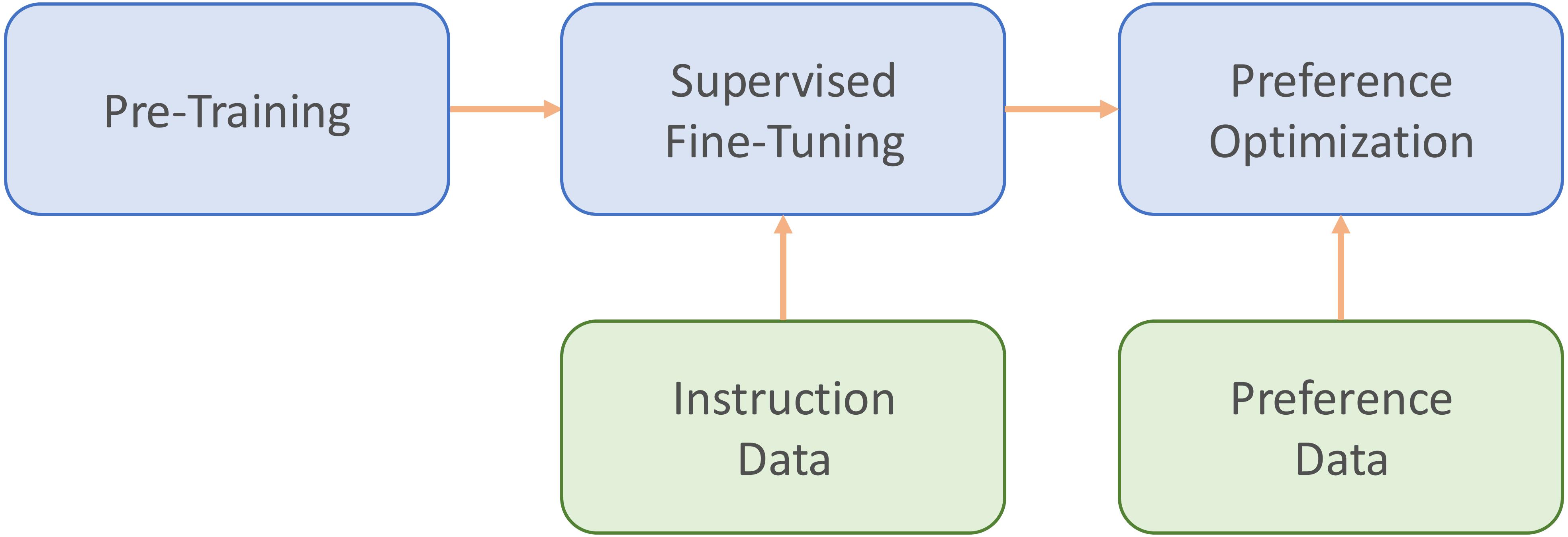
CSCE 689: Special Topics in Trustworthy NLP

Lecture 22: Human Preference Alignment (2)

