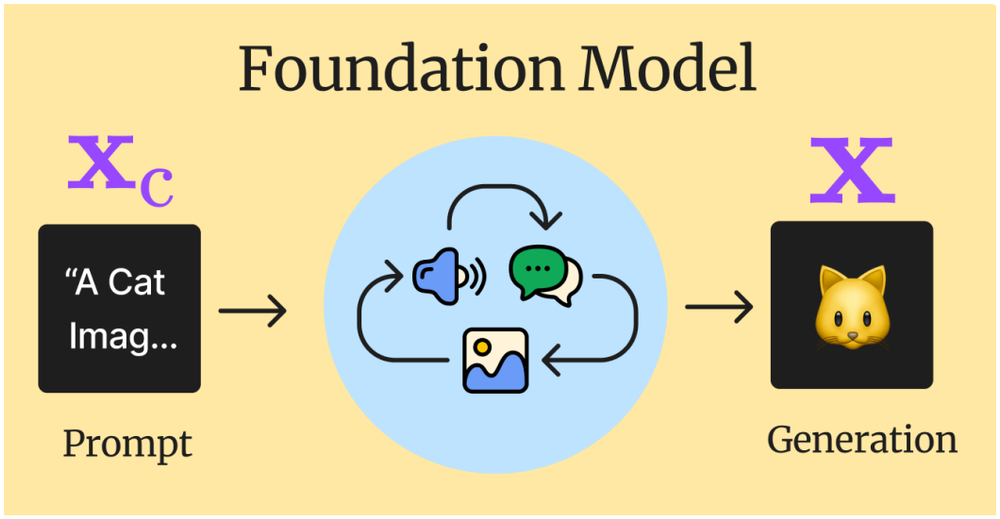
- D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, “PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment”, ICLR 2025
- Z. Wang, G. So, R. K. Vinayak, “Metric Learning from Limited Pairwise Preference Comparisons”, UAI 2024.
- G. Canal, B. Mason, R. K. Vinayak, R. Nowak, “One for All: Simultaneous Metric and Preference Learning over Multiple Users”, Neurips 2022
- G. Tatli, Y. Chen, R. K. Vinayak “Learning Population of Preferences via Pairwise Comparison Queries”, AISTATS 2024
- G. Tatli, R. Nowak, R. K. Vinayak “Metric Learning in RKHS”, UAI 2025
- Z. Wang, C. Chen, R. K. Vinayak “Bridging Lifelong Learning and Multitask Representation Learning via Algorithmic and Complexity Measure”, ALT 2026

Outline

- Motivation: Why should we care about heterogeneity in human preferences?

Generated by ChatGPT

Aligning AI/ML to human preferences



"*Alignment*":
further tuning using a lot of human feedback

Prompt: A cat Image

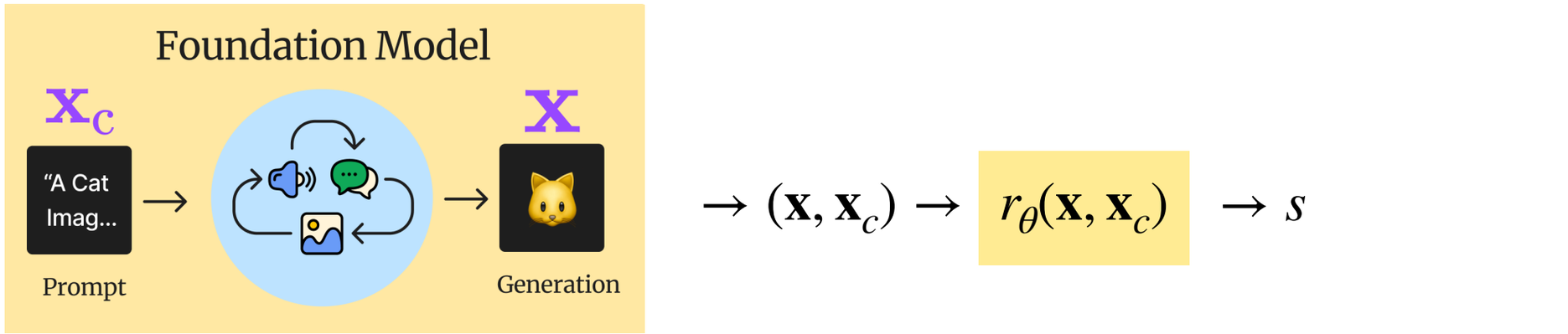
Two outputs: 🐈 , 🐈‍⬛

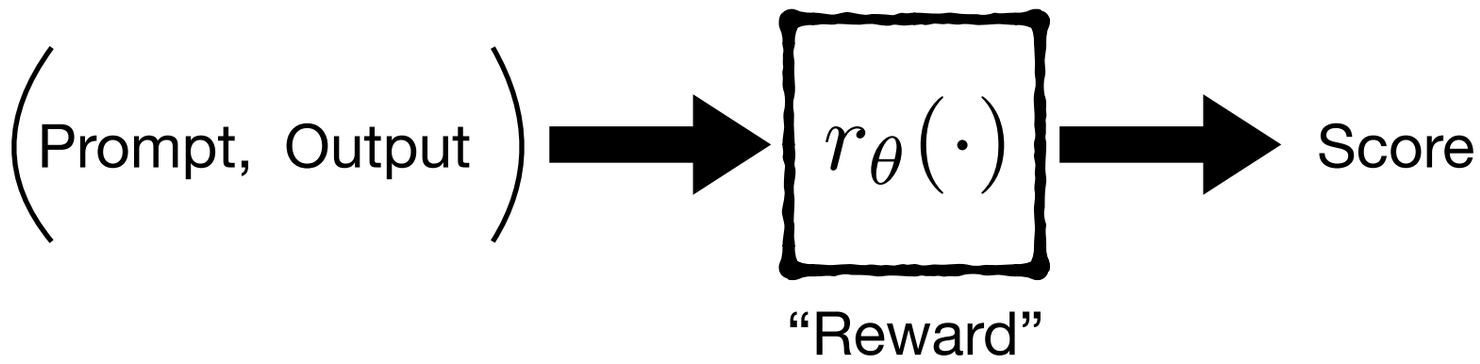
👱 : I prefer 🐈‍⬛ over 🐈

Pairwise Comparisons

- Large models trained on internet-scale data — are not readily deployable safely in the real world
- They are often very heavily fine-tuned with human feedback

Aligning AI/ML to human preferences



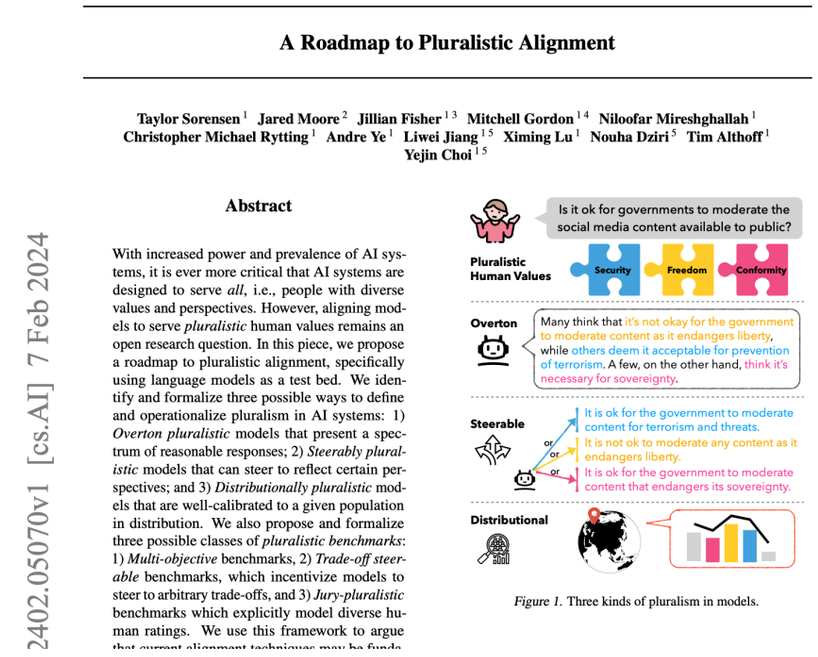


- Learn a reward function that will score the "outputs" from the pre-trained model that are preferred more by humans with higher score

- Use reward function to either (i) fine-tune the model to provide more preferred outputs or (ii) output the top ranked among multiple outputs during inference

“Plurality”: Human preferences are heterogeneous

- Alignment of AI: Whose preferences are we aligning AI to?
- We are not a monolith: different people have different preferences
- How should we model heterogeneous human preferences?
- Can we learn *personalizable* rewards with few samples per user?



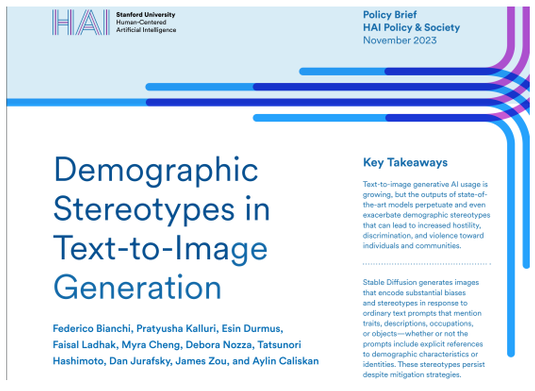
A Roadmap to Pluralistic Alignment

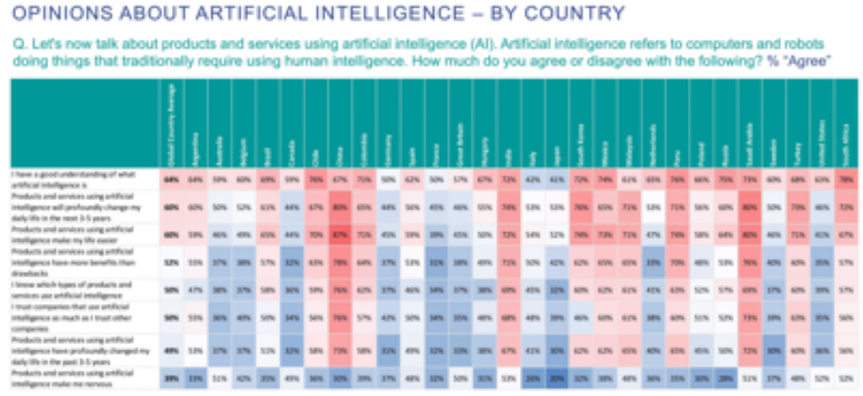
Taylor Sorensen [1] Jared Moore [2] Jillian Fisher [1 3] Mitchell Gordon [1 4] Niloofar Mireshghallah [1]
Christopher Michael Rytting [1] Andre Ye [1] Liwei Jiang [1 5] Ximing Lu [1] Nouha Dziri [5] Tim Althoff [1]
Yejin Choi [1 5]

Abstract

With increased power and prevalence of AI systems, it is ever more critical that AI systems are designed to serve *all*, i.e., people with diverse values and perspectives. However, aligning models to serve *pluralistic* human values remains an open research question. In this piece, we propose a roadmap to pluralistic alignment, specifically using language models as a test bed. We identify and formalize three possible ways to define and operationalize pluralism in AI systems: 1) *Overton pluralistic* models that present a spectrum of reasonable responses; 2) *Steerably pluralistic* models that can steer to reflect certain perspectives; and 3) *Distributionally pluralistic* models that are well-calibrated to a given population in distribution. We also propose and formalize three possible classes of *pluralistic benchmarks*: 1) *Multi-objective* benchmarks, 2) *Trade-off steerable* benchmarks, which incentivize models to steer to arbitrary trade-offs, and 3) *Jury-pluralistic* benchmarks which explicitly model diverse human ratings. We use this framework to argue

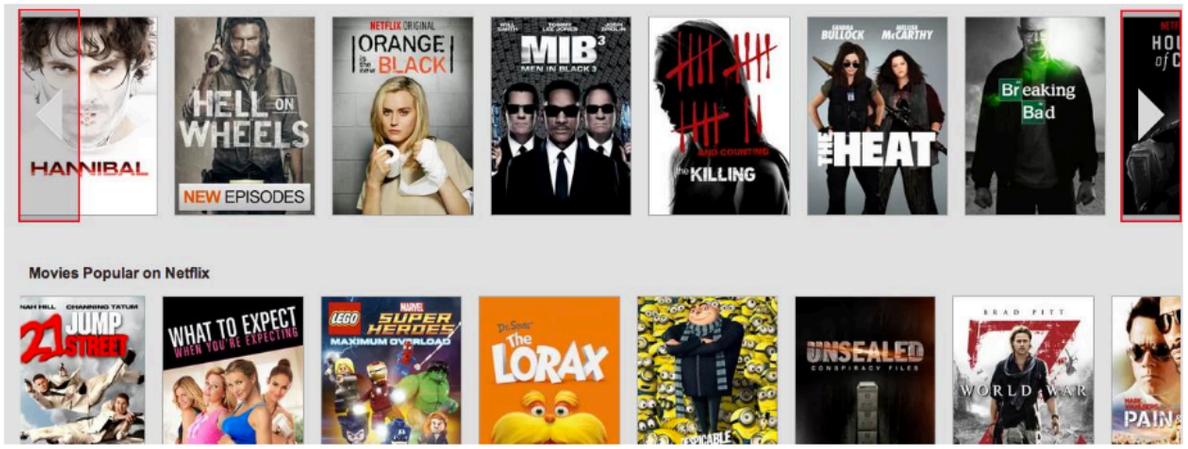
Figure 1. Three kinds of pluralism in models.

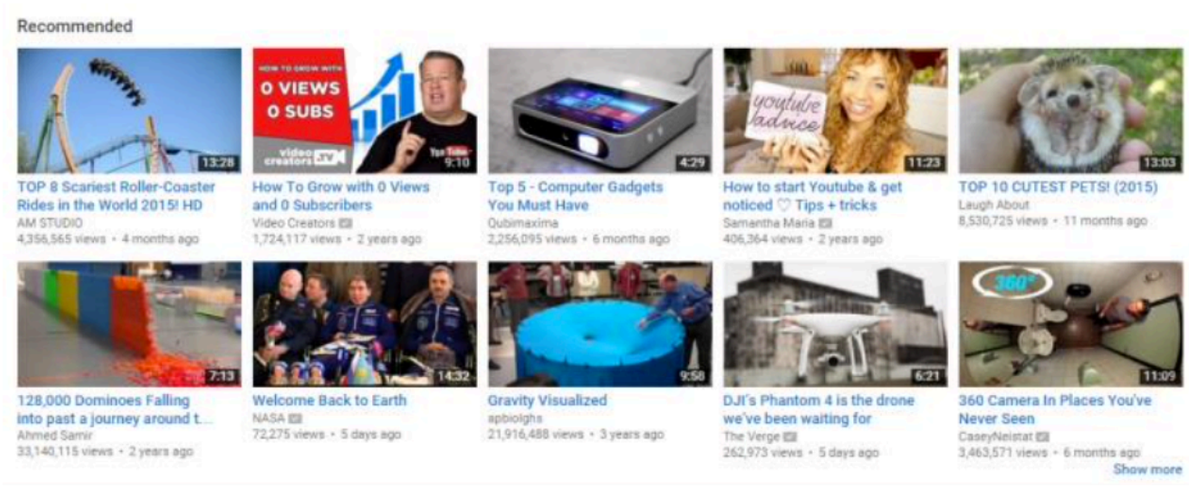




Tao et al., Cultural bias and cultural alignment of large language models." *PNAS nexus* 3.9, 2024
Bianchi et. al., Demographic Stereotypes in Text-to-image models, Stanford University HAI Policy Brief, 2023
World Economic Forum. "Artificial intelligence: Do we trust it?," *World Economic Forum*, Jan. 12, 2022. https://www.weforum.org/stories/2022/01/artificial-intelligence-ai-technology-trust-survey/

Recommendations are everywhere





Facebook thought pandemic online shopping would last forever. It didn't.

