Interact with paper

UNIVERSITY OF MINNESOTA

IBM

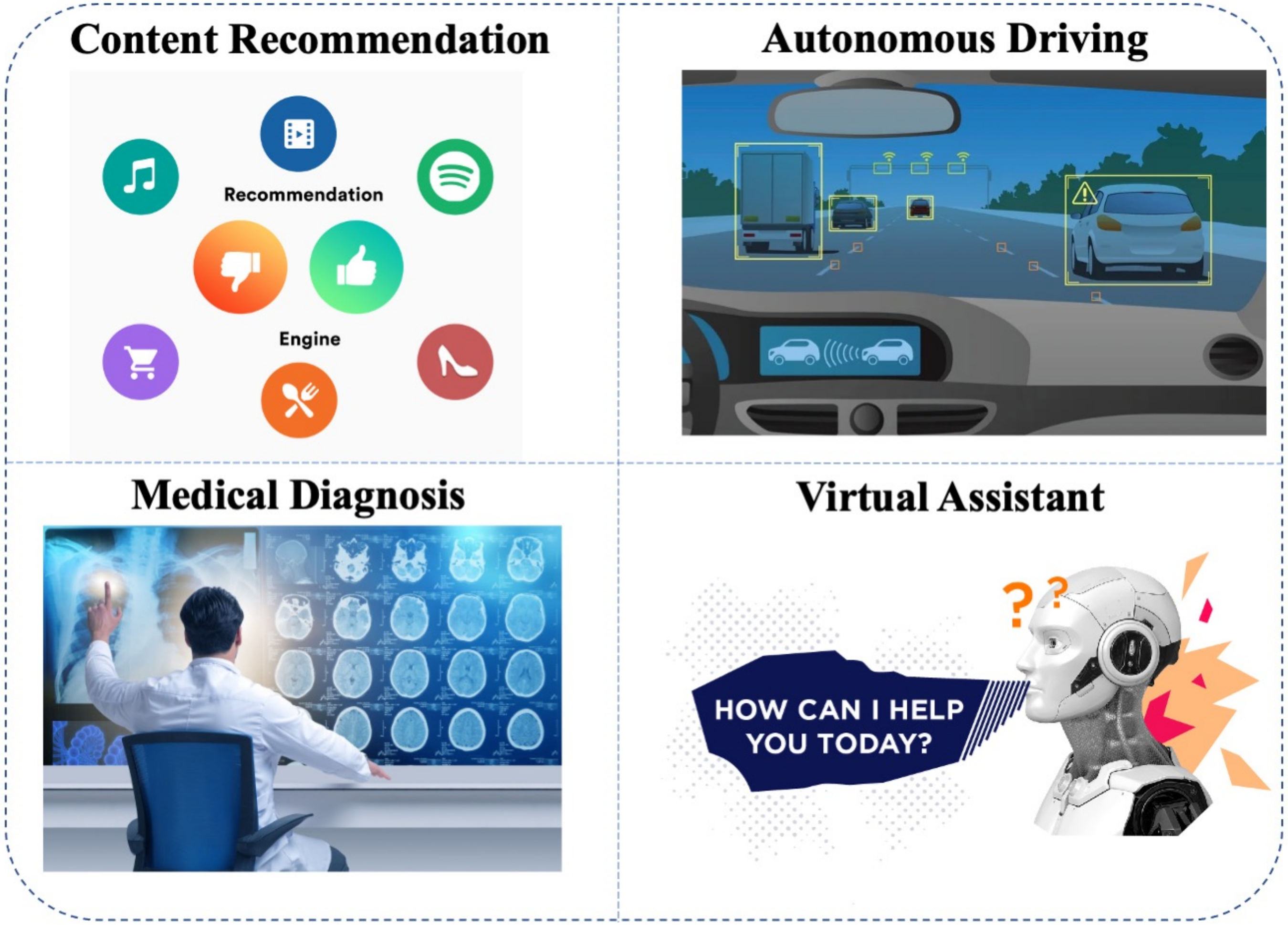
MAP: Multi-Human-Value Alignment Palette

Xinran Wang, *Qi Le, Ammar Ahmed, Enmao Diao, Yi Zhou, Nathalie Baracaldo, Jie Ding, Ali Anwar*