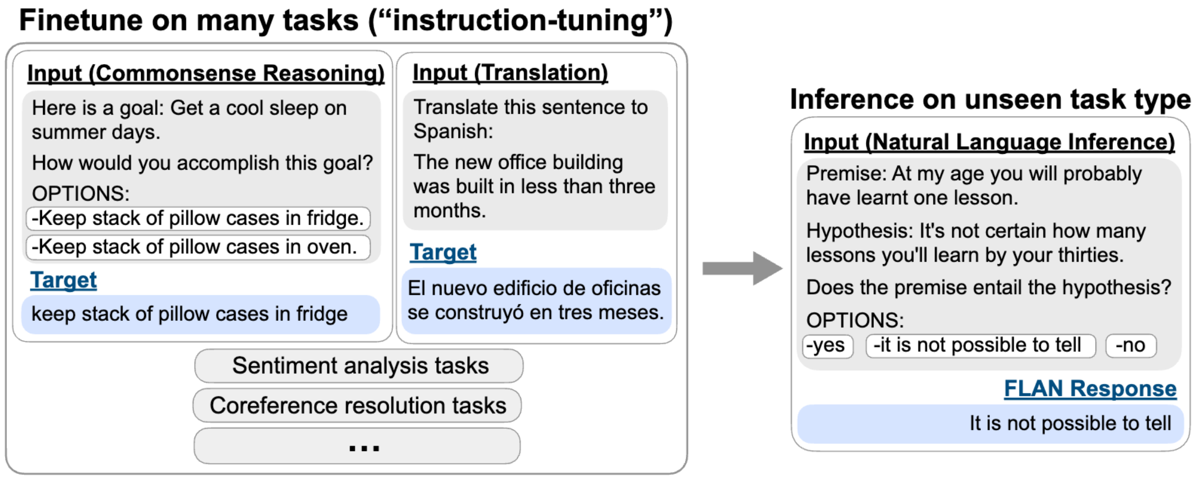
Kuan-Hao Huang
khhuang@tamu.edu

(Some slides adapted from Rafael Rafailov, Archit Sharma, and Eric Mitchell)

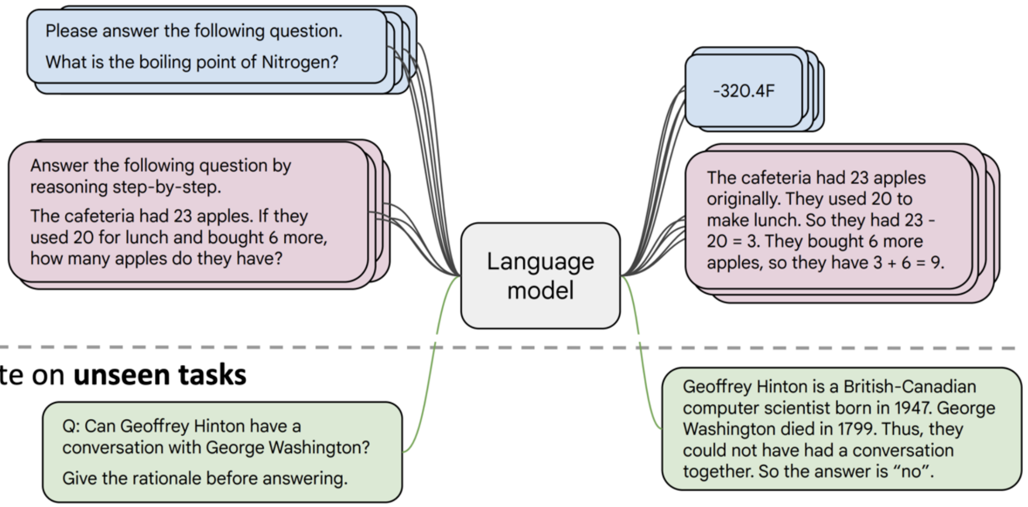
Recap: Alignment Pipeline



Recap: Instruction Fine-Tuning



- **Collect examples** of (instruction, output) pairs across many tasks and finetune an LM



- Evaluate on **unseen tasks**

Recap: Reinforcement Learning from Human Feedback

- Finally, we have everything we need:
  - A pretrained (possibly instruction-finetuned) LM $p^{PT}(s)$
  - A reward model $RM_\phi(s)$ that produces scalar rewards for LM outputs, trained on a dataset of human comparisons
  - A method for optimizing LM parameters towards an arbitrary reward function.
- Now to do RLHF:
  - Initialize a copy of the model $p_\theta^{RL}(s)$, with parameters θ we would like to optimize
  - Optimize the following reward with RL:

