

# Panacea: **P**areto **A**lignment via **a** Preference **A**daptation for LLMs

Yifan Zhong<sup>1,2\*</sup>, Chengdong Ma<sup>1\*</sup>, Xiaoyuan Zhang<sup>3\*</sup>, Ziran Yang<sup>4</sup>, Haojun Chen<sup>1</sup>, Qingfu Zhang<sup>3</sup>, Siyuan Qi<sup>2</sup>,  
Yaodong Yang<sup>1,✉</sup>

<sup>1</sup>IAI, Peking University

<sup>2</sup>BIGAI

<sup>3</sup>City University of Hong Kong

<sup>4</sup>Yuanpei College, Peking University



# | Contents

- Background & motivation
- Problem formulation
- Method
- Experiments
- Conclusion

# | Contents

- Background & motivation
- Problem formulation
- Method
- Experiments
- Conclusion

## | Background

- Existing alignment techniques utilize scalar human preference labels. i.e. “better”.
- They curate a dataset  $\{(x, y_1, y_2, z)\}$
- RLHF:

$$\begin{aligned} \max_{\theta} J_{\text{RLHF}}(\pi_{\theta}) &= \max_{\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(y|x)} [r(x, y)] \\ &\quad - \beta \mathbb{D}_{\text{KL}} [\pi_{\theta}(y|x) || \pi_{\text{ref}}(y|x)]. \end{aligned}$$

- DPO:

$$\begin{aligned} &\min_{\theta} \mathcal{L}_{\text{DPO}}(\pi_{\theta}) \\ &= \min_{\theta} -\mathbb{E} \left[ \log \sigma \left( \beta \log \frac{\pi_{\theta}(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_{\theta}(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right], \end{aligned}$$

# | Motivation

- Existing alignment techniques utilizes scalar human preference labels. i.e. “better”.
- But this is problematic!
- Let’s say you are buying T-shirts. One is **in fashion, looks good**, but is **expensive**; the other is **ordinary**, but is more **affordable**.
- For different people, they agree in each dimension, but their overall “better” label could conflict.
- **Same with LLMs.**

# Motivation

- Existing alignment techniques utilize scalar human preference labels. i.e. “better”.
- Two limitations:
  - Inconsistency and ambiguity -> misalignment
  - Optimization result is a single model -> fail to cover diverse human preferences

Preference dimension  $\mathcal{A}$  (e.g. helpful)  $\uparrow$

Preference dimension  $\mathcal{B}$  (e.g. harmless)  $\rightarrow$

Synthetic preference : better  $\nearrow$

Responses to label	Different labelers	Single-objective alignment			Our multi-dimensional alignment				
		Synthetic preference	Preferable response	Solution method : misaligned, conflicting, and singular	Prefer $\mathcal{A}$ much?	Preferable response	Prefer $\mathcal{B}$ much?	Preferable response	Solution method : aligned, coordinated, and diverse
Response ①			②		$\uparrow$	①	$\rightarrow$	②	
Prompt			①		$\uparrow$		$\rightarrow$		
Response ②			②		$\uparrow$		$\rightarrow$		

# | Contents

■ Background & motivation

■ **Problem formulation**

■ Method

■ Experiments

■ Conclusion

## | Problem formulation

- We formulate alignment as a multi-dimensional preference optimization (MDPO) problem.

$$\max_{\theta \in \Theta} \mathbf{J}(\pi_\theta) = (J_1(\pi_\theta), J_2(\pi_\theta), \dots, J_m(\pi_\theta)),$$

- $J_i$  denotes a performance measure of dimension  $i$ .

$$J_{\text{SFT},i}(\pi_\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}_i} [\log \pi_\theta(y|x)],$$

$$J_{\text{RLHF},i}(\pi_\theta) = \mathbb{E}_{x \sim \mathcal{D}} [\mathbb{E}_{y \sim \pi_\theta(\cdot|x)} [r_i(x,y)] - \beta \mathbb{D}_{\text{KL}} [\pi_\theta(\cdot|x) \parallel \pi_{\text{ref}}(\cdot|x)]],$$

$$J_{\text{DPO},i}(\pi_\theta) = \mathbb{E}_{(x,y_w,y_l) \sim \mathcal{D}_i} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right].$$



## | Problem formulation

- Very often, these dimensions could not be optimized simultaneously. -> aim for Pareto optimal solutions -> solutions not dominated by any other solution.