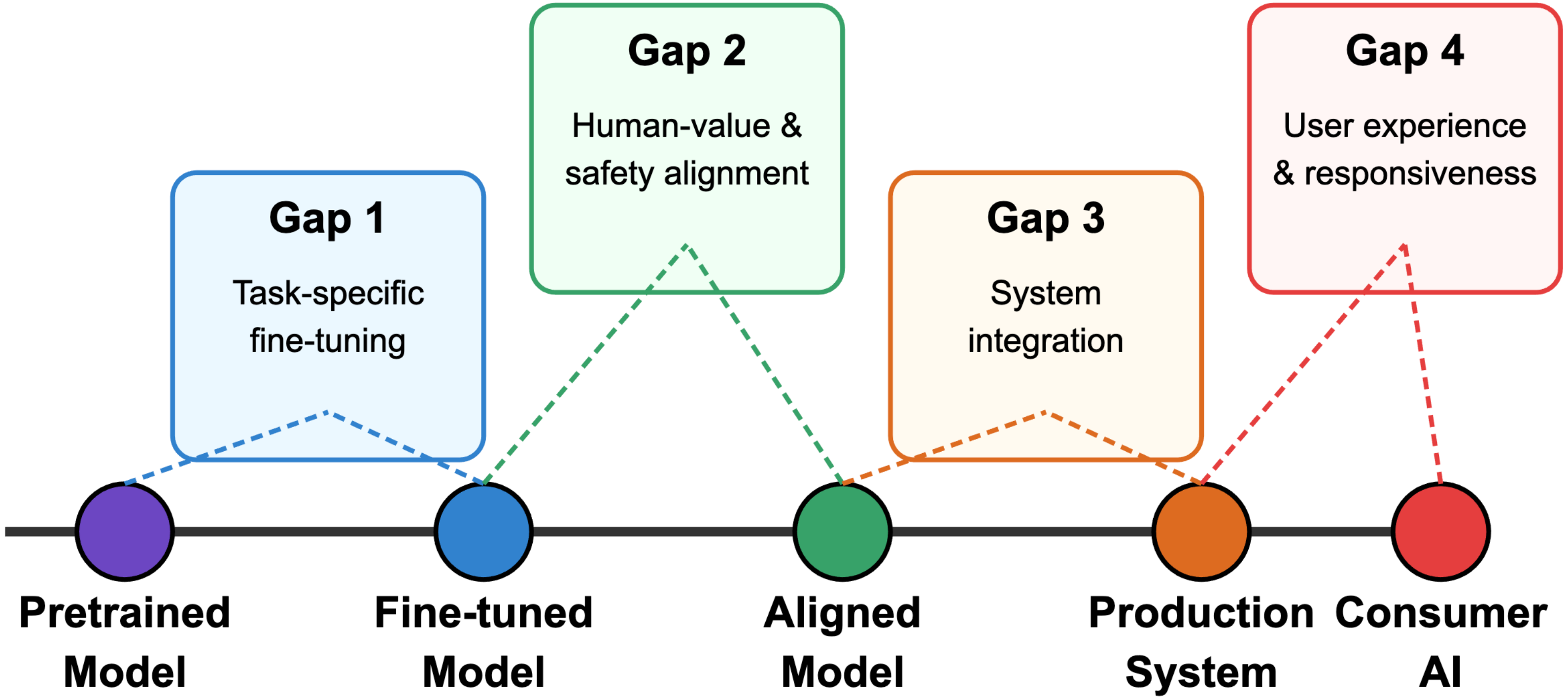
ICLR 2025 Oral Presentation

NSF 3M SAMSUNG MorphMind

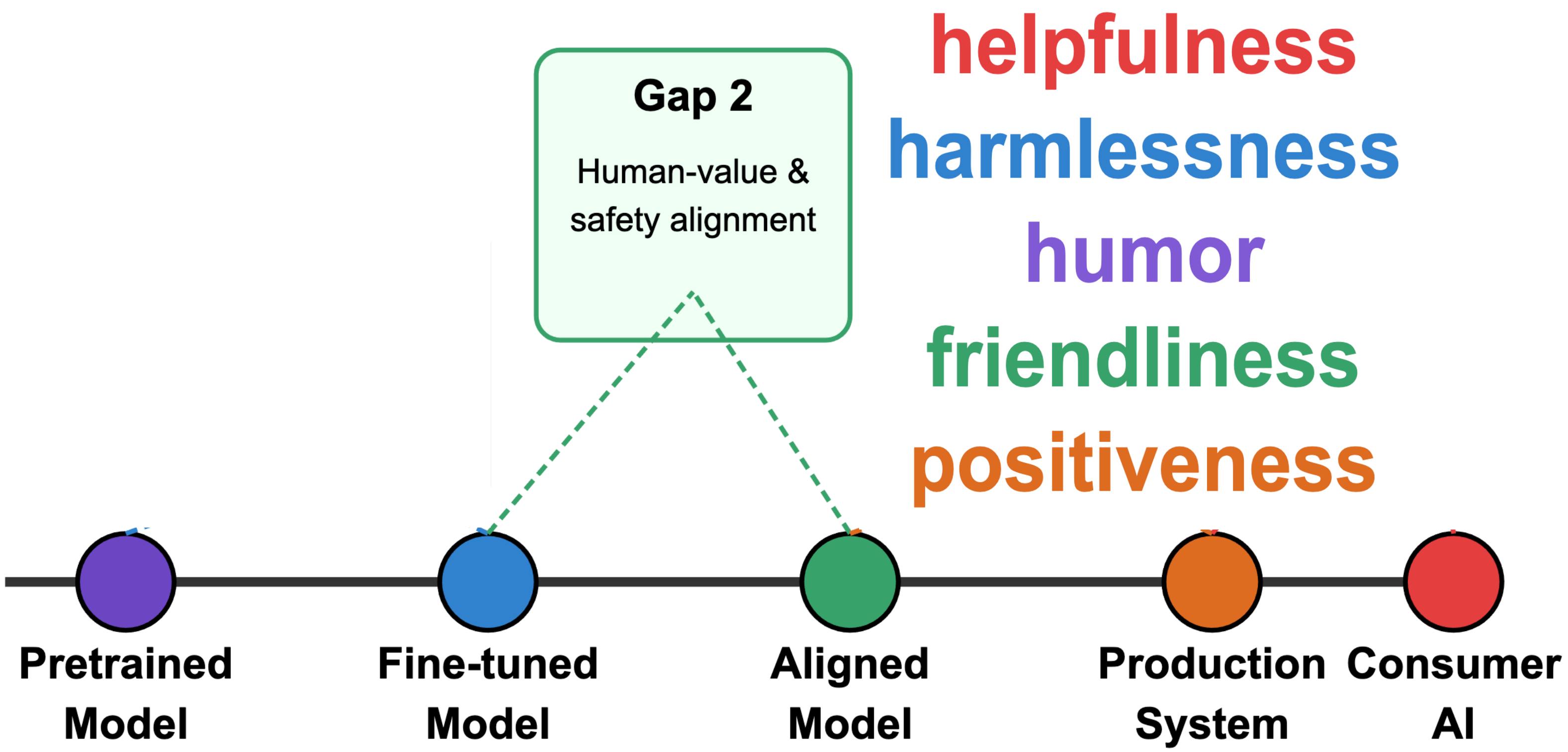
LLMs are applied to various applications



Gaps exist between pretrained models and consumer AI



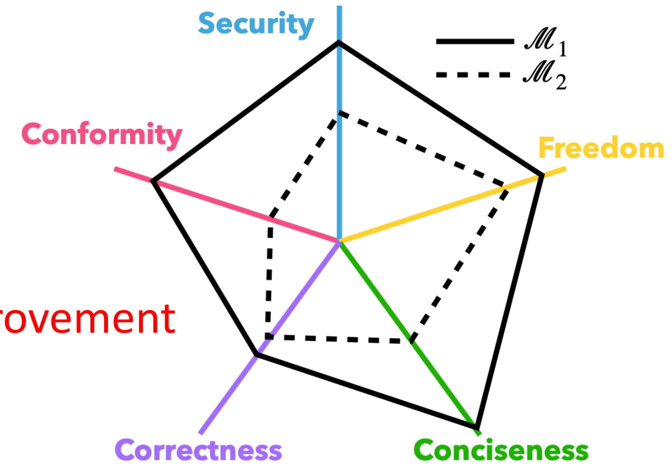
Gaps exist between pretrained models and consumer AI



Human preferences are multi-dimensional

- Users have various priorities
- Same user may value different things in different contexts
- Values can conflict in ways that require nuanced balancing

Hope:
Pareto improvement



Reality:
Tradeoffs

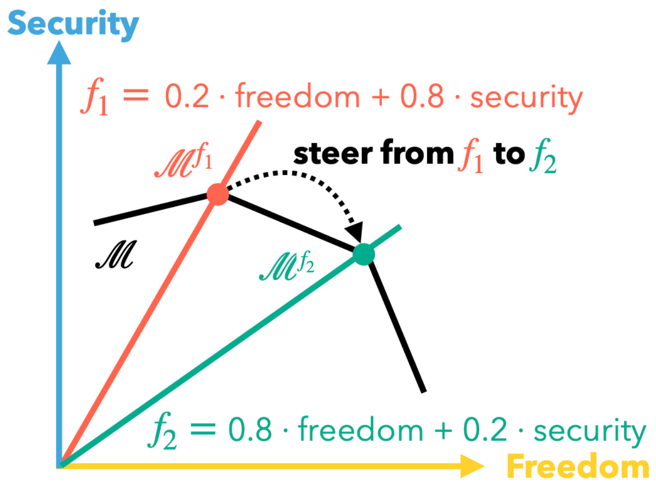


Image Source: Sorensen et al. "Position: A Roadmap to Pluralistic Alignment." ICML 2024.

Key challenges in multi-human-value alignment

1: How would one know there is still **room for Pareto improvement**, and how to approach it?

2: How to set the objective to cater **to personalized priorities** among all the values?

3: Is there a "magical" way to better account for multiple human values (in terms of reward models) than **linear combinations**?

Contributions of MAP

1: How would one know there is still room for Pareto improvement, and how to approach it?