By Naomi Nix and Jaclyn Peiser

Slide courtesy: Greg Canal

Outline

- Motivation: Why should we care about heterogeneity in human preferences?
- Reward modeling for preference learning

Generated by ChatGPT

How is alignment done from human preferences?

2.2.3 Fitting the Reward Function

We can interpret a reward function estimate $\hat{r}$ as a preference-predictor if we view $\hat{r}$ as a latent factor explaining the human's judgments and assume that the human's probability of preferring a segment σ^i depends exponentially on the value of the latent reward summed over the length of the clip:[3]

$$\hat{P}\left[\sigma^1 \succ \sigma^2\right] = \frac{\exp \sum \hat{r}\left(o_t^1, a_t^1\right)}{\exp \sum \hat{r}(o_t^1, a_t^1) + \exp \sum \hat{r}(o_t^2, a_t^2)}. \tag{1}$$

We choose $\hat{r}$ to minimize the cross-entropy loss between these predictions and the actual human labels:

$$\text{loss}(\hat{r}) = -\sum_{(\sigma^1,\sigma^2,\mu)\in\mathcal{D}} \mu(1) \log \hat{P}\left[\sigma^1 \succ \sigma^2\right] + \mu(2) \log \hat{P}\left[\sigma^2 \succ \sigma^1\right].$$

Deep Reinforcement Learning from Human Preferences

Paul F Christiano OpenAI paul@openai.com
Jan Leike DeepMind leike@google.com
Tom B Brown nottombrown@gmail.com
Miljan Martic DeepMind miljanm@google.com
Shane Legg DeepMind legg@google.com
Dario Amodei OpenAI damodei@openai.com

Training language models to follow instructions with human feedback

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* Ryan Lowe*
OpenAI

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov*† **Archit Sharma***† **Eric Mitchell***†
Stefano Ermon†‡ **Christopher D. Manning**† **Chelsea Finn**†
†Stanford University ‡CZ Biohub
{rafailov,architsh,eric.mitchell}@cs.stanford.edu

- Bradley-Terry Model, 1952

$$\Pr(\mathbf{x}_l \succ \mathbf{x}_r | \mathbf{x}_c) = \frac{e^{r_\theta(\mathbf{x}_l;\mathbf{x}_c)}}{e^{r_\theta(\mathbf{x}_l;\mathbf{x}_c)} + e^{r_\theta(\mathbf{x}_r;\mathbf{x}_c)}}$$

$\mathbf{x}_c$: prompt or context

$\mathbf{x}_l$, $\mathbf{x}_r$: generated outputs

RANK ANALYSIS OF INCOMPLETE BLOCK DESIGNS

I. THE METHOD OF PAIRED COMPARISONS

BY RALPH ALLAN BRADLEY AND MILTON E. TERRY*
Virginia Agricultural Experiment Station of the Virginia Polytechnic Institute

Limitations

- Bradley-Terry Model, 1952

RANK ANALYSIS OF INCOMPLETE BLOCK DESIGNS

I. THE METHOD OF PAIRED COMPARISONS

By RALPH ALLAN BRADLEY and MILTON E. TERRY*

Virginia Agricultural Experiment Station of the Virginia Polytechnic Institute

Probability that *i* beats *j*:
(*i* is preferred over *j*)

$$P(i \succ j) = \frac{e^{r_i}}{e^{r_i} + e^{r_j}}$$

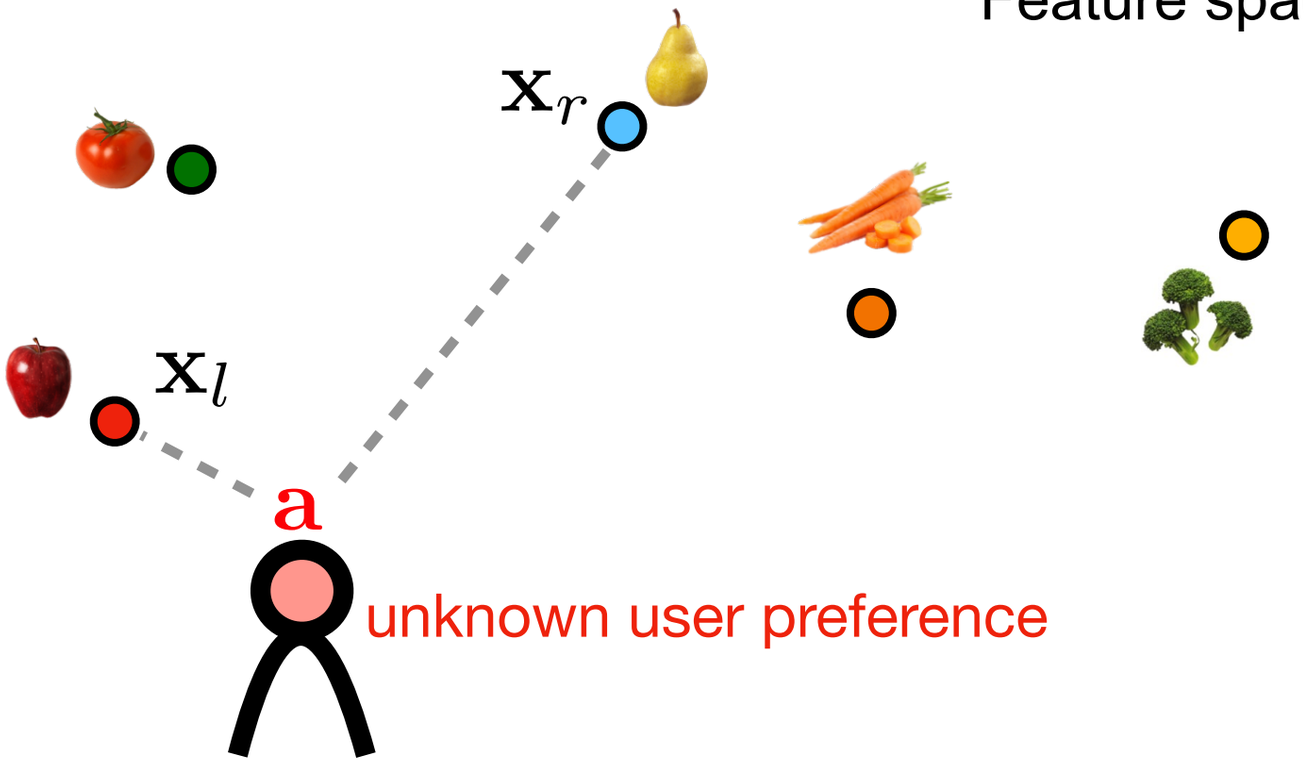
- **Assumes a single ordering** of the items/alternatives being compared
- By using BTL model, **the alignment methods are inherently assuming a single ordering of the outputs by humans**
- Differences in elicited preferences are “noisy” answers where the level of noise depends on how close the weights/scores of the two alternates are with a specific **sigmoid noise model**

Thurstone 1927, Bradley & Terry 1952, Luce 1959, Joe 1988, Hunter 2004, Ammar & Shah 2012, Neghabhan et. al. 2012, Heckel et. al. 2016 ….

Preference models: Ideal Point Model

- Coombs, 1950

Feature space $\mathcal{X} \subseteq \mathbb{R}^D$



PSYCHOLOGICAL SCALING WITHOUT A UNIT OF MEASUREMENT

BY CLYDE H. COOMBS[1]

University of Michigan

"Do you prefer  or  ?"

 preferred if $\mathrm{dist}(\mathbf{x}_l, \mathbf{a}) < \mathrm{dist}(\mathbf{x}_r, \mathbf{a})$

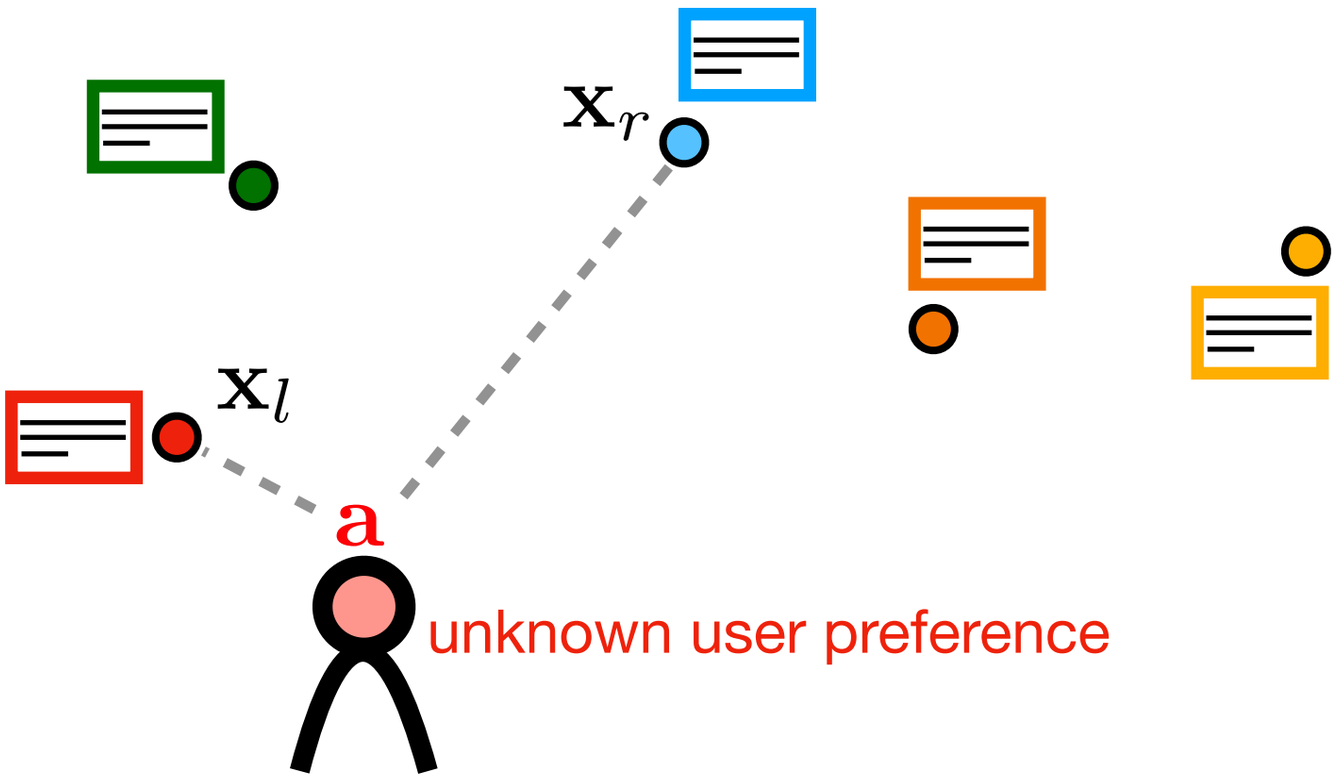
 preferred if $\mathrm{dist}(\mathbf{x}_l, \mathbf{a}) > \mathrm{dist}(\mathbf{x}_r, \mathbf{a})$

P( preferred over ) = $h(\mathrm{dist}(\mathbf{x}_l, \mathbf{a}) - \mathrm{dist}(\mathbf{x}_r, \mathbf{a}))$

h : strict monotonic function; when diff of dist is zero, the prob is 1/2, otherwise noise < 1/2

Coombs 1950, Huber 1976, Jamieson & Nowak 2011, Singla et. al 2016, Xu & Davenport 2022 ….

Preference models: Ideal Point Model



$$\mathbf{x}_l := \begin{bmatrix} \mathbf{x}_l \\ \mathbf{x}_c \end{bmatrix} \quad \mathbf{x}_r := \begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_c \end{bmatrix}$$

$\mathbf{x}_c$: context

“Do you prefer [red] or [blue] ?”

unknown user preference

[red] preferred if $\text{dist}(\mathbf{x}_l, \mathbf{a}) < \text{dist}(\mathbf{x}_r, \mathbf{a})$

[blue] preferred if $\text{dist}(\mathbf{x}_l, \mathbf{a}) > \text{dist}(\mathbf{x}_r, \mathbf{a})$

P([red] preferred over [blue]) = $h(\text{dist}(\mathbf{x}_l, \mathbf{a}) - \text{dist}(\mathbf{x}_r, \mathbf{a}))$

h : strict monotonic function; when diff of dist is zero, the prob is 1/2, otherwise noise < 1/2

P([red] preferred over [blue]) = $\dfrac{1}{1 + e^{\text{dist}(\mathbf{x}_r, \mathbf{a}) - \text{dist}(\mathbf{x}_l, \mathbf{a})}}$

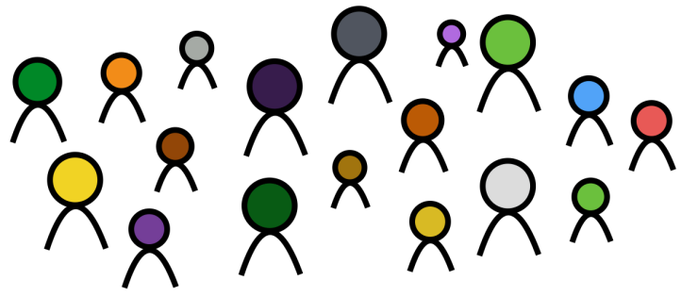
Difference of scores

Past works

- Universal models: a model (preference point) that fits on average for all individuals
- Distance function is assumed to be known

However….