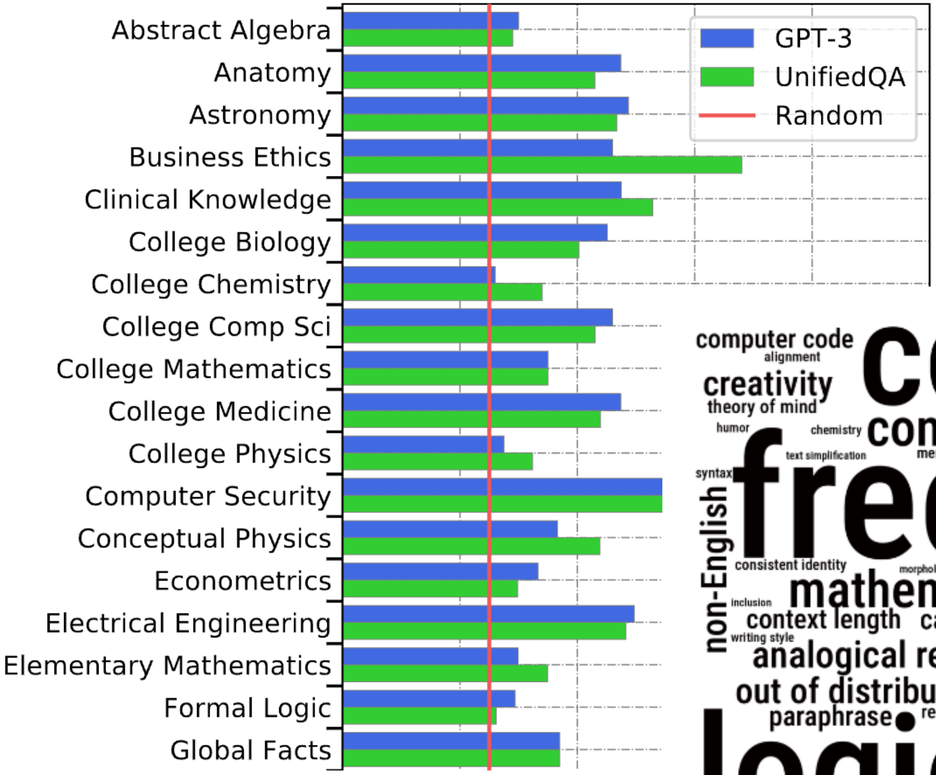
$$R(s) = RM_\phi(s) - \underbrace{\beta \log \left(\frac{p_\theta^{RL}(s)}{p^{PT}(s)} \right)}$$

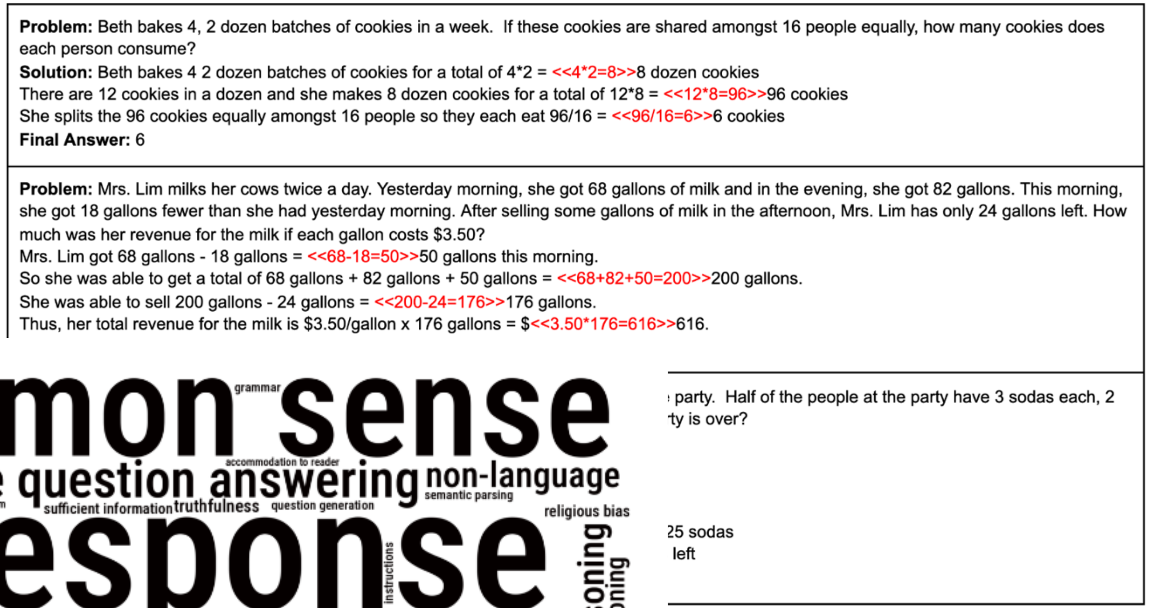
Pay a price when $p_\theta^{RL}(s) > p^{PT}(s)$