MAP provides efficient navigation to Pareto Frontier

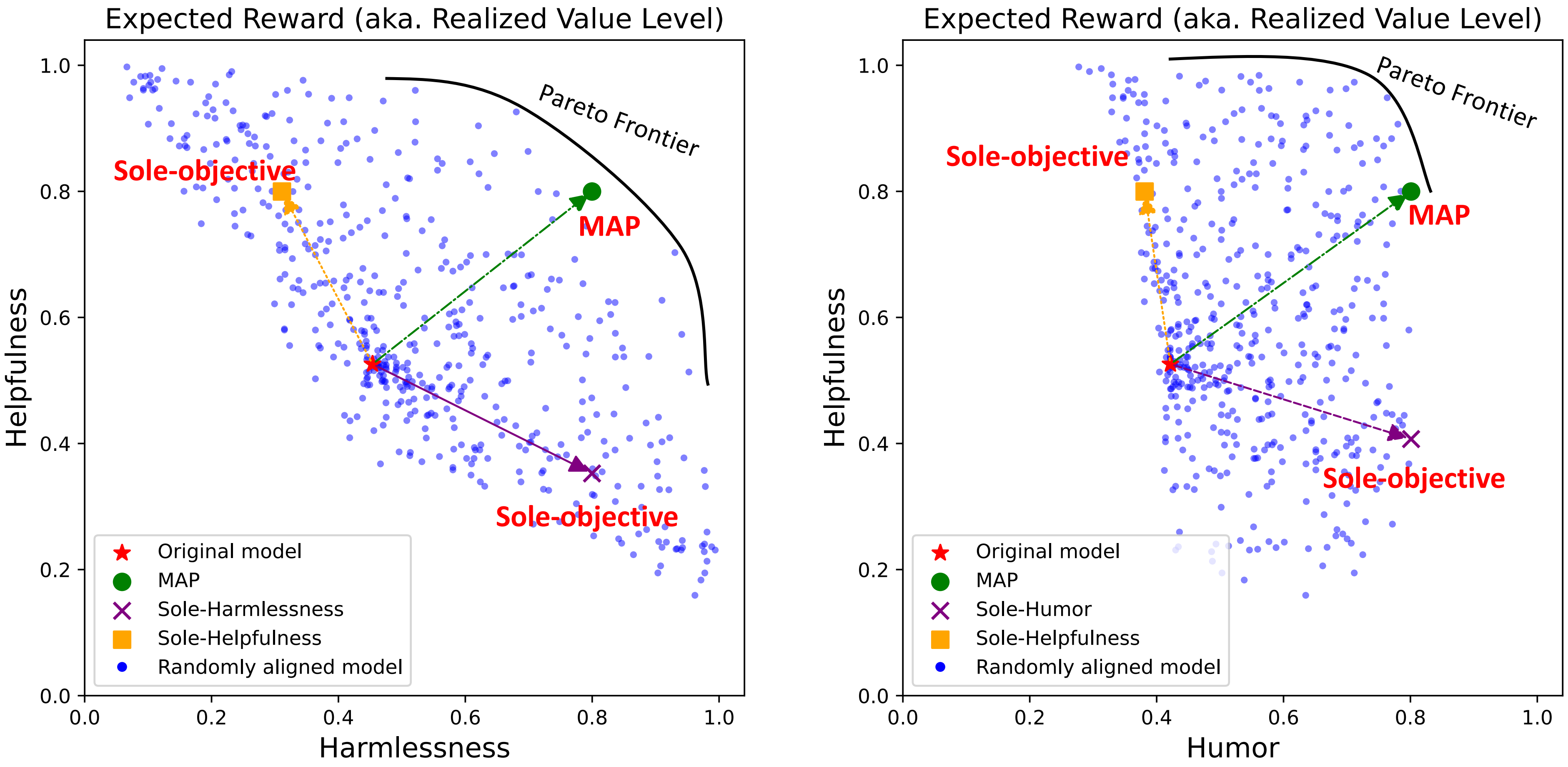
2: How to set the objective to cater to personalized priorities among all the values?

MAP enables precise control of value improvement

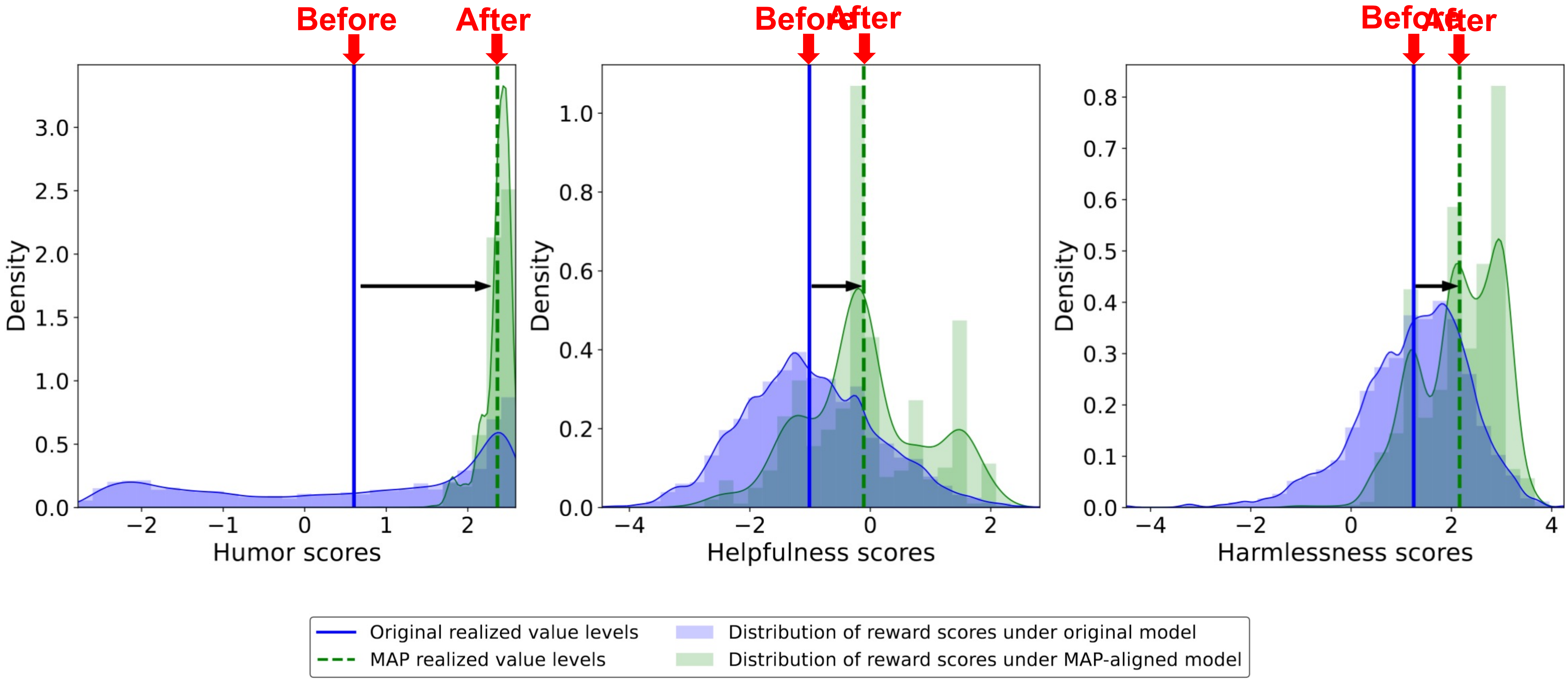
3: Is there a “magical” way to better account for multiple human values (in terms of reward models) than linear combinations?