- People have heterogeneous tastes (preferences)
- Distance function is often unknown



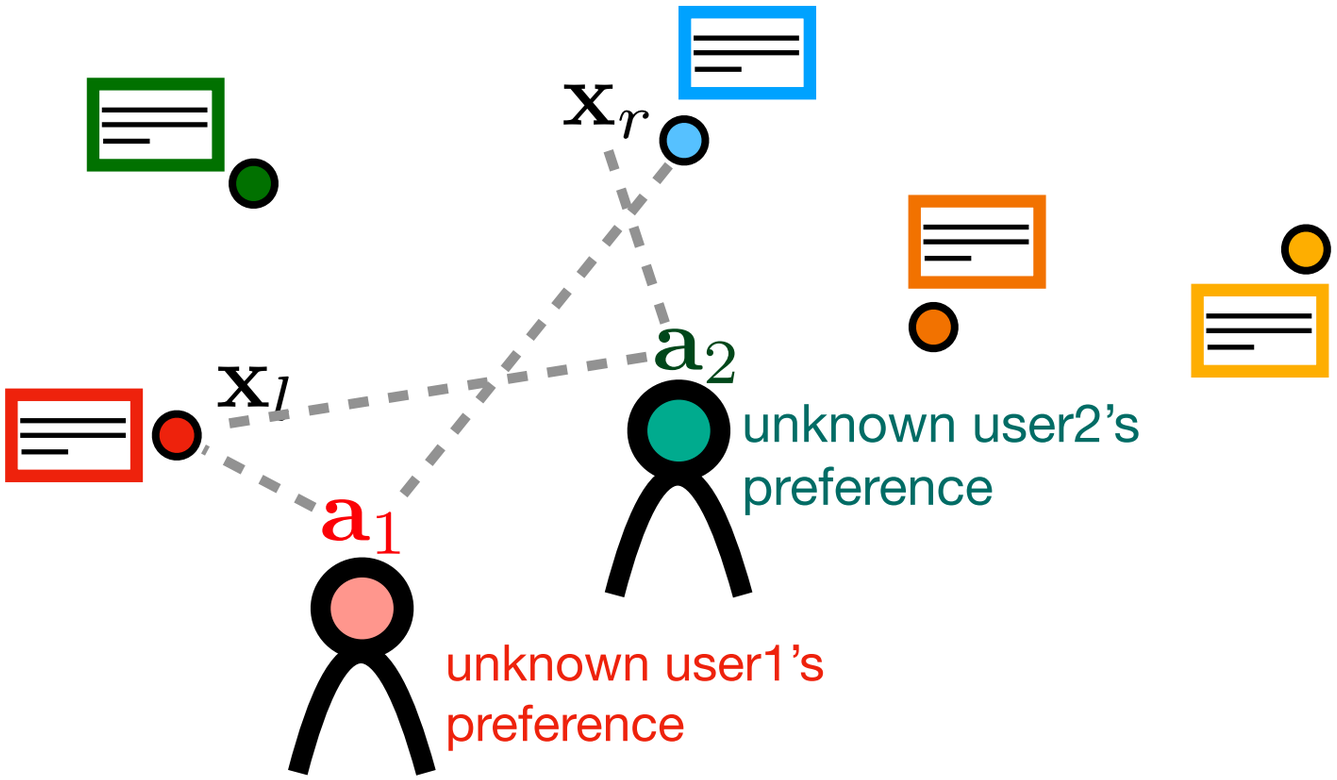
Coombs 1950, Huber 1976, Jamieson & Nowak 2011, Singla et. al 2016, Massimino & Davenport 2021 ….

Outline

- Motivation: Why should we care about heterogeneity in human preferences?
- Reward modeling for preference learning
- Personalizable reward modeling

Generated by ChatGPT

Personalized Preferences



$$\mathbf{x}_l := \begin{bmatrix} \mathbf{x}_l \\ \mathbf{x}_c \end{bmatrix} \quad \mathbf{x}_r := \begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_c \end{bmatrix}$$

$\mathbf{x}_c$: context

"Do you prefer  or  ?"

 preferred if $\mathrm{dist}(\mathbf{x}_l, \mathbf{a}_i) < \mathrm{dist}(\mathbf{x}_r, \mathbf{a}_i)$

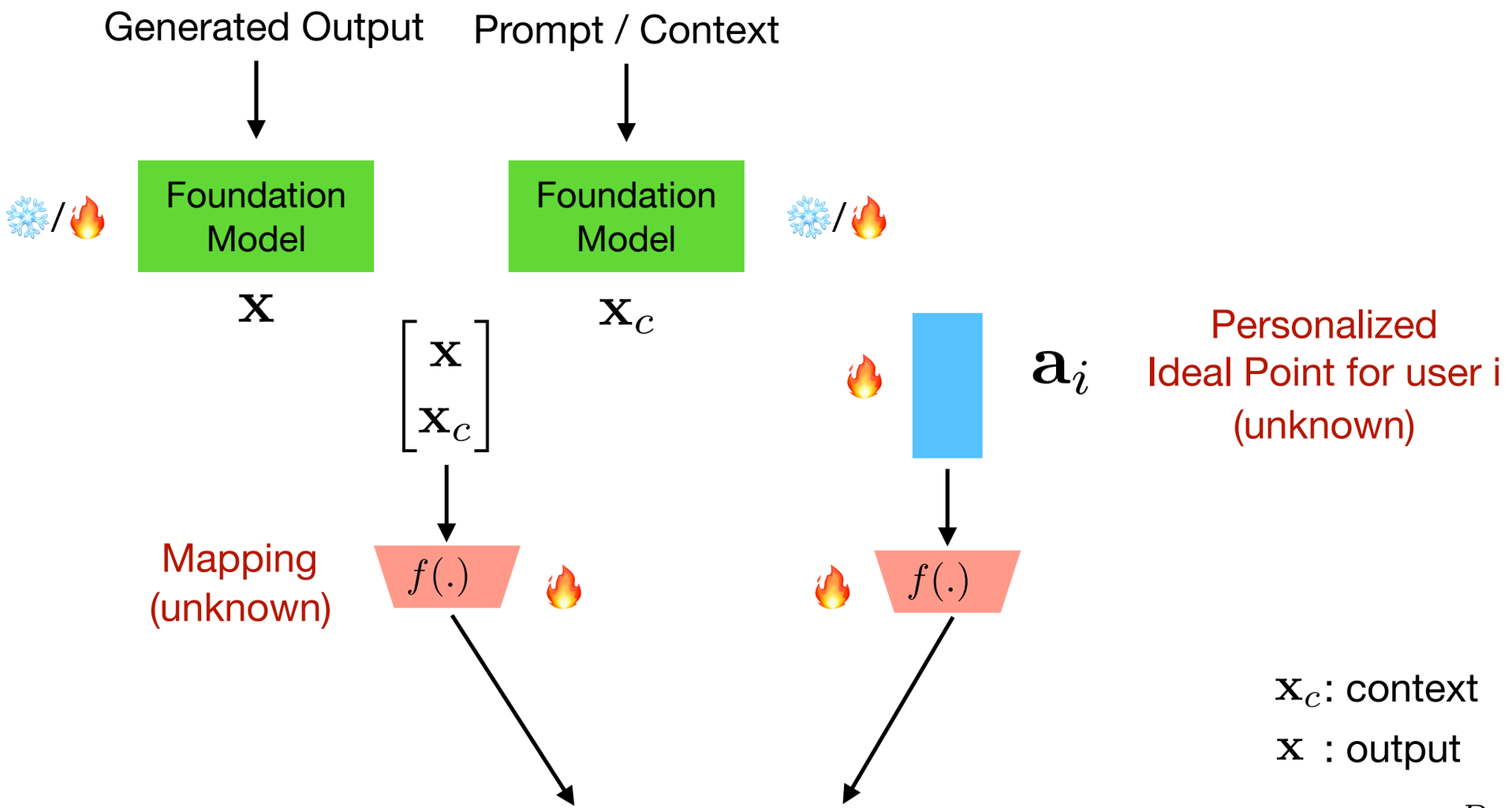
 preferred if $\mathrm{dist}(\mathbf{x}_l, \mathbf{a}_i) > \mathrm{dist}(\mathbf{x}_r, \mathbf{a}_i)$

$i = 1, 2$

P( preferred over.  | user i) = $h(\mathrm{dist}(\mathbf{x}_l, \mathbf{a}_i) - \mathrm{dist}(\mathbf{x}_r, \mathbf{a}_i))$

h : strict monotonic function; when diff of dist is zero, the prob is 1/2, otherwise noise < 1/2

Personalizable Reward Modeling



Personalized Reward For user i:

$$r^{(i)}(\mathbf{x};\mathbf{x}_c) = ||f([\mathbf{x};\mathbf{x}_c]) - f(\mathbf{a}_i)||^2$$

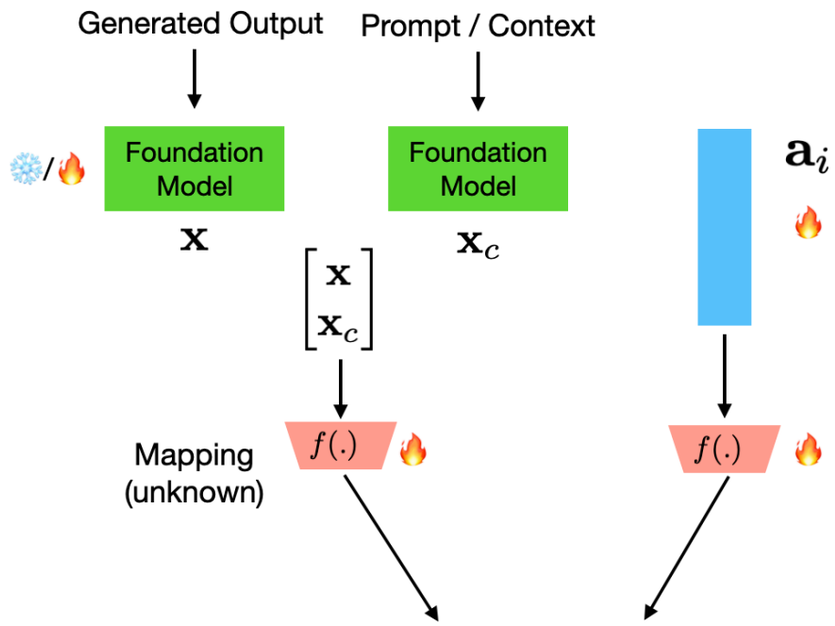
$\mathbf{x}_c$: context

$\mathbf{x}$: output

$\mathbf{x}, \mathbf{x}_c \in \mathbb{R}^D \qquad a_i \in \mathbb{R}^{2D}$

$f : \mathbb{R}^{2D} \rightarrow \mathbb{R}^d$

Learning Personalized Preferences



Personalized Reward For user i:

$$r^{(i)}(\mathbf{x};\mathbf{x}_c) = ||f([\mathbf{x};\mathbf{x}_c]) - f(\mathbf{a}_i)||^2$$

$\mathbf{x}, \mathbf{x}_c \in \mathbb{R}^D \qquad a_i \in \mathbb{R}^{2D}$

$f : \mathbb{R}^{2D} \rightarrow \mathbb{R}^d$

- Learn the unknown mapping f and unknown preference points from pairwise comparison data

$$\underset{f,\mathbf{a}_i}{\arg\min} \sum_p \text{loss}\left(y_{p,i}\left(||f([\mathbf{x}_{l_{p,i}};\mathbf{x}_{c_{p,i}}]) - f(\mathbf{a}_i)||^2 - ||f([\mathbf{x}_{r_{p,i}};\mathbf{x}_{c_{p,i}}]) - f(\mathbf{a}_i)||^2\right)\right)$$

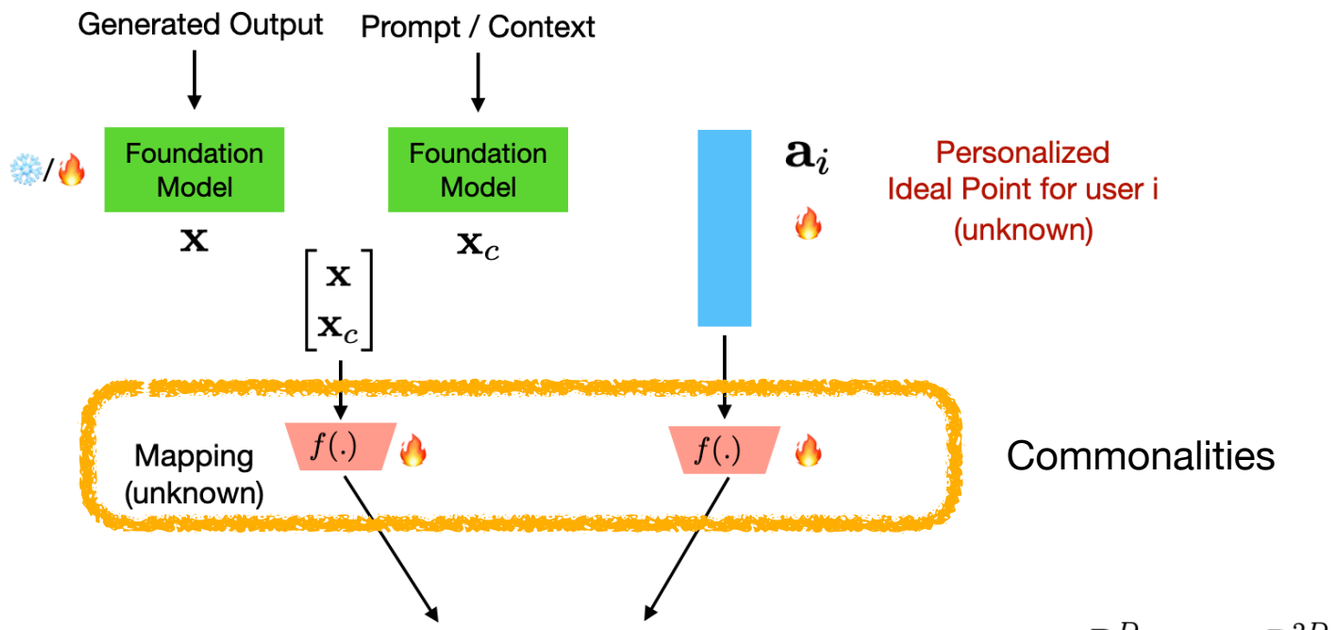
- If we learn individually, each user needs to provide too much data to learn all the components

Outline

- Motivation: Why should we care about heterogeneity in human preferences?
- Reward modeling for preference learning
- Personalizable reward modeling
- Learning metric and multiple user preferences simultaneously

Generated by ChatGPT

Simultaneous Metric and Preference Learning



Personalized Reward For user i:

$$r^{(i)}(\mathbf{x};\mathbf{x}_c) = ||f([\mathbf{x};\mathbf{x}_c]) - f(\mathbf{a}_i)||^2$$