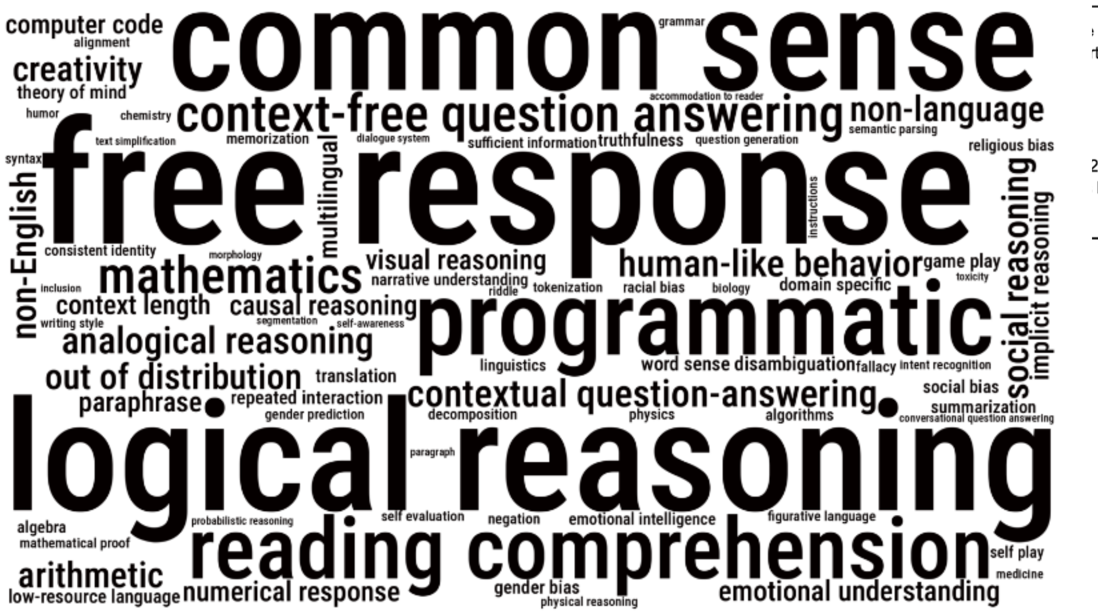
This is a penalty which prevents us from diverging too far from the pretrained model. In expectation, it is known as the **Kullback-Leibler** (**KL**) divergence between $p_\theta^{RL}(s)$ and $p^{PT}(s)$.

Recap: Evolution Benchmark

- MMLU, BIG-Bench, GSM8K, etc.







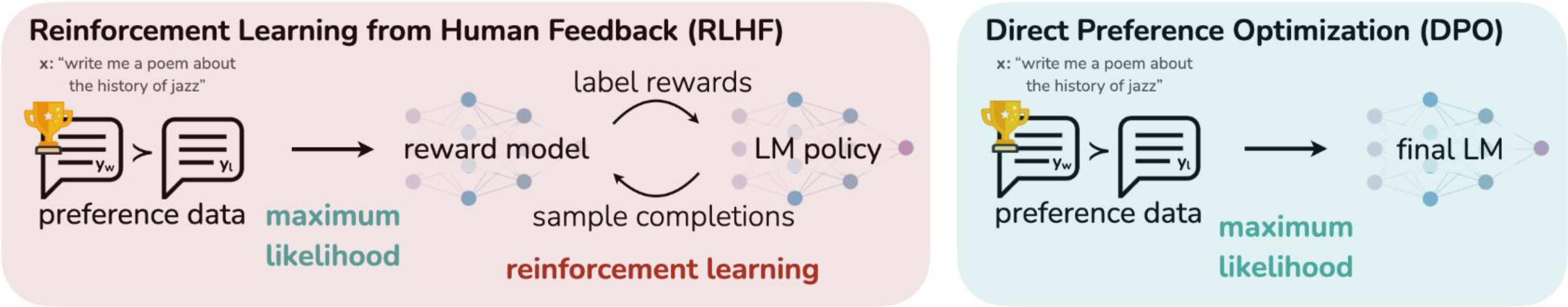
Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Rafael Rafailov[*†] **Archit Sharma**[*†] **Eric Mitchell**[*†]