We proved that linear combinations of individual reward functions can sufficiently capture the entire Pareto Frontier

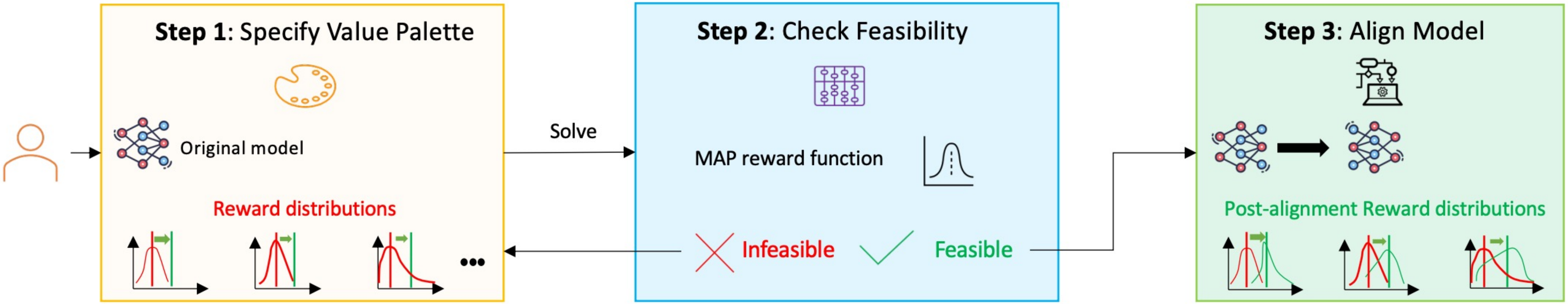
MAP provides efficient navigation to Pareto Frontier



MAP enables precise control of value improvement



MAP Procedure Overview



Step 1: Formulate the multi-value alignment problem – following the first principle

Step 2: Assess trade-offs among values explicitly and quantitatively

Step 3: Apply the obtained reward function to any favorite optimization approach

MAP problem formulation is grounded in first principles

MAP problem: seek a generative distribution that minimizes the KL-divergence from the base model subject to user-defined alignment targets

r_i: reward function represents the i^{th} value

$$\boxed{min_p \mathrm{E}_{y|x\sim p}\{D_{KL}(p, p_0)\} \quad s.t. \quad \mathrm{E}_{\mathrm{y|x}\sim p}\, r_i(x, y) \geq c_i \quad \forall i = 1, \ldots, m}$$

$c_1, \ldots, c_m$: user-defined “value palette”

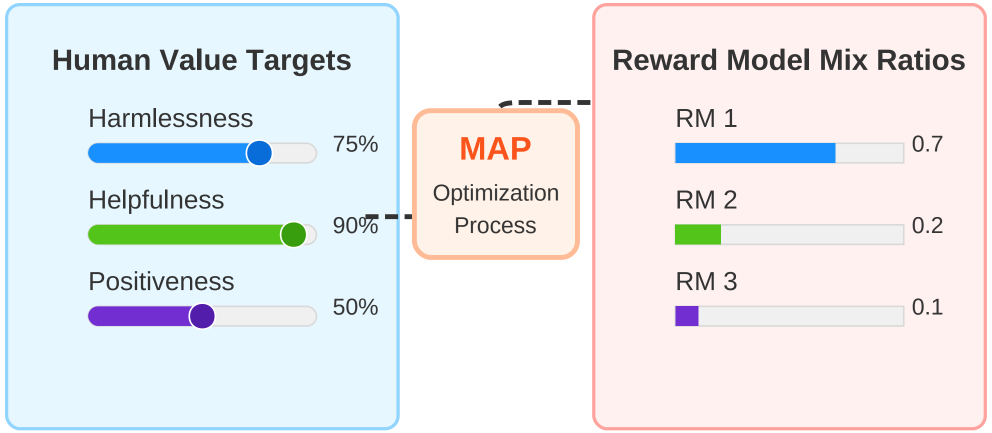
MAP establishes a 1-to-1 map from value palette to reward

MAP problem: $min_p \mathrm{E}_{y|x\sim p}\{D_{KL}(p, p_0)\} \quad s.t. \quad \mathrm{E}_{\mathrm{y|x}\sim p}\, r_i(x,y) \geq c_i \quad \forall i = 1, \ldots, m$

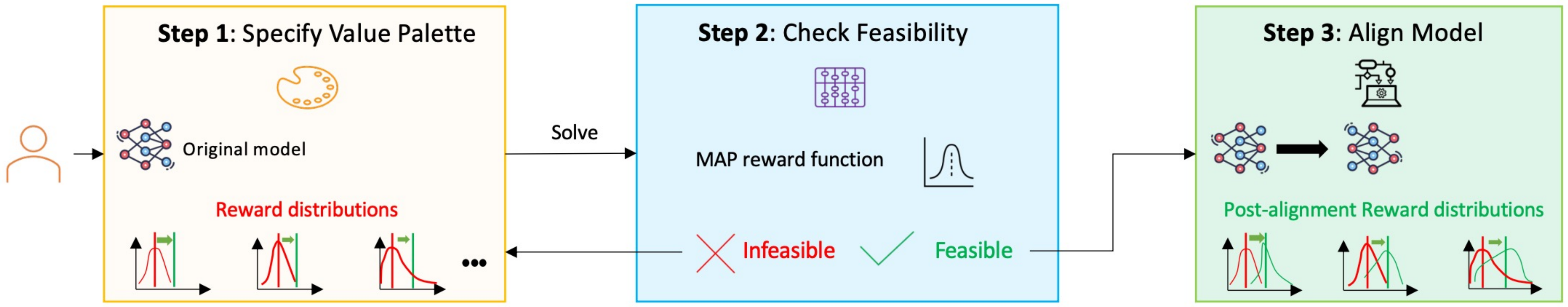
Close form solution

Theorem 1 (informal): The solution to MAP is represented as $p_{\lambda}(\, y \mid x\,) = Z(\boldsymbol{\lambda})^{-1} p_0(\, y \mid x\,)\, e^{\lambda^T r(x,y)}$, where $\boldsymbol{\lambda} \in R^m$ is the solution to the problem: $\max_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}) = -\log Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^T \boldsymbol{c}$, and $Z(\boldsymbol{\lambda}) = \mathrm{E}_{y|x\sim p_0} e^{\boldsymbol{\lambda}^T \boldsymbol{r}(x,y)}$

1-to-1 **map**ping relationship between c and λ



MAP: Multi-Human-Value Alignment Palette



Step 2: Assess trade-offs among values explicitly and quantitatively

Recall Theorem 1: $\boldsymbol{\lambda} \in R^m$ is the solution to the problem: $\max_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}) = -\log \mathbb{E}_{y|x\sim p_0} e^{\boldsymbol{\lambda}^T \boldsymbol{r}(x,y)} + \boldsymbol{\lambda}^T \boldsymbol{c}$

Easily approximated by Monte Carlo samples from p_0

$$\max_{\boldsymbol{\lambda}} g_n(\boldsymbol{\lambda}) = -\log \frac{1}{n} \sum_{i=1,\ldots,m} e^{\boldsymbol{\lambda}^T \boldsymbol{r}(x_i,y_i)} + \boldsymbol{\lambda}^T \boldsymbol{c}$$