$\mathbf{x}, \mathbf{x}_c \in \mathbb{R}^D \qquad a_i \in \mathbb{R}^{2D}$

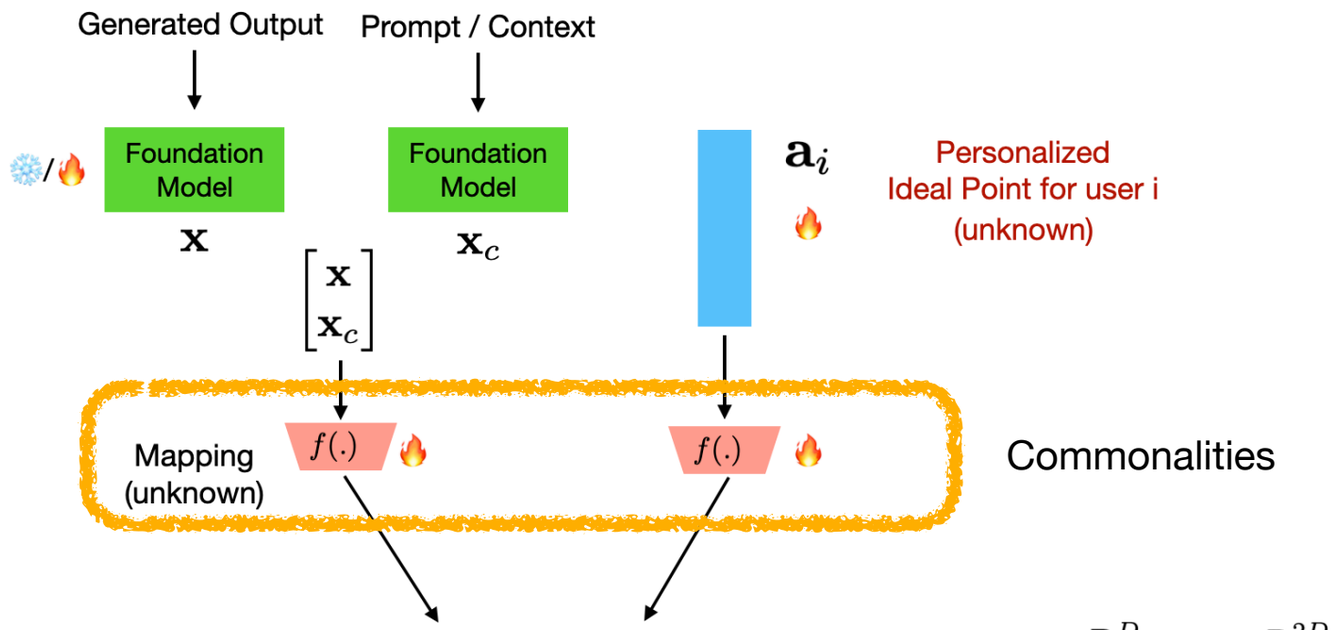
$f : \mathbb{R}^{2D} \to \mathbb{R}^d$

- Leverage everyone's data to learn the common mapping f
- Use individual data to learn the personalized preference

$$\underset{f,\mathbf{a}_1,\ldots,\mathbf{a}_N}{\arg\min} \sum_i \sum_p \text{loss}\left(y_{p,i}\left(||f([\mathbf{x}_{l_{p,i}};\mathbf{x}_{c_{p,i}}]) - f(\mathbf{a}_i)||^2 - ||f([\mathbf{x}_{r_{p,i}};\mathbf{x}_{c_{p,i}}]) - f(\mathbf{a}_i)||^2\right)\right)$$

Canal, Mason, Vinayak, Nowak, "One for all: Simultaneous metric and preference learning over multiple users", Neurips 2022

Simultaneous Metric and Preference Learning

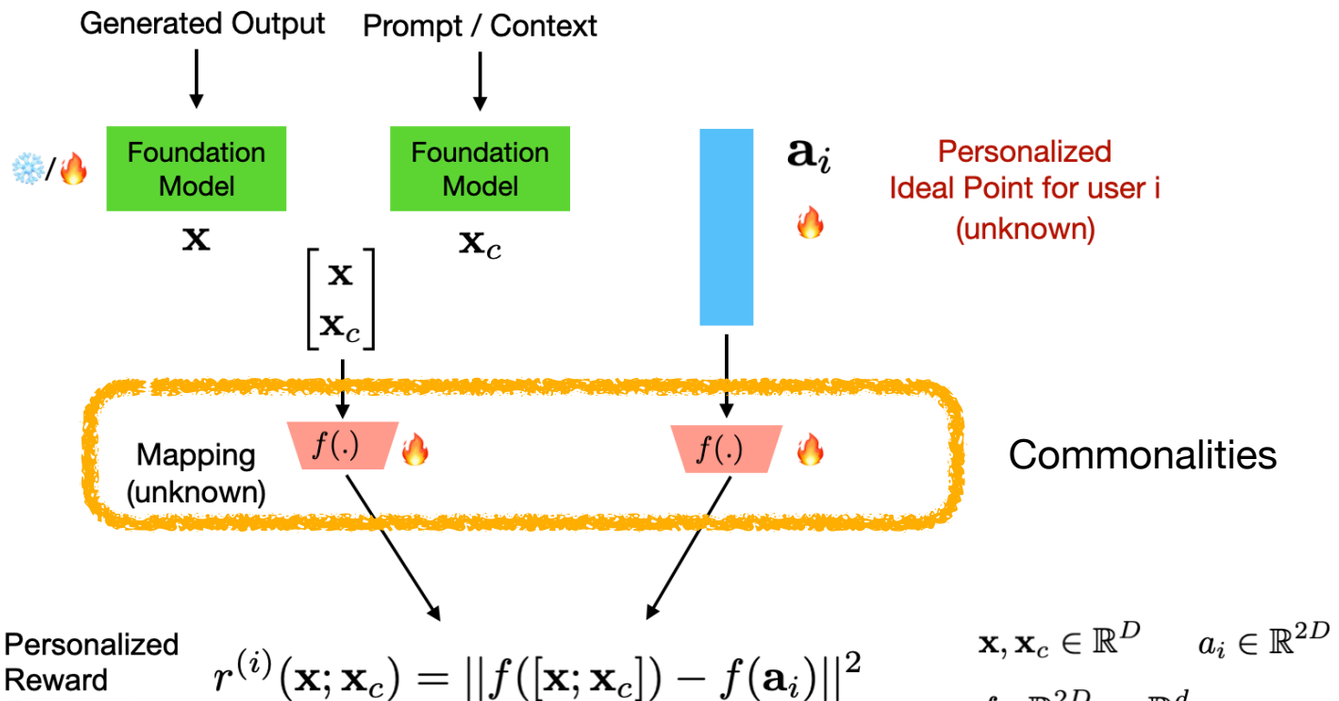


- Cost of learning f is amortized across all users
- With sufficiently large number of users in the dataset, each user only needs to provide enough data to learn their latent preference, i.e., ideal point

O(#dimensions)

Canal, Mason, Vinayak, Nowak, “One for all: Simultaneous metric and preference learning over multiple users”, Neurips 2022

Metric Learning from Pairwise Comparisons



- Each user needs to provide enough data to learn their latent ideal point

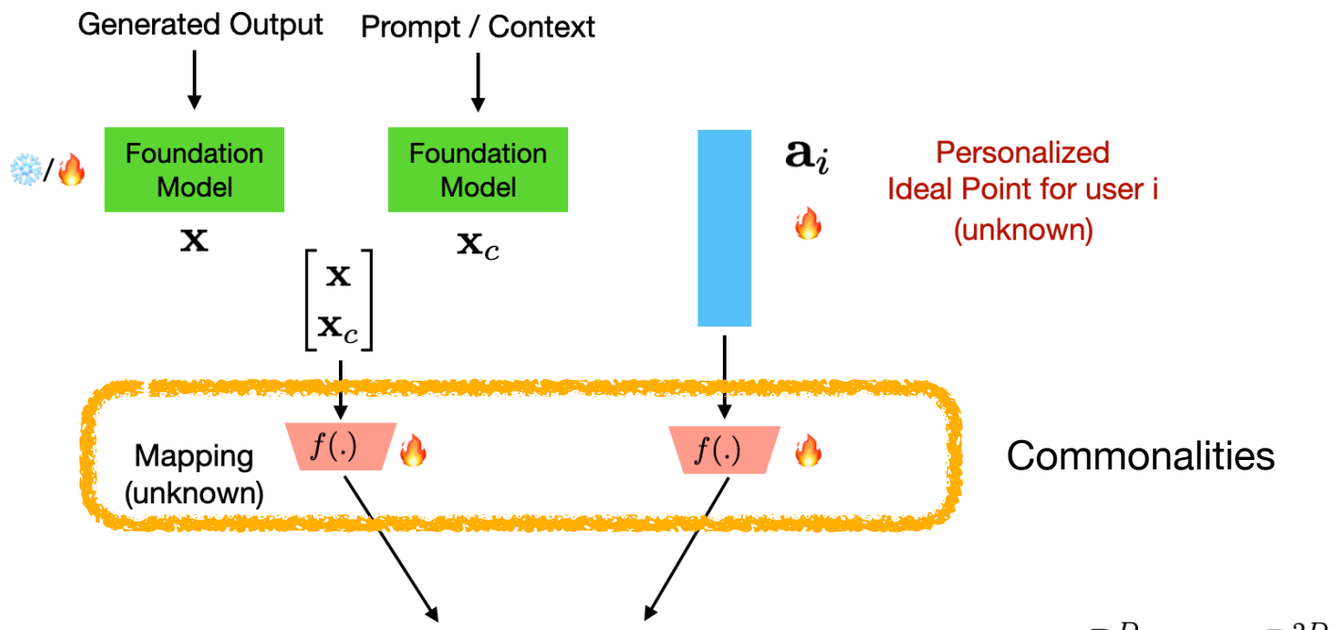
O(#dimensions)

- High-dimensional — costly

Can we learn the unknown mapping with fewer samples than O(#dimensions)?

Wang, So, Vinayak, “Metric Learning from Limited Pairwise Preference Comparisons”, UAI 2024, arXiv:2403.19629, 2024.

Metric Learning from Pairwise Comparisons



Can we learn the unknown mapping without localizing the preferences?

NO when there is no low-dimensional structure to leverage

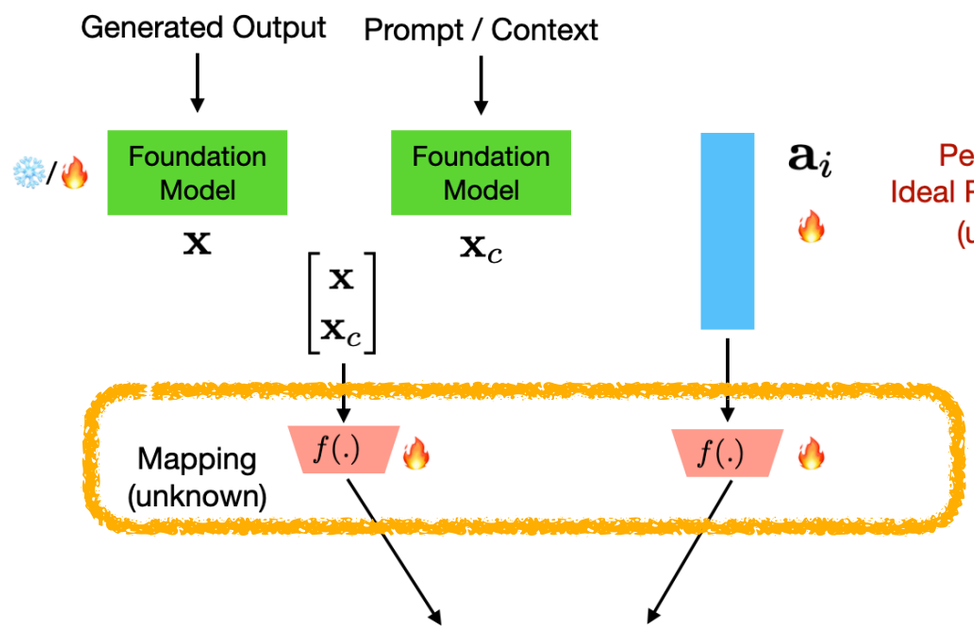
Wang, So, Vinayak, "Metric Learning from Limited Pairwise Preference Comparisons", UAI 2024, arXiv:2403.19629, 2024.

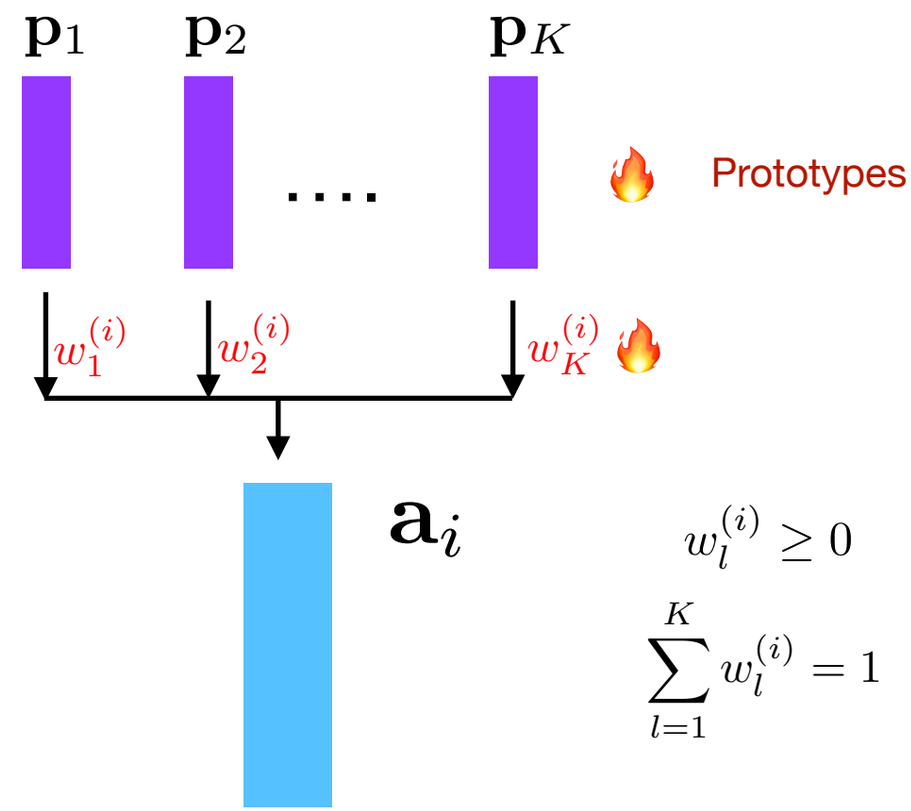
Outline

- Motivation: Why should we care about heterogeneity in human preferences?
- Reward modeling for preference learning
- Personalizable reward modeling
- Learning metric and multiple user preferences simultaneously
- Sample efficiency by leveraging commonalities:
  Personalizable reward modeling for pluralistic alignment (PAL)