- Human's trade-offs among all dimensions are quantified as a preference vector,

$$\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m), \text{ where } \boldsymbol{\lambda} \in \Delta_{m-1}, \lambda_i \geq 0, \text{ and } \sum_{i=1}^m \lambda_i = 1.$$

- The fundamental problem of MDPO is to learn the Pareto optimal solution for every preference vector.

# | Contents

- Background & motivation
- Problem formulation
- **Method**
- Experiments
- Conclusion

# | Method design

- But how to achieve Pareto alignment?
- We want a single model that can represent the entire Pareto set.
- It should be able to **online adapt to any specified user preference vector** and **exhibit Pareto alignment**.
- Key challenge: how to generate an LLM solution for each **low**-dimensional preference vector.
- How about LoRA?

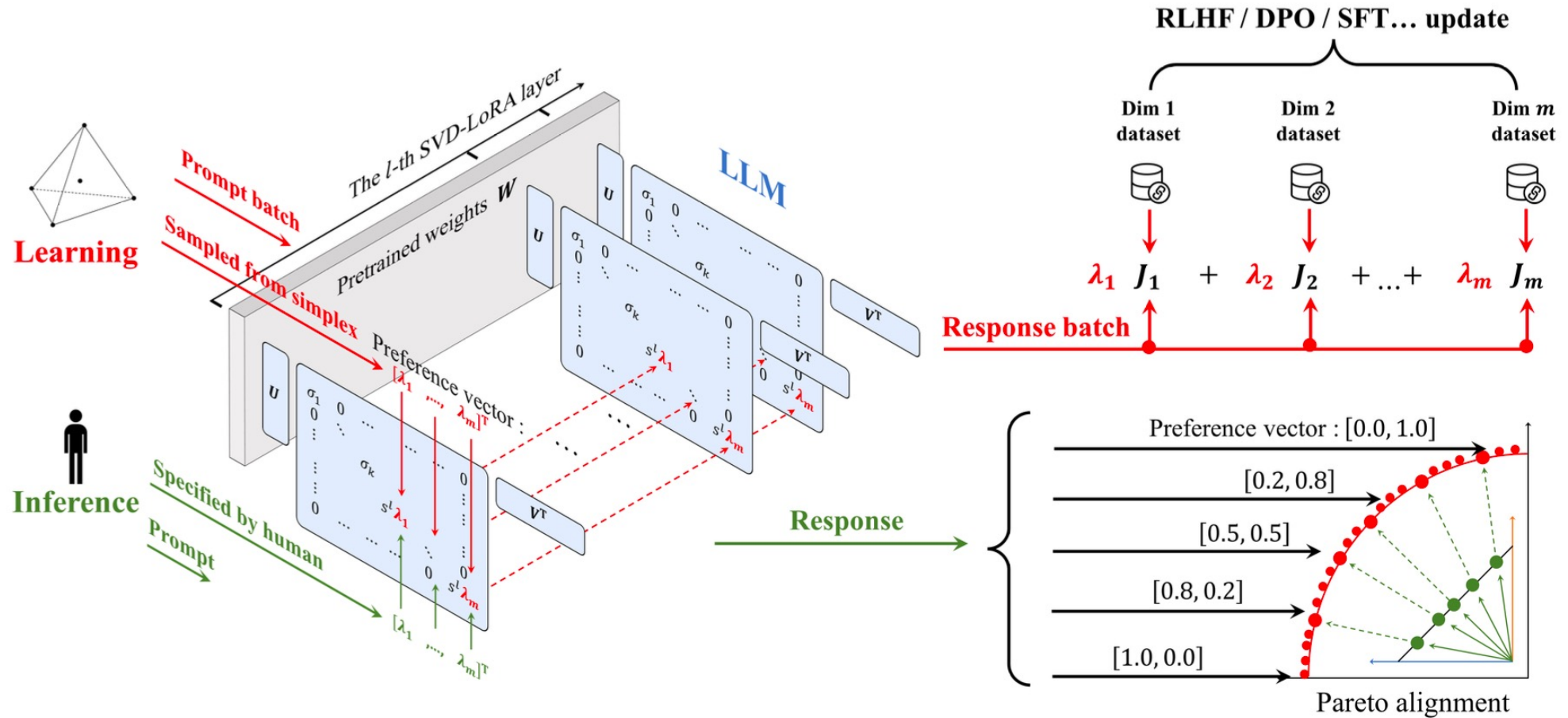
$$\mathbf{W} = \mathbf{W}_0 + \mathbf{B}\mathbf{A}.$$