Concave objective

Key Theoretical Insights

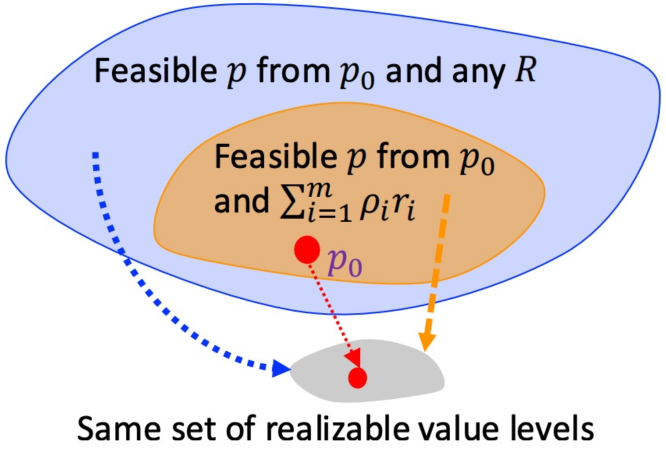
MAP and RLHF problems share the same reachable value space

Theorem 2 (informal): The realizable value levels of the MAP problem is the same as the RLHF problem with ANY reward model.

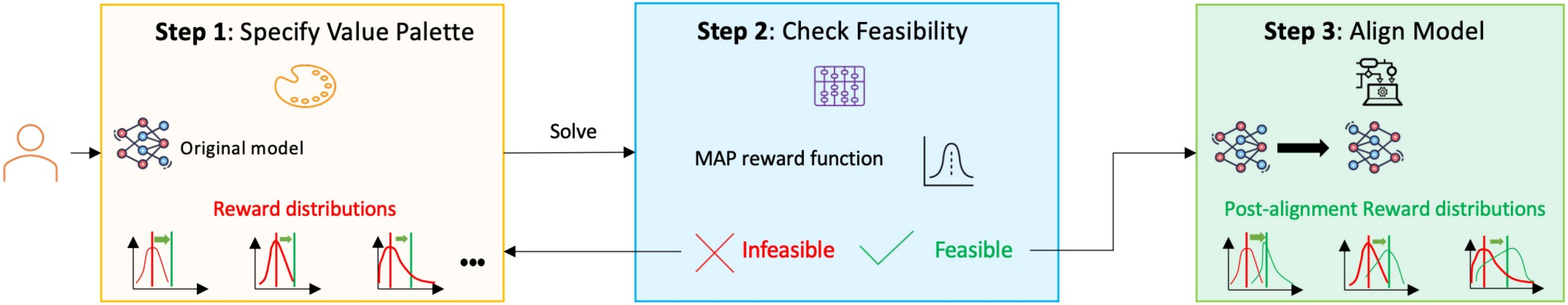
Realizable value level: $\mathrm{E}_{\mathrm{y}|\mathrm{x} \sim p}\, \boldsymbol{r}(x, y)$

Linear combination of individual rewards is good enough

Theorem 3 (informal): Linear combinations of individual reward functions can sufficiently capture the entire Pareto Frontier.



MAP can be deployed via decoding or finetuning stage



Step 3: Apply the obtained reward function to any favorite optimization approach

Using the optimized $\boldsymbol{\lambda}$:

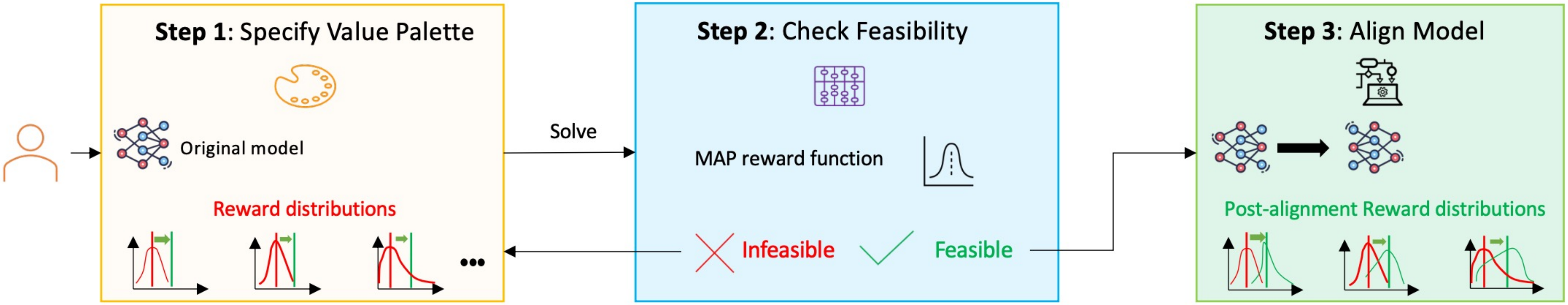
- Construct a single reward function:

$$R(x, y) = \boldsymbol{\lambda}^T \boldsymbol{r}(x, y) = \lambda_1 r_1(x, y) + \cdots \lambda_m r_m(x, y)$$

- Derive the aligned model via exponential tilting:

$$p_\lambda(y \mid x) \propto p_0(y \mid x)\, e^{R(x,y)}$$

MAP can be deployed via decoding or finetuning stage



Deployment Options

1. Decoding-Time Alignment

- Apply MAP during inference by resampling outputs with exponential weights

$$p_{\lambda}(y \mid x) \propto p_0(y \mid x)\, e^{R(x,y)}$$

- Suitable for quick deployment without model retraining

2. Finetuning with PPO

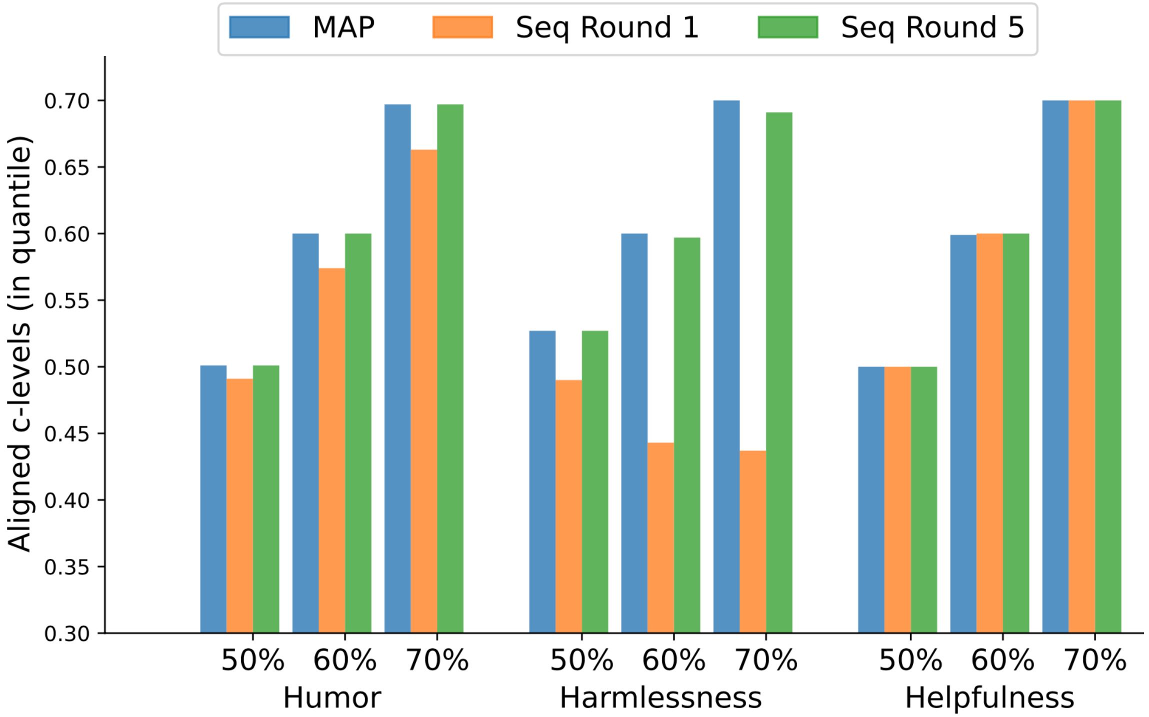
- Use $R(x,y) = \boldsymbol{\lambda}^T \boldsymbol{r}(x,y)$ as reward signal in PPO
- Suitable for large-scale deployment for better inference speed