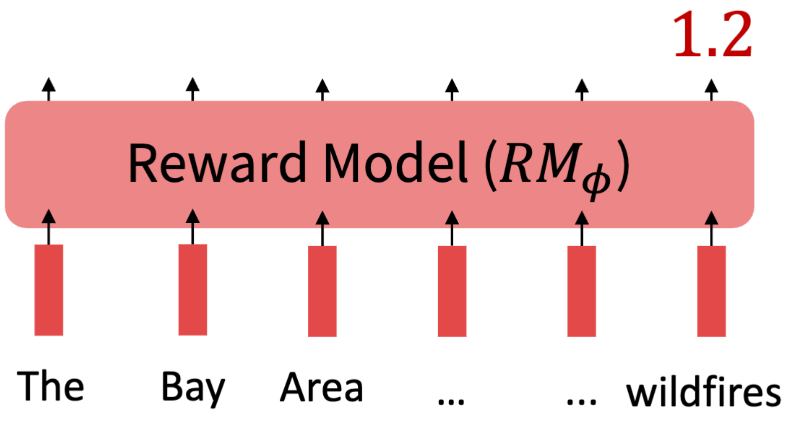
Stefano Ermon[†‡] **Christopher D. Manning**[†] **Chelsea Finn**[†]

[†]Stanford University [‡]CZ Biohub
{rafailov,architsh,eric.mitchell}@cs.stanford.edu

Direct Preference Optimization (DPO)



RLHF: Proximal Policy Optimization (PPO)



An earthquake hit San Francisco. There was minor property damage, but no injuries.

s_1

$>$

The Bay Area has good weather but is prone to earthquakes and wildfires.

s_2

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right]$$

Direct Preference Optimization (DPO)

RLHF Objective

(get **high reward**, stay **close** to reference model)

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi(y|x)}\left[r(x,y)\right] - \beta \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid x) \| \pi_{\mathrm{ref}}(\cdot \mid x))$$

Maximize reward

Keep similar behavior

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi}\left[r(x,y)\right] - \beta \mathbb{D}_{\mathrm{KL}}\left[\pi(y|x) \mid\mid \pi_{\mathrm{ref}}(y|x)\right]$$

$$= \max_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[r(x,y) - \beta \log \frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log \frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)} - \frac{1}{\beta} r(x,y)\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi(y|x)}\left[\log \frac{\pi(y|x)}{\frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}r(x,y)\right)} - \log Z(x)\right]$$

$$Z(x) = \sum_{y} \pi_{\mathrm{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x,y)\right)$$

Direct Preference Optimization (DPO)

RLHF Objective

(get **high reward**, stay **close** to reference model)

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi(y|x)} \left[r(x,y)\right] - \beta \mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid x) \| \pi_{\mathrm{ref}}(\cdot \mid x))$$