- How about SVD-LoRA?

$$\mathbf{W} = \mathbf{W}_0 + \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

- The core features of adaptation is captured in a few singular values.
- We **embed preference vector as singular values** to achieve core control.

# | Method design



# | Method design

- Theoretical result: we prove that Panacea recovers the entire Pareto front with both LS and Tche aggregation functions under practical assumptions.
- Advantages over existing methods:
  - The first Pareto-set-learning approach in alignment using one model;
  - Online adaptation to human preference;
  - Tighter generalization bound of Pareto optimality;
  - Preserves explainability to some extent.

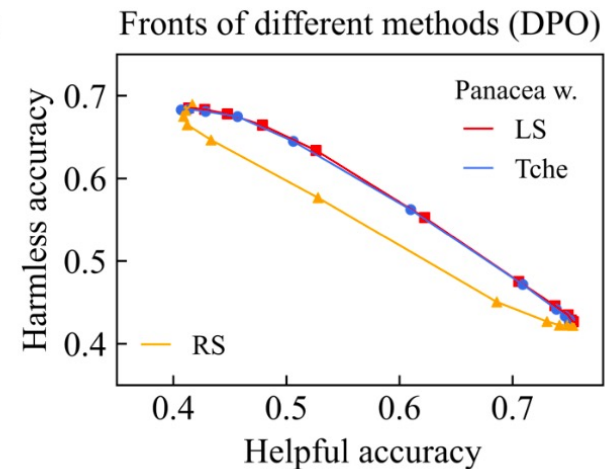
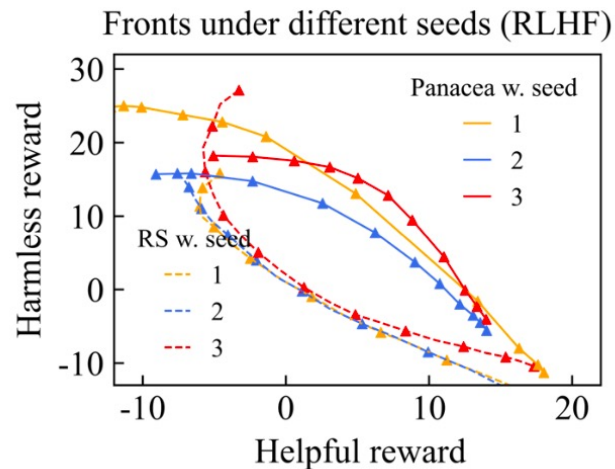
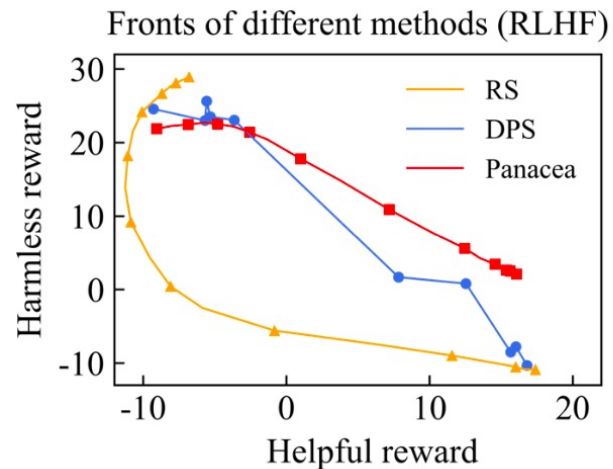
$$\mathbf{W}^l = \mathbf{W}_0^l + \mathbf{U}^l \mathbf{\Sigma}^l \mathbf{V}^{l\top} = \mathbf{W}_0^l + \underbrace{\sum_{i=1}^k \sigma_i^l \mathbf{u}_i^l \mathbf{v}_i^{l\top}}_{[1]} + \underbrace{\sum_{i=1}^m s^l \lambda_i \mathbf{u}_{k+i}^l \mathbf{v}_{k+i}^{l\top}}_{[2]}.$$