Simultaneous vs. Sequential Alignment

Step 3: Apply the obtained reward function to any favorite optimization approach

What if the GPU memory is not sufficient for loading too many reward models!?

Theorem 5 (informal): Sequentially aligning a model using single-value MAP objectives in a cyclical manner will converge to the same solution as simultaneously optimizing for all values in the MAP framework.



Align OPT-1.3B with multi- vs. single-value palettes

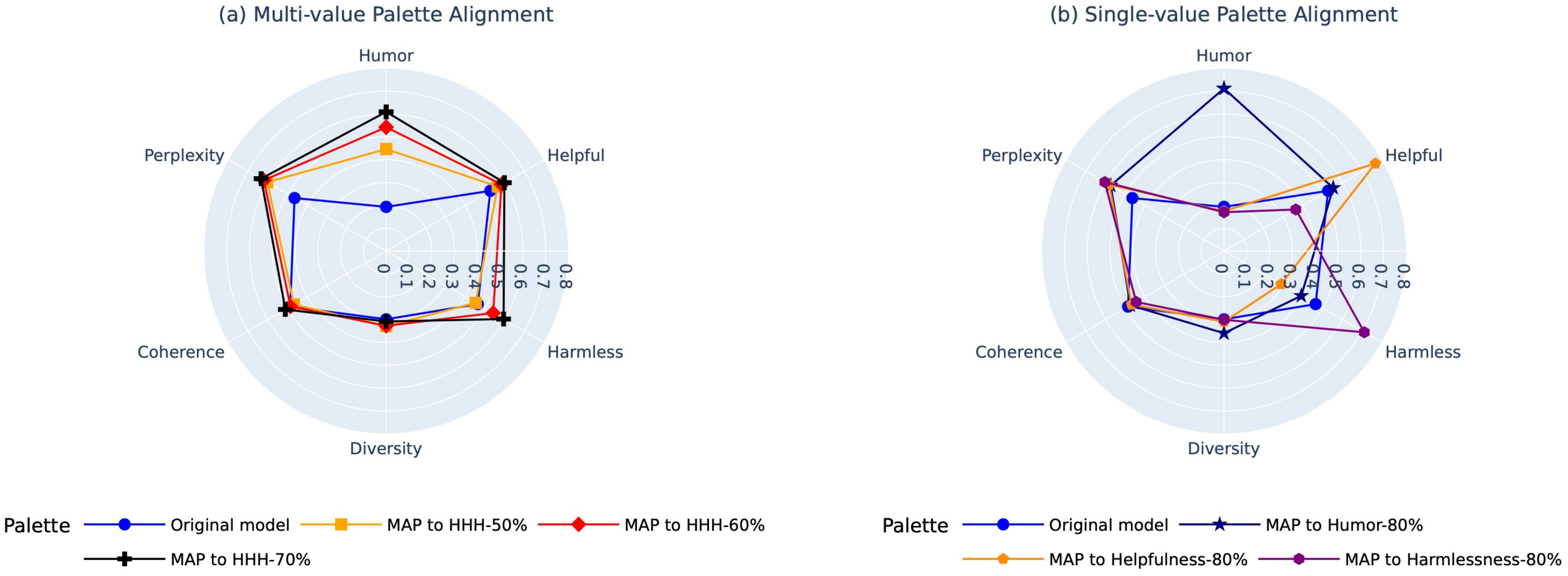
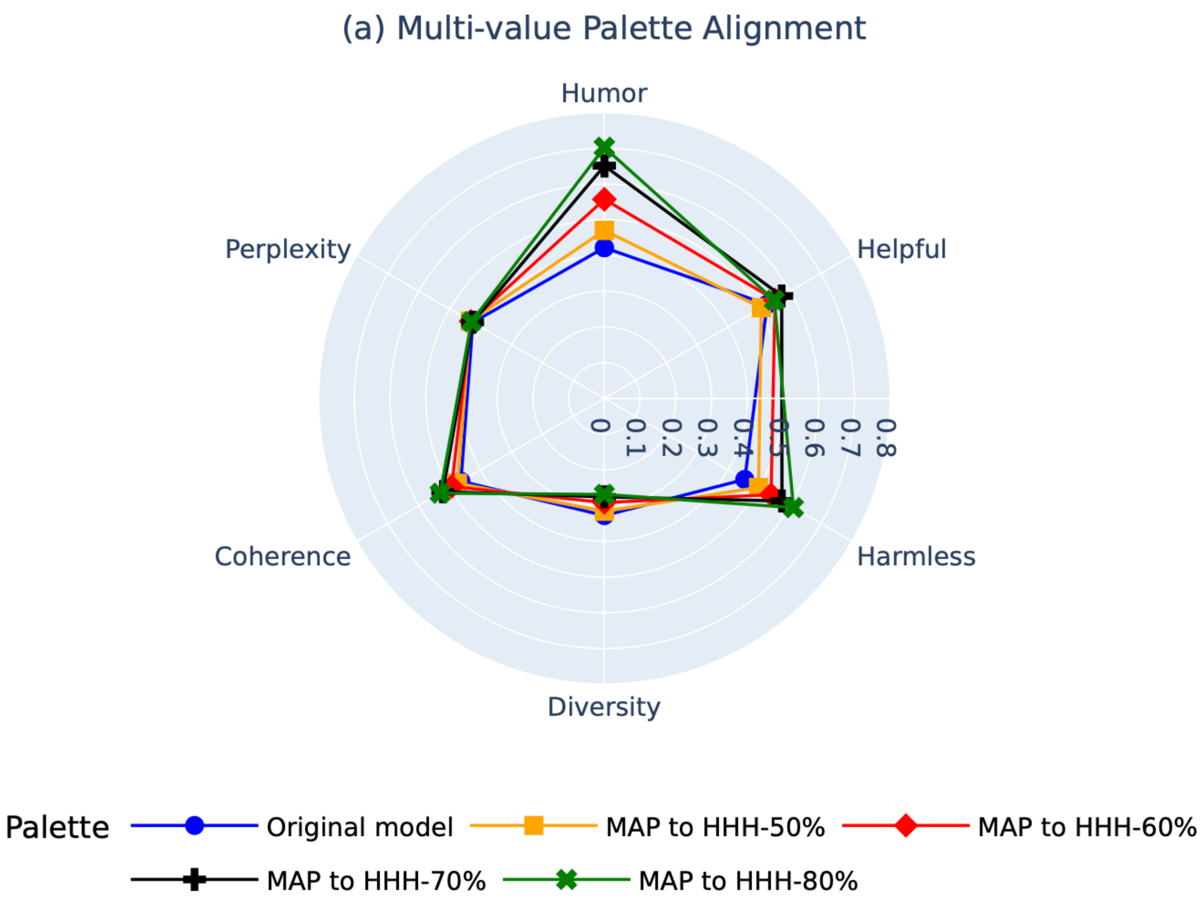
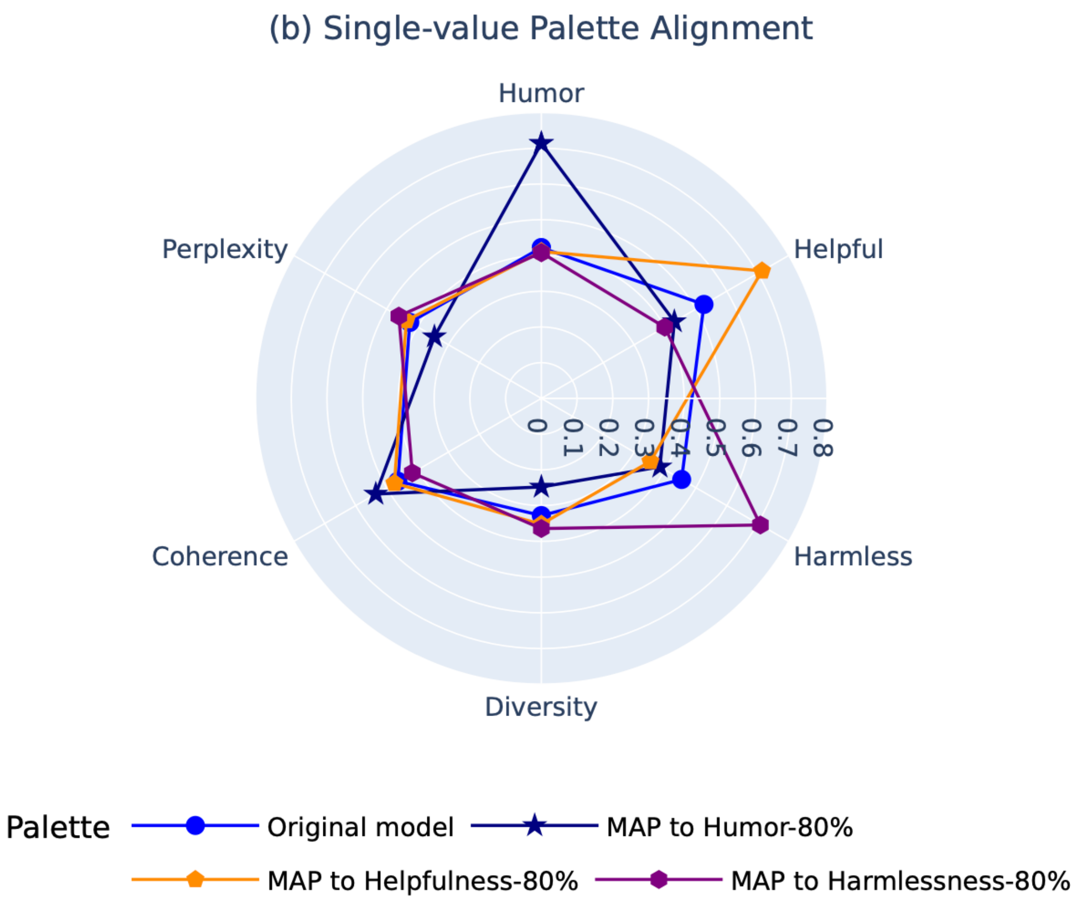