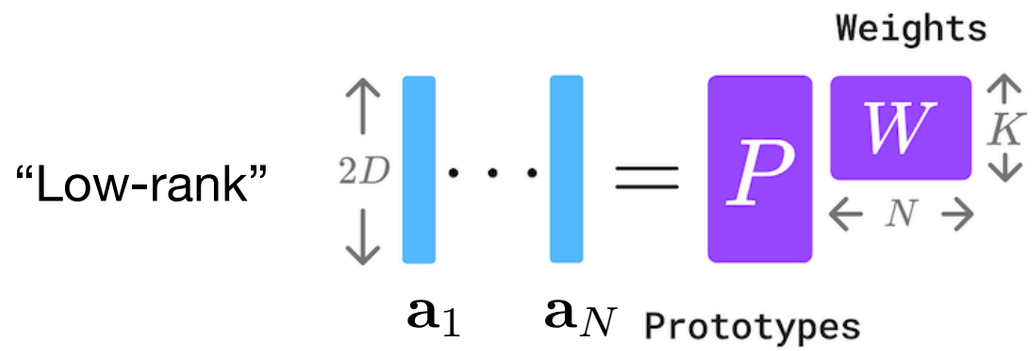
Generated by ChatGPT

Personalizable Reward Modeling: PAL A



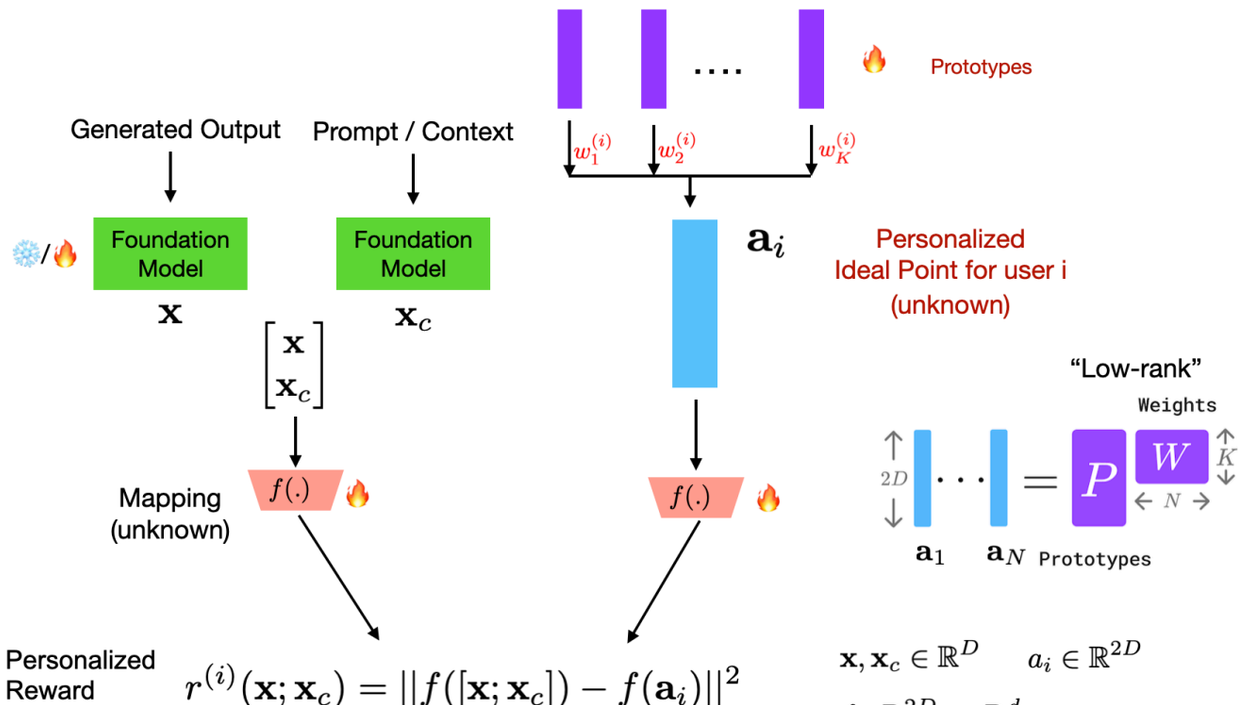


- Each user's preference is modeled as a mixture of K prototypical preferences



D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, "PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment", ICLR 2025

Learning Personalizable Rewards: PAL A



- With sufficient number of users in the dataset, each user needs to only provide enough data to learn their ***weights***

O(# of Prototypes)

$$K \ll D$$

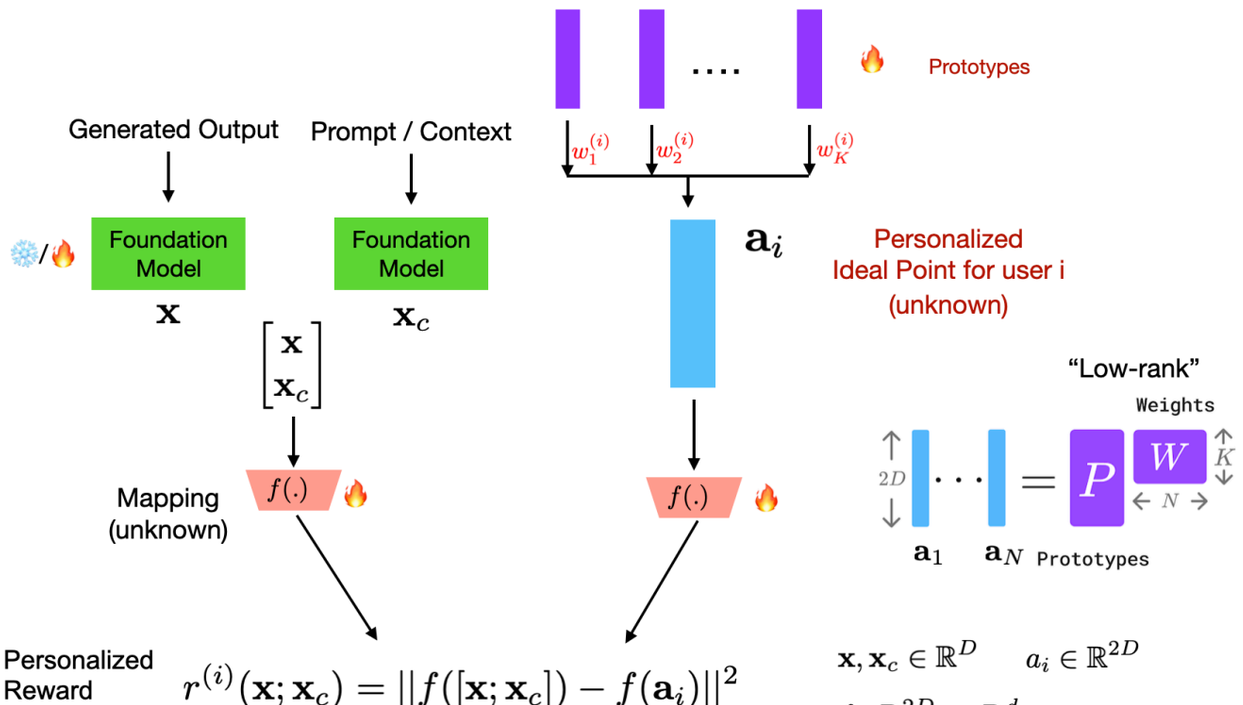
- Leveraging commonalities and low-rank structure in user preferences further reduces the samples needed per user

$$\underset{f,\mathbf{p}_1,\ldots,\mathbf{p}_K,w^{(1)},\ldots,w^{(N)}}{\arg\min} \sum_i \sum_p \text{loss}\left(y_{p,i}\left(||f([\mathbf{x}_{l_{p,i}};\mathbf{x}_{c_{p,i}}]) - f(\mathbf{a}_i)||^2 - ||f([\mathbf{x}_{r_{p,i}};\mathbf{x}_{c_{p,i}}]) - f(\mathbf{a}_i)||^2\right)\right)$$

$$\text{where} \quad \mathbf{a}_i = \sum_{l=1}^{K} w_l^{(i)} p_l \qquad \sum_{l=1}^{K} w_l^{(i)} = 1 \qquad w_l^{(i)} \geq 0 \qquad \forall l, i$$

D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, "PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment", ICLR 2025

Few-shot Generalization to New Users



- New user needs to only provide enough data to learn their ***weights***

O(# of Prototypes)

$$K \ll D$$

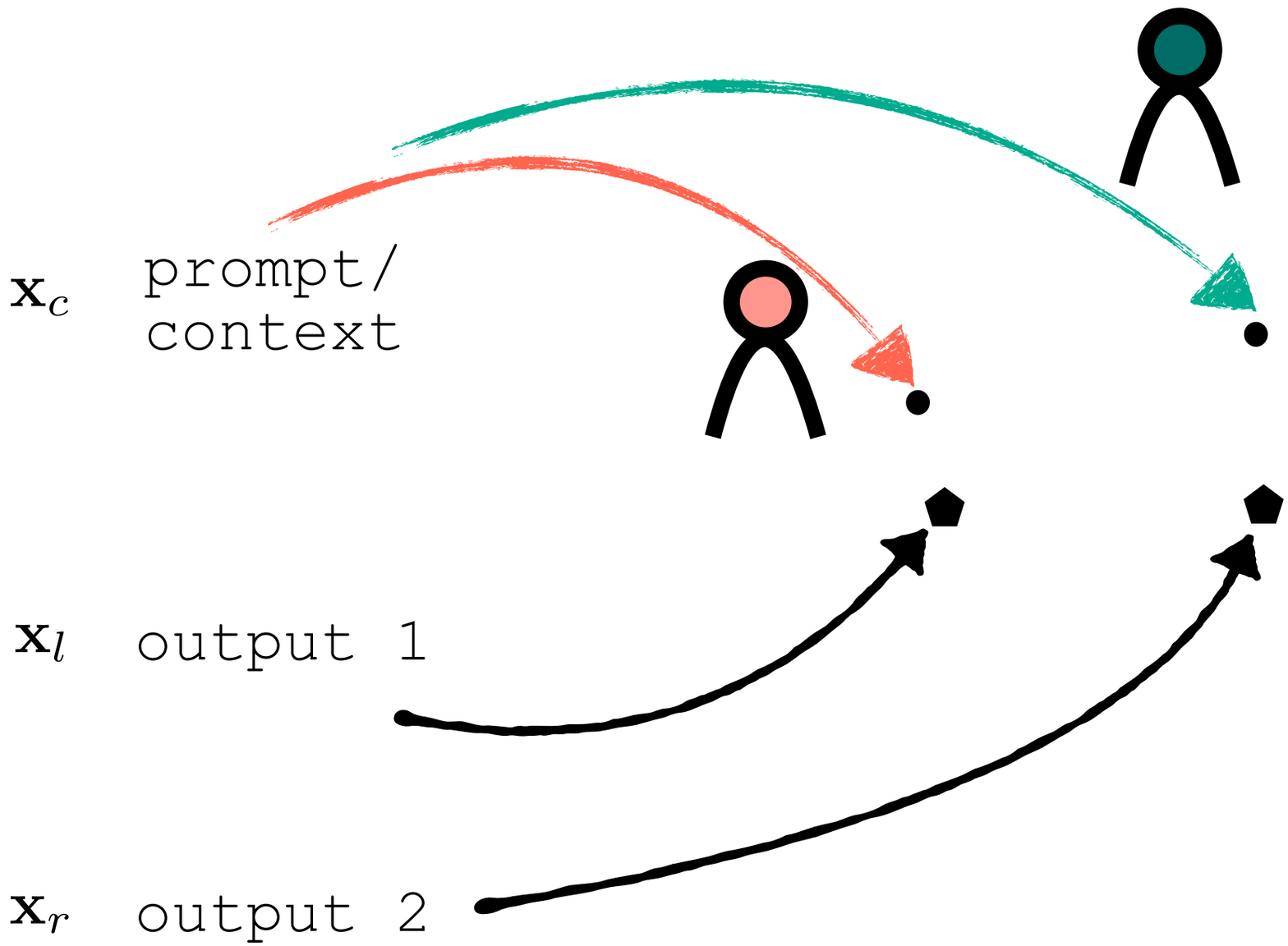
- Mapping f and prototypes are kept fixed
- Few-shot learning to only learn the weights for the new user

$$\arg\min_{w^{(i)}} \sum_p \text{loss}\left(y_{p,i}\left(||f([\mathbf{x}_{l_{p,i}};\mathbf{x}_{c_{p,i}}]) - f(\mathbf{a}_i)||^2 - ||f([\mathbf{x}_{r_{p,i}};\mathbf{x}_{c_{p,i}}]) - f(\mathbf{a}_i)||^2\right)\right)$$

$$\text{where} \quad \mathbf{a}_i = \sum_{l=1}^{K} w_l^{(i)} p_l \qquad \sum_{l=1}^{K} w_l^{(i)} = 1 \qquad w_l^{(i)} \geq 0 \qquad \forall l, i$$

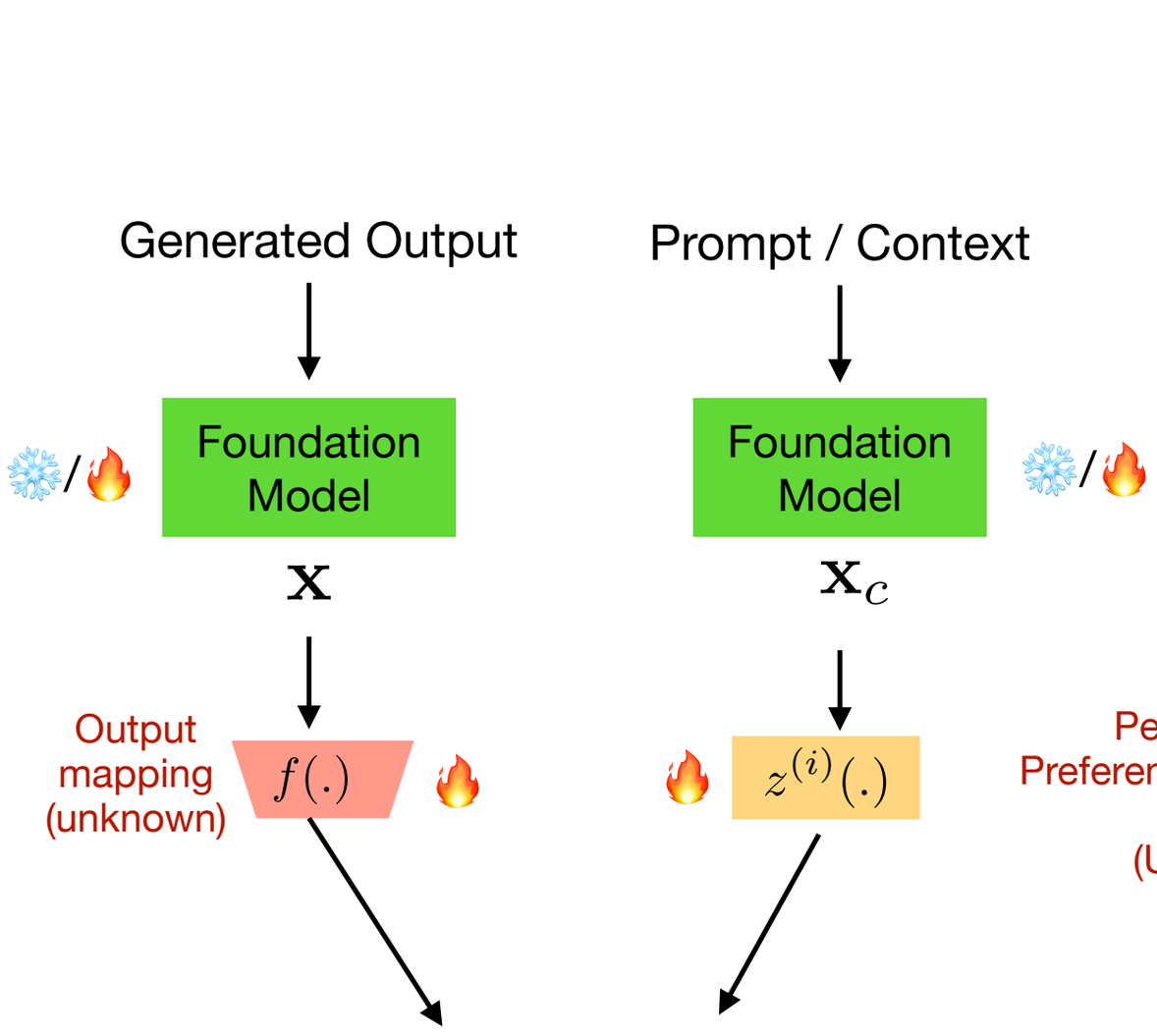
D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, "PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment", ICLR 2025