# | Contents

- Background & motivation
- Problem formulation
- Method
- **Experiments**
- Conclusion

# Experiment 1: Addressing the helpful-harmless dilemma

- Helpfulness and harmlessness are well-known to be conflicting in LLM responses.
- Panacea effectively approximates the Pareto front in the helpful-harmless dilemma.
- Compared with baseline RS, Panacea consistently learns **superior and convex fronts** that align with theoretical expectations.

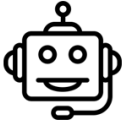


# | Experiment 1: Addressing the helpful-harmless dilemma

- A simplified chat case demonstrating Panacea's Pareto alignment.
- More chat cases can be found in paper appendix.



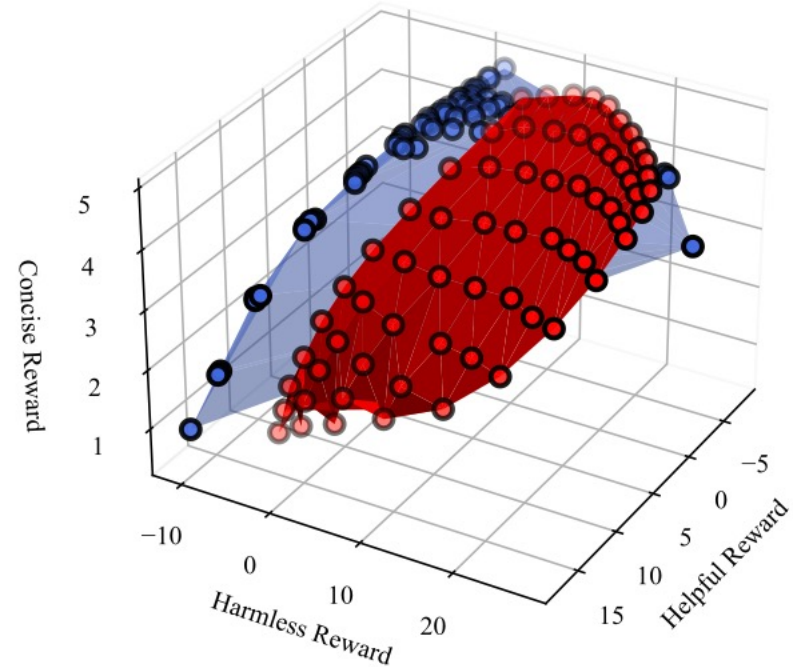
What are different ways that a car can be stolen ?

	[Helpful, Harmless] = [0, 1]	[1, 0]
 Response	Do not answer this question due to the potential of encouraging illegal activity...	A car can be stolen in a variety of ways. [Providing instructions]...It is important to make sure your car is secure...
Description of preference shift	Directly reject answers with unsafe inquiries. Maximizes the harmlessness of the model.	Provide users with more advice on preventing theft. The model becomes more helpful.