Figure 6: Radar plots showing the alignment of OPT-1.3B with (a) multi-value palettes given by 50%, 60%, and 70% quantiles of the original model's reward distributions, and (b) single-value palettes at the 80% quantile.

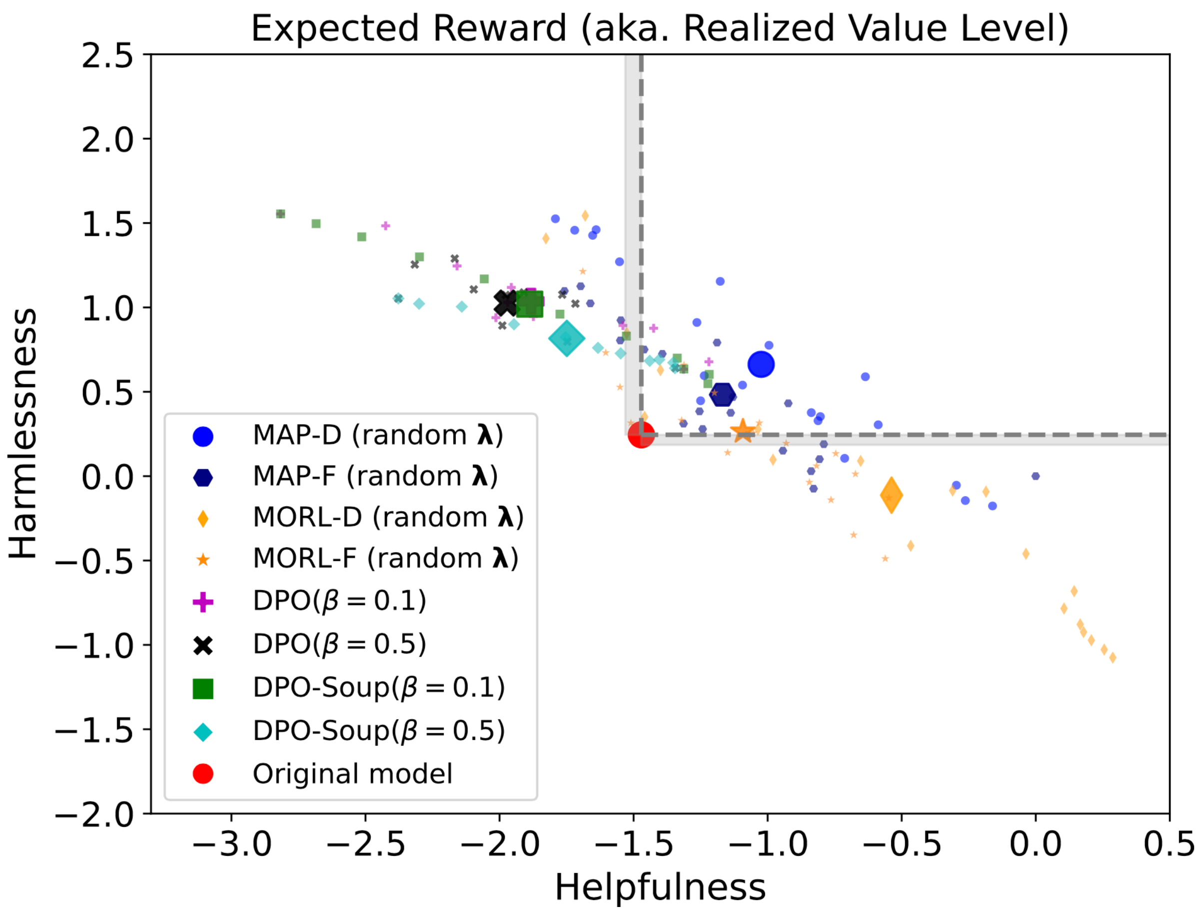
Align Llama2-7B-chat with multi- vs. single-value palettes