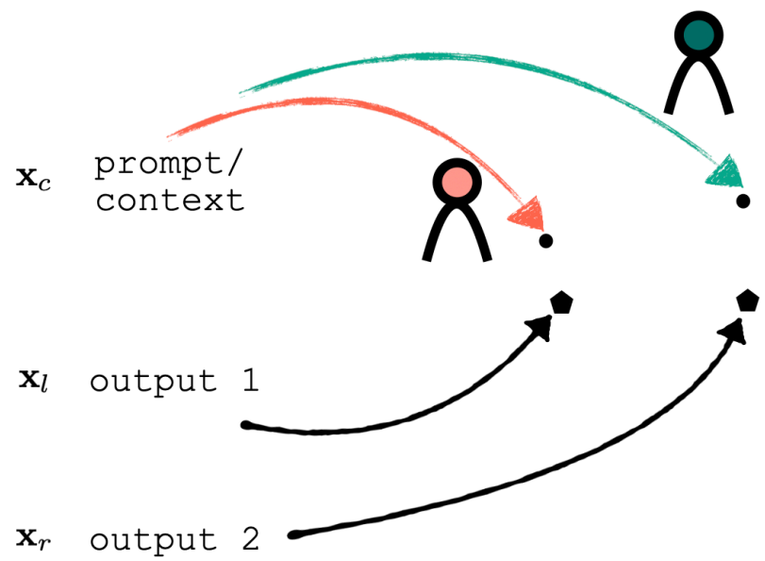
Capturing diverse human preferences via preference mappings



D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, “PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment”, ICLR 2025

Personalizable Reward Modeling: PAL B





Preference mappings

Personalized Reward For user i:

$$r_\theta^{(i)}(\mathbf{x}; \mathbf{x}_c) = \langle f(\mathbf{x}), z^{(i)}(\mathbf{x}_c) \rangle$$

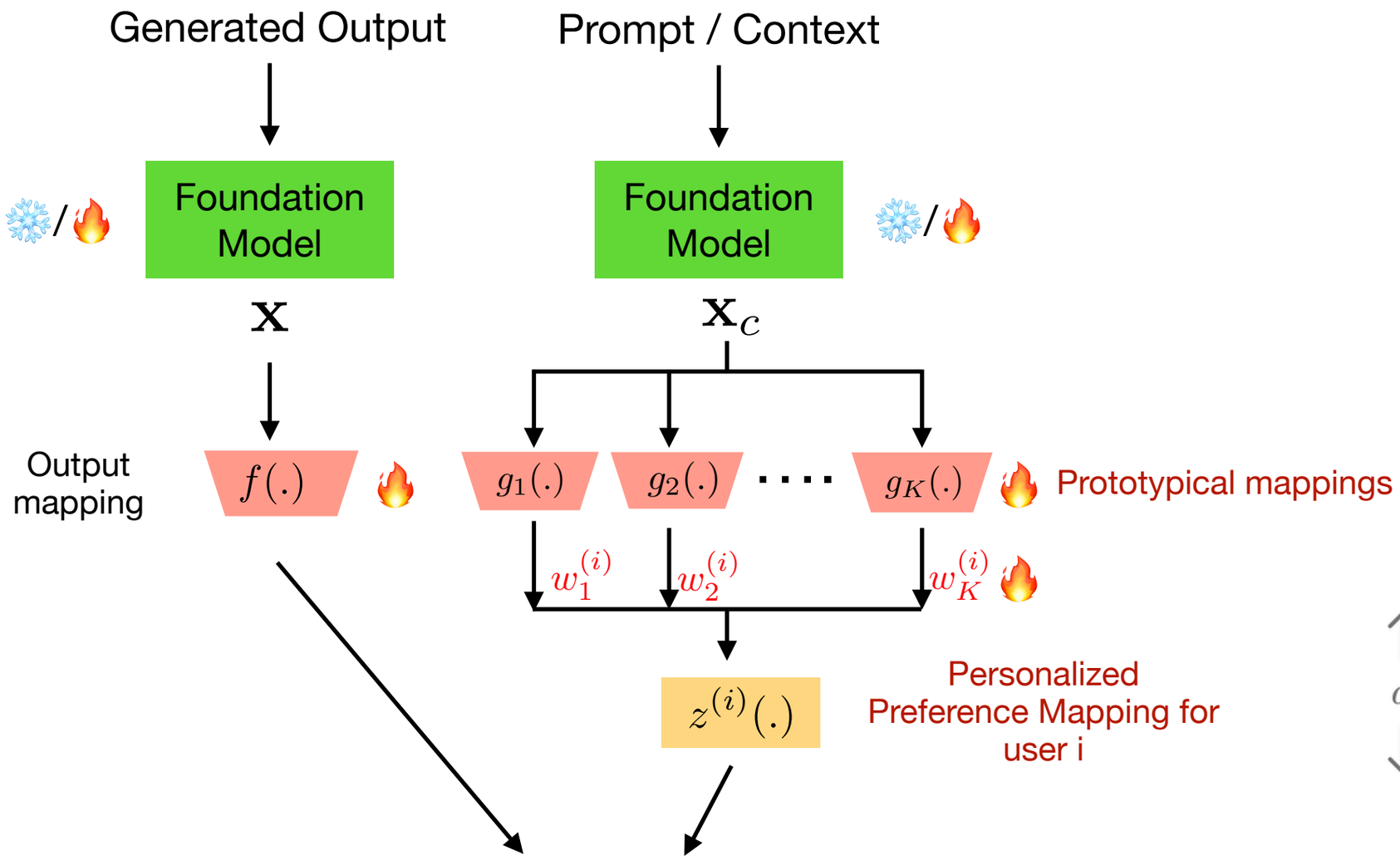
$\mathbf{x}_c$: context

$\mathbf{x}$: output

$\mathbf{x}, \mathbf{x}_c \in \mathbb{R}^D \quad f, z^{(i)} : \mathbb{R}^D \to \mathbb{R}^d$

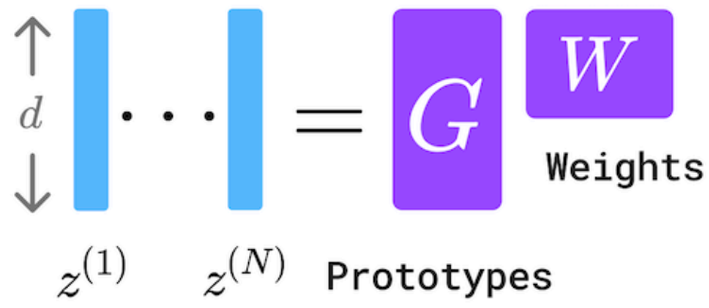
D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, “PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment”, ICLR 2025

Personalizable Reward Modeling: PAL B



$$w_l^{(i)} \geq 0$$

$$\sum_{l=1}^{K} w_l^{(i)} = 1$$



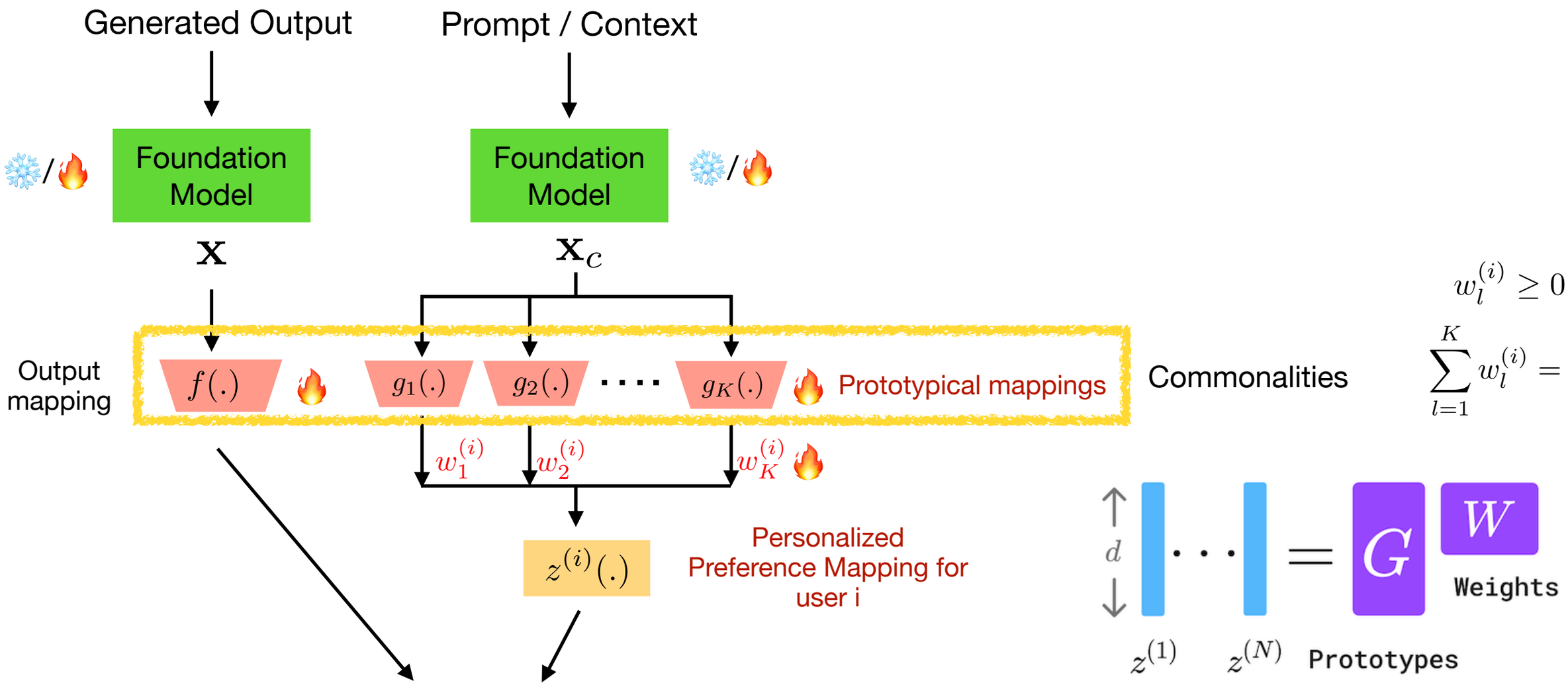
Personalized Reward For user i:

$$r_\theta^{(i)}(\mathbf{x}; \mathbf{x}_c) = \langle f(\mathbf{x}), z^{(i)}(\mathbf{x}_c) \rangle$$

$$\mathbf{x}, \mathbf{x}_c \in \mathbb{R}^D \quad f, z^{(i)} : \mathbb{R}^D \to \mathbb{R}^d$$

D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, “PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment”, ICLR 2025

Personalizable Reward Modeling: PAL B



$$w_l^{(i)} \geq 0$$

$$\sum_{l=1}^{K} w_l^{(i)} = 1$$

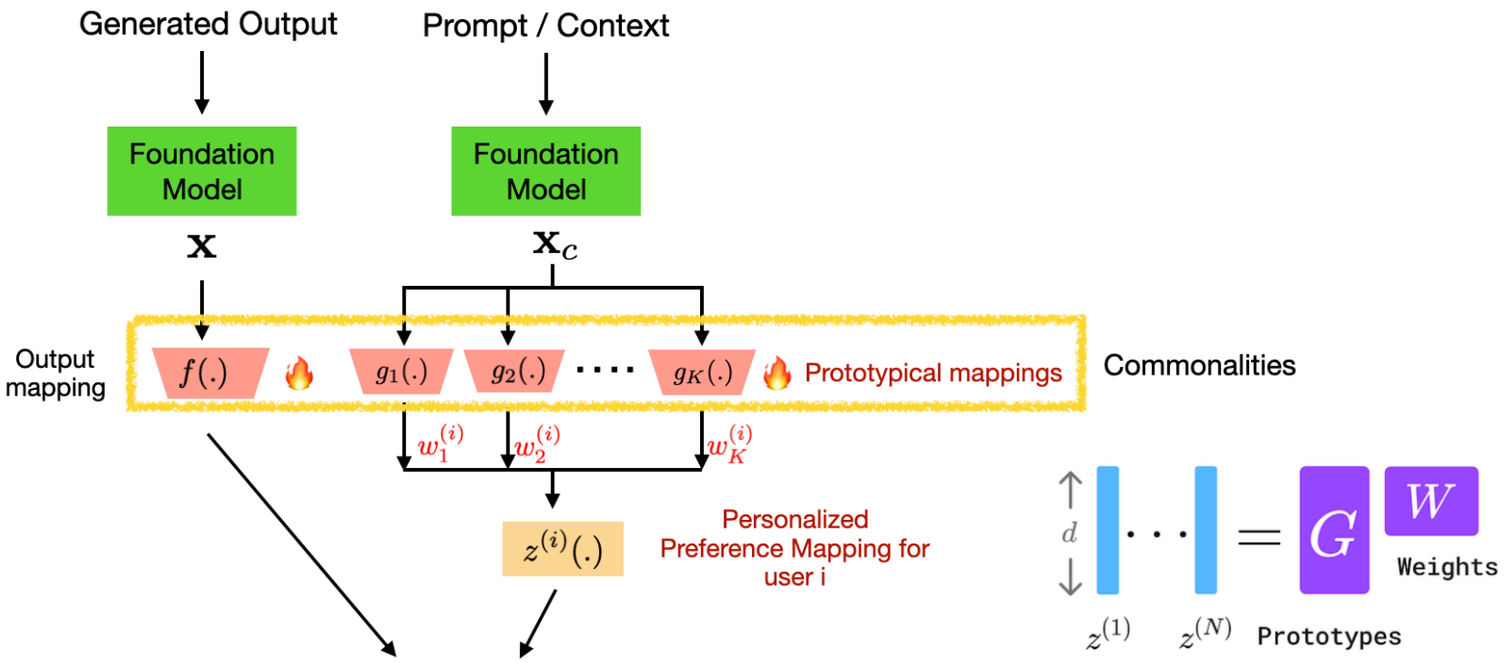
Personalized Reward For user i:

$$r_\theta^{(i)}(\mathbf{x}; \mathbf{x}_c) = \langle f(\mathbf{x}), z^{(i)}(\mathbf{x}_c) \rangle$$

$$\mathbf{x}, \mathbf{x}_c \in \mathbb{R}^D \quad f, z^{(i)} : \mathbb{R}^D \to \mathbb{R}^d$$