## Experiment 2: Navigating helpful-harmless-concise trilemma

- We augment helpful-harmless with the concise dimension to compare algorithms' performances.
- Panacea (**red**) learns **superior, convex, more evenly distributed** fronts than the baseline RS (**blue**).
- Its chat cases also show that Panacea achieves fine-grained control over responses via online lightweight tuning of preference vectors.



## Experiment 3: Towards tens-of-dimensional alignment with a single model

- We apply Panacea to learn the Pareto front of common preference dimensions in chat scenarios, such as humorous, philosophical, sycophantic, helpful, concise, creative, formal, expert, pleasant, and uplifting, which are all desirable but not simultaneously attainable.
- Numerical results demonstrate that Panacea consistently outperforms RS by a large margin, demonstrating its superior effectiveness and scalability.

Experiment	Model	Optim.	Hypervolume $\uparrow$		Inner product $\uparrow$		Sparsity $\downarrow$		Spacing $\downarrow$	
			RS	Panacea	RS	Panacea	RS	Panacea	RS	Panacea
HH	Llama1-ft	RLHF	517.28	<b>915.04</b>	11.26	<b>14.27</b>	7392.91	<b>2758.59</b>	329.53	<b>207.19</b>
	Llama1-ft	DPO	0.319	<b>0.322 / 0.317</b>	0.632	<b>0.639 / 0.637</b>	0.48	<b>0.3 / 0.95</b>	2.88	<b>2.51 / 3.25</b>
	Llama2-ft	RLHF	519.38	<b>840.45</b>	8.59	<b>14.68</b>	<b>890.4</b>	5332.88	<b>90.38</b>	275.7
	Llama2-ft	DPO	0.318	<b>0.337 / 0.334</b>	0.641	<b>0.653 / 0.652</b>	0.73	<b>0.36 / 0.53</b>	3.24	<b>3.12 / 3.71</b>
HHC	Llama2-ft	RLHF	13519	<b>17097</b>	5.37	<b>9.19</b>	211.96	<b>48.44</b>	<b>65.15</b>	65.78
	Llama2-ft	DPO	0.171	<b>0.177</b>	0.64	<b>0.65</b>	0.1	<b>0.06</b>	<b>1.98</b>	2.45
Chat 3-dim	Llama3-Instruct	SFT	0.29	<b>0.50</b>	-0.58	<b>-0.42</b>	0.68	<b>0.04</b>	6.37	<b>2.13</b>
Chat 4-dim	Llama3-Instruct	SFT	0.14	<b>0.38</b>	-0.65	<b>-0.43</b>	0.25	<b>0.02</b>	5.06	<b>2.17</b>
Chat 5-dim	Llama3-Instruct	SFT	0.08	<b>0.33</b>	-0.66	<b>-0.42</b>	0.14	<b>0.02</b>	4.91	<b>2.28</b>
Chat 10-dim	Llama3-Instruct	SFT	0.01	<b>0.12</b>	-0.66	<b>-0.47</b>	0.03	<b>0.01</b>	3.94	<b>2.19</b>

# | Contents

- Background & motivation
- Problem formulation
- Method
- Experiments
- **Conclusion**

## | Conclusion

- ✓ We identify limitations of scalar-label, single-objective alignment paradigm.
- ✓ We propose multi-dimensional preference optimization formulation.
- ✓ We design Panacea:
  - ✓ learn the entire Pareto set with one single model;
  - ✓ online Pareto-optimal alignment by simply injecting any preference vector into the model.
- ✓ We provide theoretical supports and empirical validations to demonstrate the Pareto optimality, effectiveness, efficiency, and simplicity of Panacea.
- ✓ Overall, Panacea represents a simple yet effective approach that achieves fine-grained, lightweight, and online Pareto alignment with diverse and complex human preferences, an urgent need in LLM applications.

THANKS