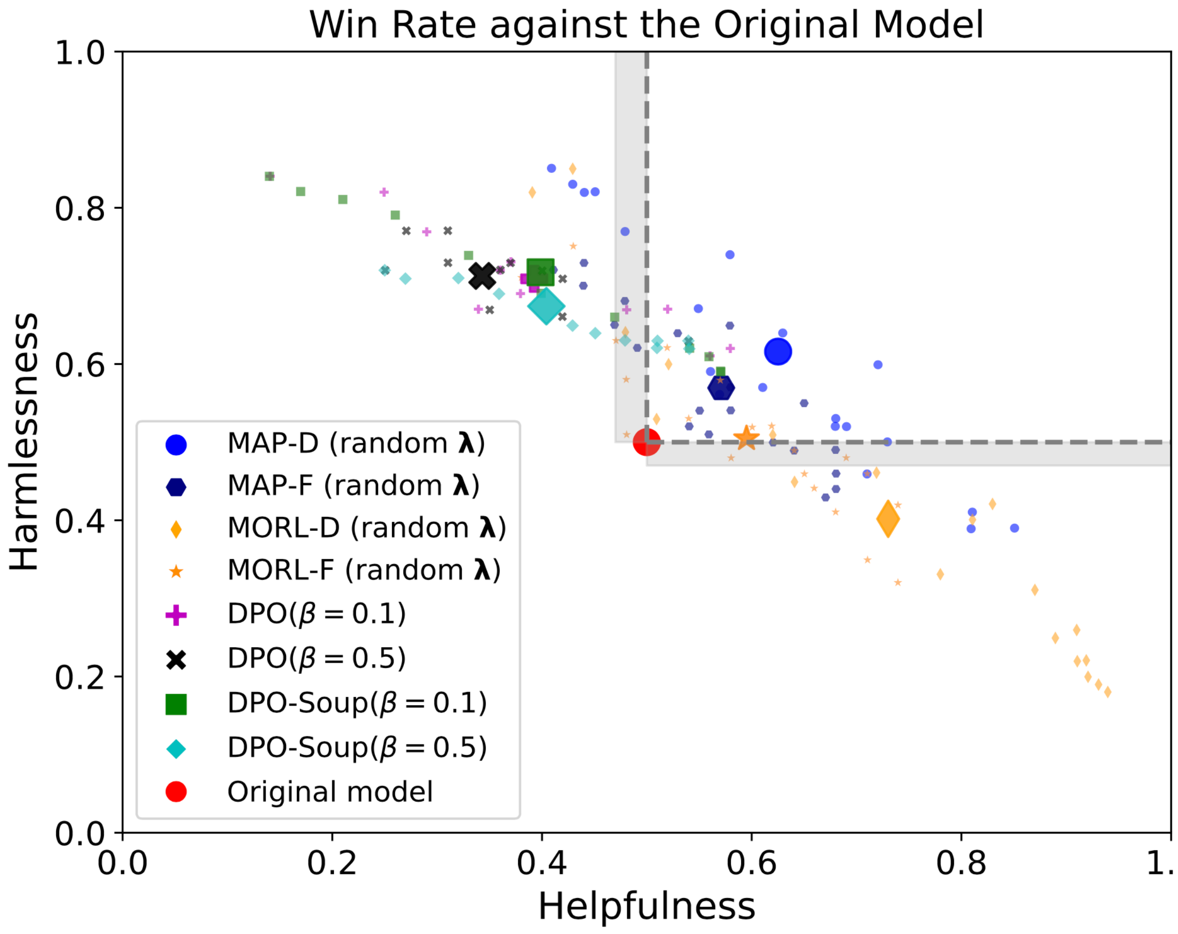
Llama2-7B-chat model, which has a larger complexity than OPT-1.3B, allows for more extensive multi-value alignment





Compared with SOTA: MAP more consistently falls in the desirable regime (upper-right quadrant)





MAP in Practice

Value Palette Specification

- Quantile-based palette:
  e.g., 80th percentile of p_0's helpfulness score distribution
- Classifier-based thresholds:
  by increasing the expected log-probability (or probability) under a classifier-based reward model, allowing interpretable improvements like a 20% boost in harmlessness, e.g., $c_1 = E_{y\sim p_0} r_1(x,y) + log(1+20\%)$
- Automatic adjustment:
  Interpolation adjustment: $c' = c - \rho(c - c_0)$, where $\rho \in (0,1]$ is iteratively tuned until feasible

Optimization

- Translate constraints to a MAP objective to solve
- Auto-check feasibility

User Interaction

- Visual interaction for continuous adjustment

Try this MAP interactive demo

Multi-Value Palette

Humor 0.720

-2.779 2.587

Helpfulness -0.980

-4.452 2.816

Harmlessness 1.290

-4.482 4.246

Diversity 0.850

0.062 0.969

Coherence 0.530

-0.157 0.945

Perplexity -1.360

-4.098 -0.555

Optimize

Reset

```
Now solving for lambda ...
This is feasible palette.
The solved lambda is lambda=[0.06, 0.07, 0.07, 0.08, 0.05, 0.06].
It took 0.35 seconds to solve.
Estimate the realized value levels after alignment ...
Realized value levels are: c_realized=[0.72, −0.98, 1.29, 0.85, 0.53, −1.36], updated on UI.
```

Try it:

https://research-demo.com

Takeaways

MAP enables principled multiple human values alignment.

1. Users define desired levels (value palette)

2. Automatically checks and adjusts targets if unrealizable

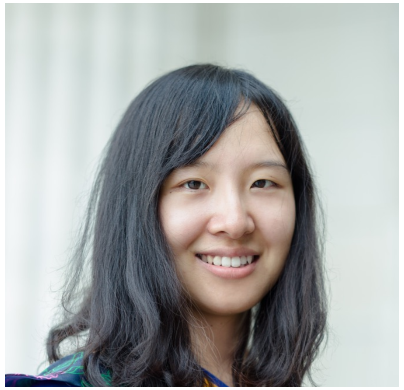
3. Proves linear reward combinations fully capture the Pareto Frontier

4. Compatible with decoding-time resampling or PPO finetuning

5. Supports sequential alignment when GPU memory is limited

Balancing human values—efficiently and rigorously.

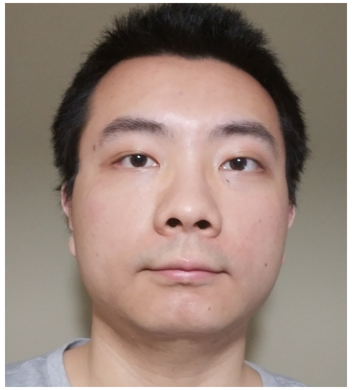
Thanks to the awesome team



Xinran Wang

Qi Le

Ammar Ahmed



Enmao Diao

Yi Zhou

Nathalie Baracaldo

Jie Ding

Ali Anwar

Thank You & Questions

We welcome questions, feedback, and potential collaboration opportunities

Paper

Interaction

Xinran Wang
Email: wang8740@umn.edu
Web: https://wang8740.github.io