D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, “PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment”, ICLR 2025

Learning Personalizable Rewards: PAL B



Personalized Reward For user i: $r_\theta^{(i)}(\mathbf{x};\mathbf{x}_c) = \langle f(\mathbf{x}), z^{(i)}(\mathbf{x}_c)\rangle$ $\quad \mathbf{x},\mathbf{x}_c \in \mathbb{R}^D \quad f, z^{(i)} : \mathbb{R}^D \to \mathbb{R}^d$

- Leveraging commonalities and low-rank structure in user preferences reduces the samples needed per user

$$\underset{f,g_1,\ldots,g_K,w^{(1)},\ldots,w^{(N)}}{\arg\min} \sum_i \sum_p \text{loss}\left(y_{p,i}\left(\langle f(\mathbf{x}_{l_{p,i}}), z^{(i)}(\mathbf{x}_{c_{p,i}})\rangle - \langle f(\mathbf{x}_{r_{p,i}}), z^{(i)}(\mathbf{x}_{c_{p,i}})\rangle\right)\right)$$

$$\text{where} \quad \mathbf{a}_i = \sum_{l=1}^{K} w_l^{(i)} p_l \qquad \sum_{l=1}^{K} w_l^{(i)} = 1 \qquad w_l^{(i)} \geq 0 \qquad \forall l, i$$

D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, “PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment”, ICLR 2025

Learning Personalizable Rewards: PAL B

Algorithm 2 Learning algorithm for `PAL-B`

Require: Dataset $\mathcal{D} = \bigcup_{i=1}^{N} \left\{ \left(\mathbf{x}_{j,l}^{(i)}, \mathbf{x}_{j,r}^{(i)}; \mathbf{x}_{j,c}^{(i)}, y_j^{(i)} \right) \right\}_{j=1}^{m_i}$, loss function ℓ, mapping function f_θ, prototype mapping functions $\{g_{\theta_k}\}_{k=1}^{K}$, user weights $\{\mathbf{w}^{(i)} := [w_1^{(i)}, ..., w_K^{(i)}]\}_{i=1}^{N}$.

1: **for** each iteration **do**
2: **Sample** a mini-batch $\left\{ \left(\mathbf{x}_{j,l}^{(i)}, \mathbf{x}_{j,r}^{(i)}; \mathbf{x}_{j,c}^{(i)}, y_j^{(i)} \right) \right\}$ ▷ random pairs, not ordered by users
3: **User Ideal Point (conditioned on prompts)**:
4: $\mathbf{a}^{(i)} = \left[g_{\theta_i}(\mathbf{x}_{c,j}^{(i)}), ..., g_{\theta_K}(\mathbf{x}_{c,j}^{(i)}) \right]^\top \cdot \mathbf{w}^{(i)}$
5: **Distance**:
6: $d_{l,j}^{(i)} = \langle f_\theta \left(\mathbf{x}_{l,j}^{(i)} \right), \mathbf{a}^{(i)} \rangle, \quad d_{r,j}^{(i)} = \langle f_\theta \left(\mathbf{x}_{r,j}^{(i)} \right), \mathbf{a}^{(i)} \rangle$
7: **Loss**: $\psi_j^{(i)}(\mathbf{x}_{l,j}^{(i)}, \mathbf{x}_{r,j}^{(i)}; \mathbf{x}_{c,j}^{(i)}, y_j^{(i)}) = \ell(y_j^{(i)} \cdot (d_{l,j}^{(i)} - d_{r,j}^{(i)}))$
8: **Update Step:** $\text{argmin}_{\theta, \mathbf{P}, \{\mathbf{w}^{(i)}\}_{i=1}^{N}} \sum \psi_j^{(i)}(\mathbf{x}_{l,j}^{(i)}, \mathbf{x}_{r,j}^{(i)}; \mathbf{x}_{c,j}^{(i)}, y_j^{(i)})$
9: **end for**

D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, “PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment”, ICLR 2025

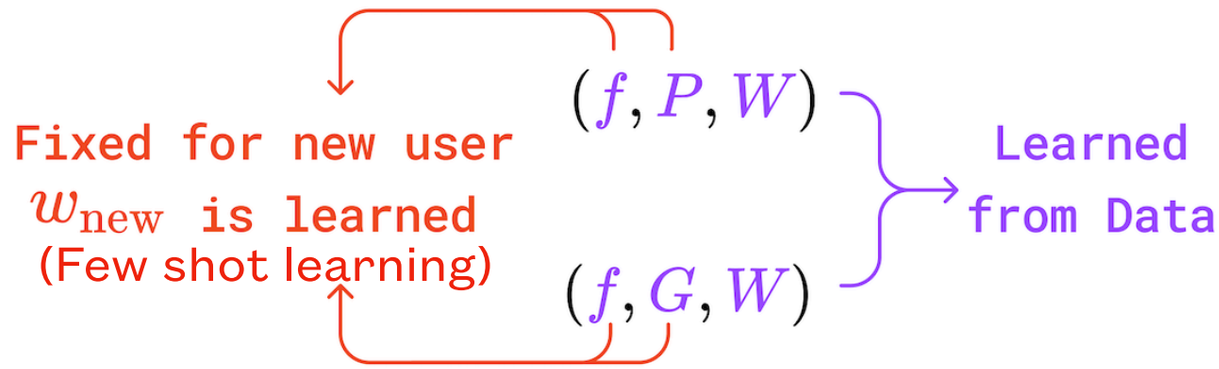
Outline

- Motivation: Why should we care about heterogeneity in human preferences?
- Reward modeling for preference learning
- Personalizable reward modeling
- Learning metric and multiple user preferences simultaneously
- Sample efficiency by leveraging commonalities:
  Personalizable reward modeling for pluralistic alignment (PAL)
- Experimental results

Generated by ChatGPT

Training and Generalization

- Learn the unknown mapping f and unknown prototypes from pairwise comparisons



Two types of generalization for preference test accuracy:

1. ***Seen*** accuracy: on users present in train set for unseen pairs
2. ***Unseen*** accuracy: on users not present in train set

Adapt to Unseen Users

- Part of data to localize users
- Only learn weights W
- Fix mappings and prototypes

D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, “PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment”, ICLR 2025

Experiments: Pick-a-pic dataset

Text-to-Image generation

Model	Params	v1 Acc (%)
CLIP-H/14	986M	59.23
ImageReward	447M	61.10
HPS	986M	66.70
PickScore	**986M**	**71.85**
PAL-A-Tiny	8.4M	69.29 ± 0.6
PAL-B-Tiny	**6.3M**	**71.13 ± 0.3**

PickScore fine tunes last 20 layers of CLIP-H

Our models: train 2-layer MLP on the CLIP-H embeddings K = 1

Model	Train Dataset	Test Accuracy on Pick-a-Pic v2 (%)	
		No-leakage	Leakage
CLIP-H/14	-	62.57	58.59
PickScore	pickapic v1	68.04	74.16*
PAL-B-Tiny on CLIP-H	**pickapic v1**	**70.02 ± 0.39**	79.32 ± 1.68*
PAL-B-Tiny on CLIP-H	**pickapic v2**	**70.51 ± 0.22**	**68.67 ± 0.51**
PAL-B-Tiny on PickScore	pickapic v2	70.16 ± 0.19	74.79 ± 0.13*

Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., and Levy, O. Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems, 36, 2023.

Experiments: Summary dataset

Language generation

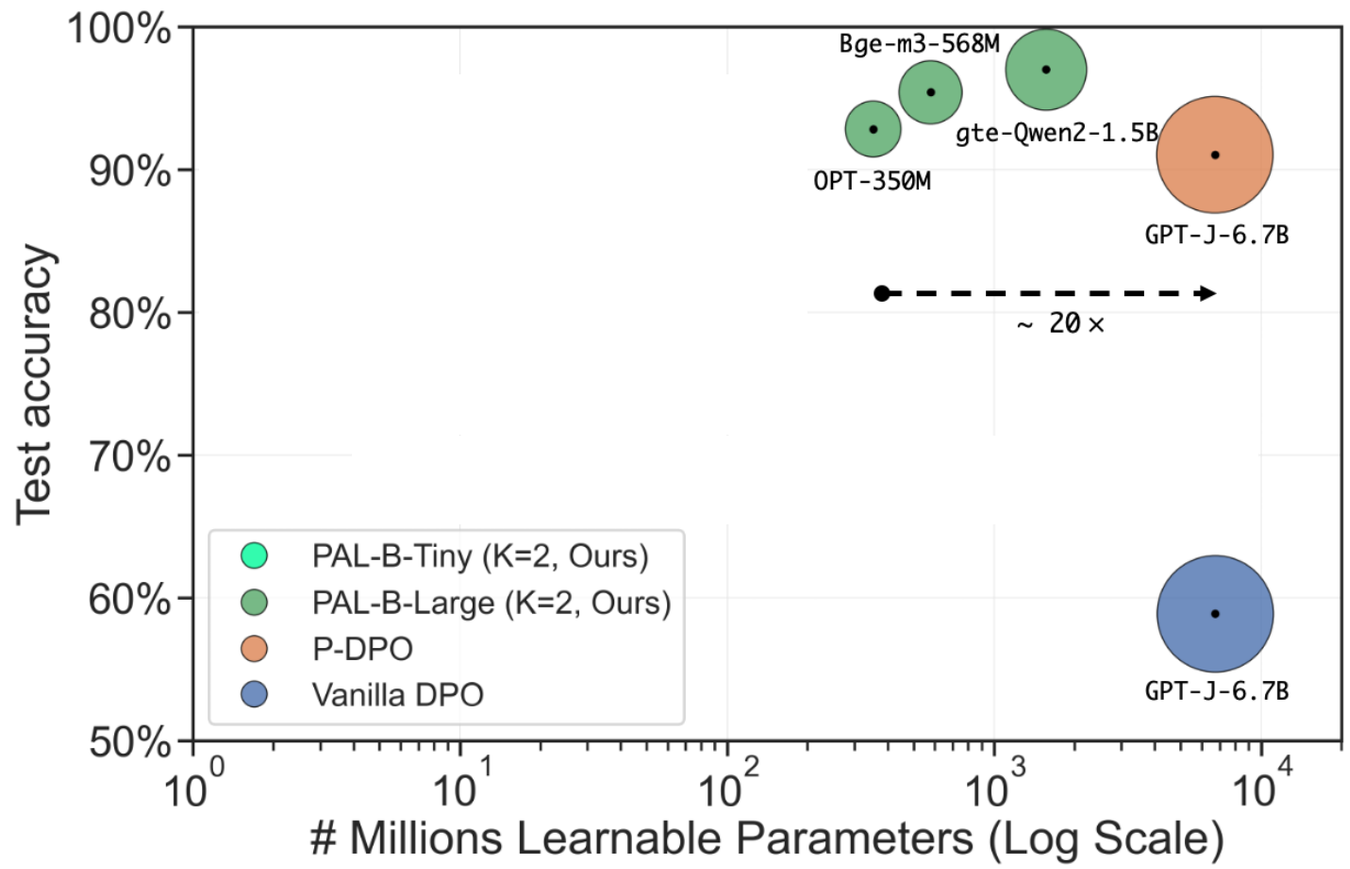
Model	Seen Acc (%)
Vanilla DPO	58.91
P-DPO Individual	91.04
P-DPO Cluster ($K = 5$)	91.12
`PAL-B-Tiny` ($K = 2$)	79.54 ± 0.54
`PAL-B-Large` ($K = 2$)	$\mathbf{92.82 \pm 0.95}$

$$r_\theta^{(i)}(\mathbf{x}; \mathbf{x}_c) = \langle f(\mathbf{x}), z^{(i)}(\mathbf{x}_c) \rangle \qquad \mathbf{x}, \mathbf{x}_c \in \mathbb{R}^D \qquad f, z^{(i)} : \mathbb{R}^D \to \mathbb{R}^d$$

$$z^{(i)}(\mathbf{x}_c) = \sum_{k=1}^{K} w_k^{(i)} g_k(\mathbf{x}_c) \qquad w_k^{(i)} \geq 0 \qquad \sum_{k=1}^{K} w_k^{(i)} = 1$$

Li et. al. uses supervised fine tuning on a *6B-parameter language model*

We train 2-layer MLP (~600K parameters) + a simpler language model for representations (with 350M parameters, OPT-350)



D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, "PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment", ICLR 2025
https://openreview.net/pdf?id=1kFDrYCuSu

Li, X., Lipton, Z. C., and Leqi, L. Personalized language modeling from personalized human feedback. arXiv preprint, 2024 arXiv:2402.05133

Experiments: Summary dataset

Language generation

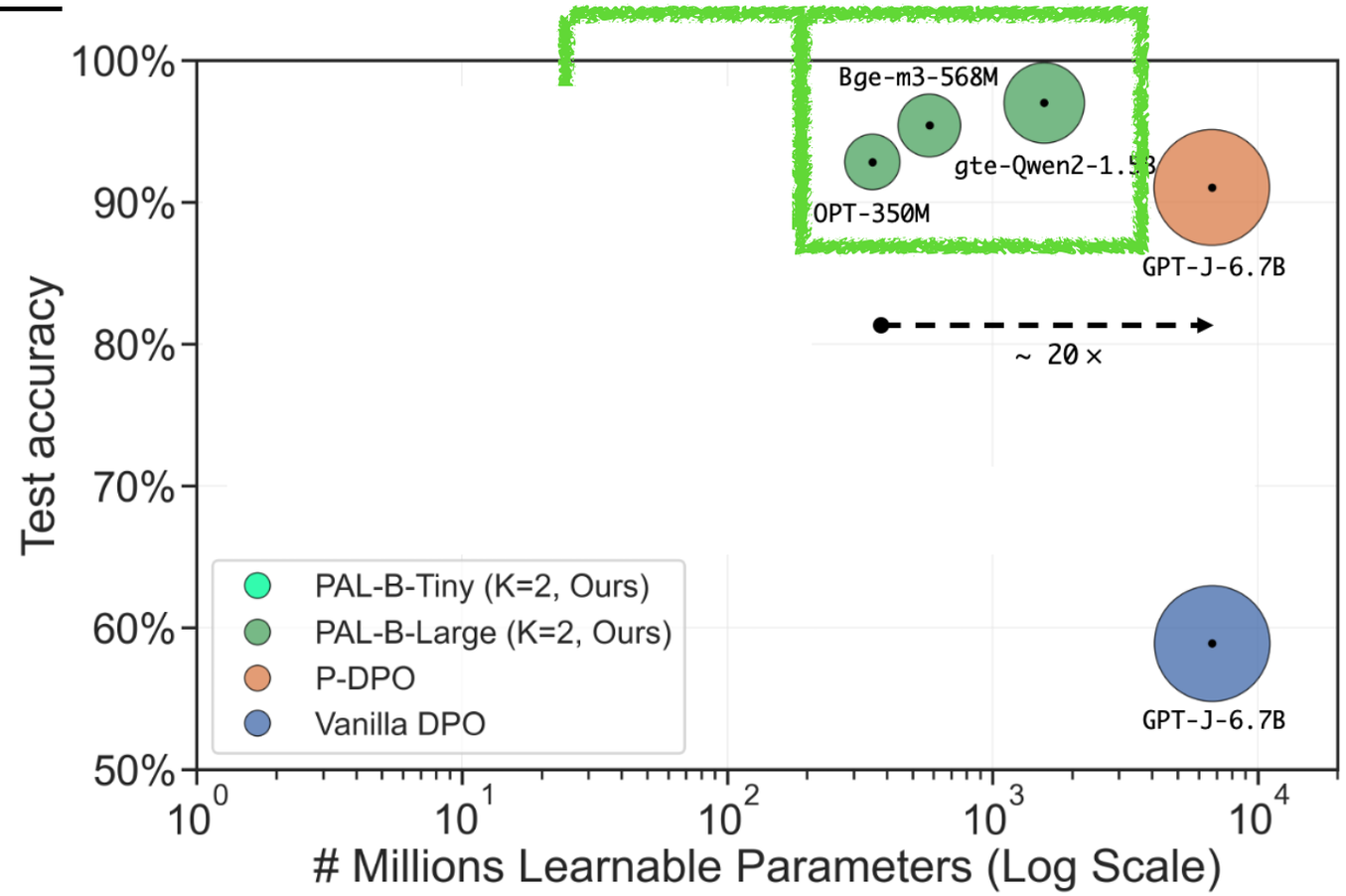
Model	Seen Acc (%)
Vanilla DPO	58.91
P-DPO Individual	91.04
P-DPO Cluster ($K = 5$)	91.12
`PAL-B-Tiny` ($K = 2$)	79.54 ± 0.54
`PAL-B-Large` ($K = 2$)	$\mathbf{92.82 \pm 0.95}$

$$r_\theta^{(i)}(\mathbf{x}; \mathbf{x}_c) = \langle f(\mathbf{x}), z^{(i)}(\mathbf{x}_c) \rangle \qquad \mathbf{x}, \mathbf{x}_c \in \mathbb{R}^D \qquad f, z^{(i)} : \mathbb{R}^D \to \mathbb{R}^d$$

$$z^{(i)}(\mathbf{x}_c) = \sum_{k=1}^{K} w_k^{(i)} g_k(\mathbf{x}_c) \qquad w_k^{(i)} \geq 0 \qquad \sum_{k=1}^{K} w_k^{(i)} = 1$$

Li et. al. uses supervised fine tuning on a *6B-parameter language model*

We train 2-layer MLP (~600K parameters) + a simpler language model for representations (with 350M parameters, OPT-350)



D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, “PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment”, ICLR 2025
https://openreview.net/pdf?id=1kFDrYCuSu

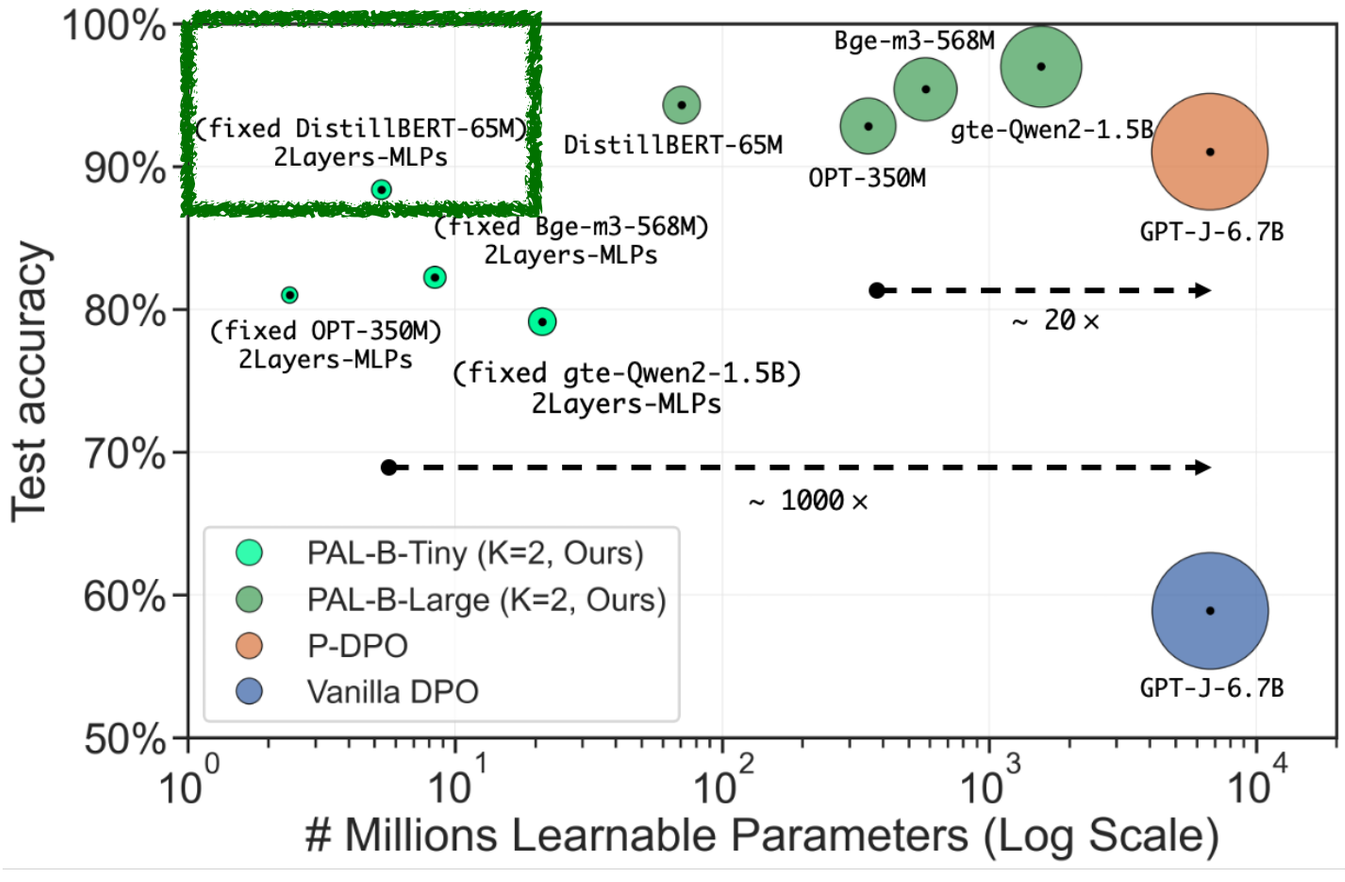
Li, X., Lipton, Z. C., and Leqi, L. Personalized language modeling from personalized human feedback. arXiv preprint, 2024 arXiv:2402.05133

Experiments: Summary dataset

Language generation

Model	Seen Acc (%)
Vanilla DPO	58.91
P-DPO Individual	91.04
P-DPO Cluster ($K = 5$)	91.12
`PAL-B-Tiny` ($K = 2$)	79.54 ± 0.54
`PAL-B-Large` ($K = 2$)	$\mathbf{92.82 \pm 0.95}$

With Fixed DistillBERT-65M, just a 2-layer MLP gets to nearly 90% accuracy!



D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, “PAL: Sample-Efficient Personalized Reward Models for Pluralistic Alignment”, ICLR 2025
https://openreview.net/pdf?id=1kFDrYCuSu

Li, X., Lipton, Z. C., and Leqi, L. Personalized language modeling from personalized human feedback. arXiv preprint, 2024 arXiv:2402.05133

Experiments: Summary dataset

Language generation

Model	Seen Acc (%)
Vanilla DPO	58.91
P-DPO Individual	91.04
P-DPO Cluster ($K = 5$)	91.12
`PAL-B-Tiny` ($K = 2$)	79.54 ± 0.54
`PAL-B-Large` ($K = 2$)	$\mathbf{92.82 \pm 0.95}$

Generalize to new users via few shot learning

Unseen user generalization

Methods	Test accuracy (Unseen user)
P-DPO individual	55.34%
P-DPO K=5	54.55%
Vanilla DPO	55.37%
PAL-B K=2 N=10	$91.63\% \pm 1.43\%$
PAL-B K=2 N=20	$91.72\% \pm 1.40\%$
PAL-B K=2 N=50	$92.02\% \pm 1.10\%$
PAL-B K=2 N=100	$91.97\% \pm 1.91\%$

Li, X., Lipton, Z. C., and Leqi, L. Personalized language modeling from personalized human feedback. arXiv preprint, 2024 arXiv:2402.05133

Advantages of our modeling approach

- **Sample-Efficient**: can be learned with few samples per user
- **Modular design:** leverages commonalities while adapting to individual preferences,
- **Versatile:** can be applied to different modalities, e.g., text-to-image, text-to-text, classical recommendation systems
- **Few-shot generalization:** can be adapted to new users with few samples
- **High-performance with fewer parameters:** we obtain SoTA results with 20x to 100x few parameters

Outline

- Motivation: Why should we care about heterogeneity in human preferences?
- Reward modeling for preference learning
- Personalizable reward modeling
- Learning metric and multiple user preferences simultaneously
- Sample efficiency by leveraging commonalities:
  Personalizable reward modeling for pluralistic alignment (PAL)
- Experimental results
- Future directions

Generated by ChatGPT

Limitations of current datasets

Currently available open datasets for alignment or reward learning suffer from severe limitations for learning under heterogeneity

- Many datasets do not contain anonymized ids
- Collected using a clear rubric on how to evaluate the outputs: so these are crowdsourcing the researchers preference rather than collecting people's preferences
- Highly imbalanced in terms of number of examples per user:
  A small subset of users provide a lot of the comparison data while a large set of people provide very few comparisons

Need better datasets and benchmarks!

Future directions

- More general approaches for reward modeling

$$r_\theta^{(i)}(\mathbf{x}, \mathbf{x}_c) = \sum_{l=1}^{K} w_l^{(i)} r_{\theta_l}(\mathbf{x}, \mathbf{x}_c)$$

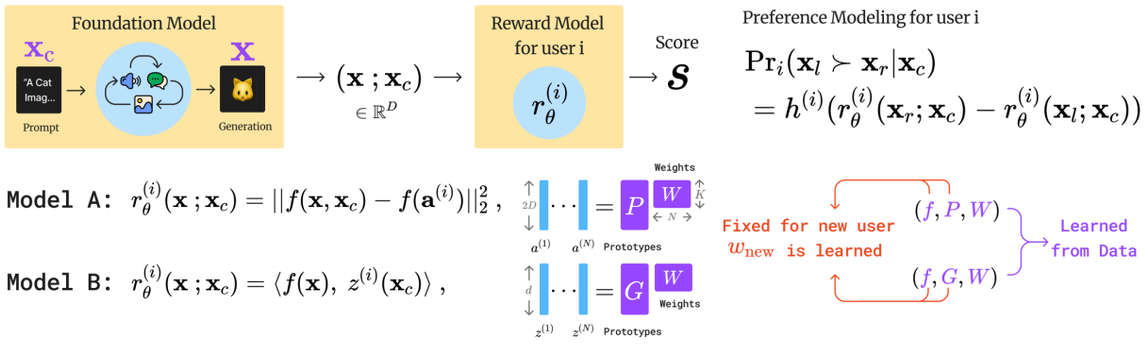
Chen et al. 2024, Ponder et al. 2024, Wang et al. 2024, Bose et al. 2025, …..

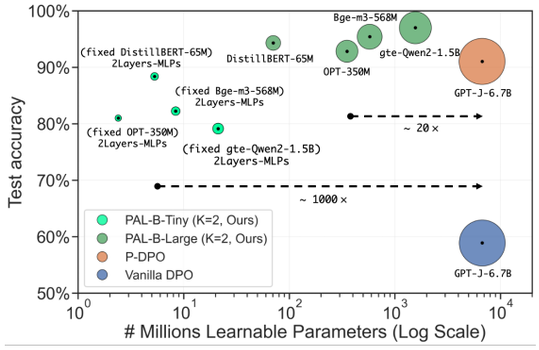
PAL-B

$$r_\theta^{(i)}(\mathbf{x}, \mathbf{x}_c) = \sum_{l=1}^{K} w_l^{(i)} \langle f(\mathbf{x}), g_l(\mathbf{x}_c) \rangle$$

- Need better datasets and benchmarks!
- Adapting on the fly
  - Z. Wang, C. Zhang, R. K. Vinayak “Bridging Lifelong and Multi-Task Representation Learning via Algorithm and Complexity Measure”, Algorithmic Learning Theory (ALT) 2026
- Need better understanding of generalization
  - OOD generalization for users
  - OOD generalization for prompts
- Global safety constraints + Personal preferences
- Conversations, dialogues and interactive collaborations with AI

Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment





- Modular
- Versatile
- Sample-Efficient

Thank you!

Questions?

ramya@ece.wisc.edu

@ramyavinayak

https://github.com/RamyaLab/pluralistic-alignment

- D. Chen, Y. Chen, A. Rege, Z. Wang, R. K. Vinayak, “PAL: Sample-Efficient Personalized Reward Modeling for Pluralistic Alignment”, ICLR 2025
- Canal, Mason, Vinayak, Nowak, “One for all: Simultaneous metric and preference learning over multiple users”, Neurips 2022
- Wang, So, Vinayak, “Metric Learning from Limited Pairwise Preference Comparisons”, UAI 2024, arXiv:2403.19629, 2024.
- G. Tatli, Y. Chen, R. K. Vinayak “Learning Population of Preferences via Pairwise Comparison Queries”, AISTATS 2024
- Z. Wang, C. Chen, R. K. Vinayak “Bridging Lifelong Learning and Multitask Representation Learning via Algorithmic and Complexity Measure”, ALT 2026