Maximize reward

Keep similar behavior

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{\mathrm{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x,y)\right)$$

$$\min_{\pi} \mathbb{E}_{x\sim\mathcal{D}} \mathbb{E}_{y\sim\pi(y|x)} \left[\log \frac{\pi(y|x)}{\frac{1}{Z(x)} \pi_{\mathrm{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x,y)\right)} - \log Z(x)\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}} \left[\mathbb{E}_{y\sim\pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi^*(y|x)}\right] - \log Z(x)\right]$$

$$= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}} \left[\mathbb{D}_{\mathrm{KL}}(\pi(y|x) \mid\mid \pi^*(y|x)) - \log Z(x)\right]$$

$$\pi(y|x) = \pi^*(y|x) = \frac{1}{Z(x)} \pi_{\mathrm{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x,y)\right)$$

Direct Preference Optimization (DPO)

RLHF Objective

(get **high reward**, stay **close** to reference model)

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi(y|x)}\left[r(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}(\pi(\cdot \mid x)\|\pi_{\mathrm{ref}}(\cdot \mid x))$$

Maximize reward

Keep similar behavior

Closed-form Optimal Policy

(write **optimal policy** as function of **reward function**; from prior work)

$$\pi^*(y \mid x) = \frac{1}{Z(x)}\pi_{\mathrm{ref}}(y \mid x)\exp\left(\frac{1}{\beta}r(x,y)\right)$$

with $Z(x) = \sum_y \pi_{\mathrm{ref}}(y \mid x)\exp\left(\frac{1}{\beta}r(x,y)\right)$

Note **intractable sum** over possible responses; can't immediately use this

Rearrange

(write **any reward function** as function of **optimal policy**)

Ratio is **positive** if policy likes response more than reference model, **negative** if policy likes response less than ref. model

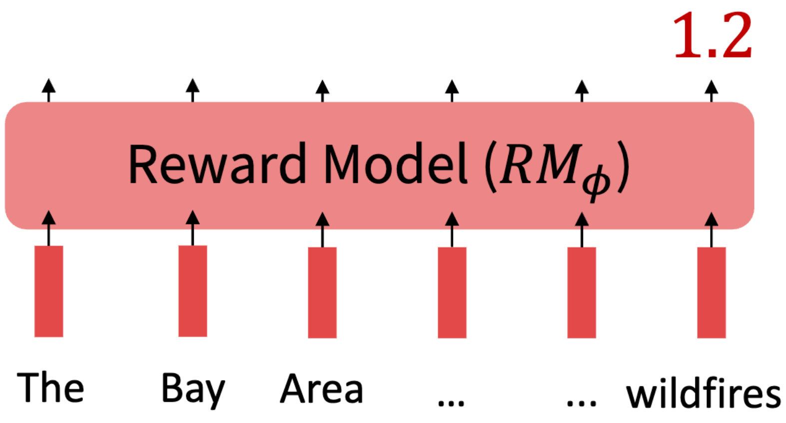
$$r(x,y) = \underbrace{\beta\log\frac{\pi^*(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} + \beta\log Z(x)}_{\text{some parameterization of a reward function}}$$

Direct Preference Optimization (DPO)

A loss function on <u>reward functions</u>

Derived from the Bradley-Terry model of human preferences:

$$\mathcal{L}_R(r, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma(r(x, y_w) - r(x, y_l)) \right]$$



An earthquake hit San Francisco. There was minor property damage, but no injuries.

S_1

$>$

The Bay Area has good weather but is prone to earthquakes and wildfires.

S_2

Direct Preference Optimization (DPO)

Derived from the Bradley-Terry model of human preferences:

A loss function on <u>reward functions</u>

$$\mathcal{L}_R(r, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma(r(x, y_w) - r(x, y_l)) \right]$$

+

A transformation between <u>reward functions</u> and <u>policies</u>

$$r_{\pi_\theta}(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x)$$

Direct Preference Optimization (DPO)

Derived from the Bradley-Terry model of human preferences:

A loss function on <u>reward functions</u>

$$\mathcal{L}_R(r, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma(r(x, y_w) - r(x, y_l)) \right]$$

+

A transformation between <u>reward functions</u> and <u>policies</u>

$$r_{\pi_\theta}(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x)$$

=

A loss function on <u>policies</u>

Reward of **preferred** response

Reward of **dispreferred** response

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

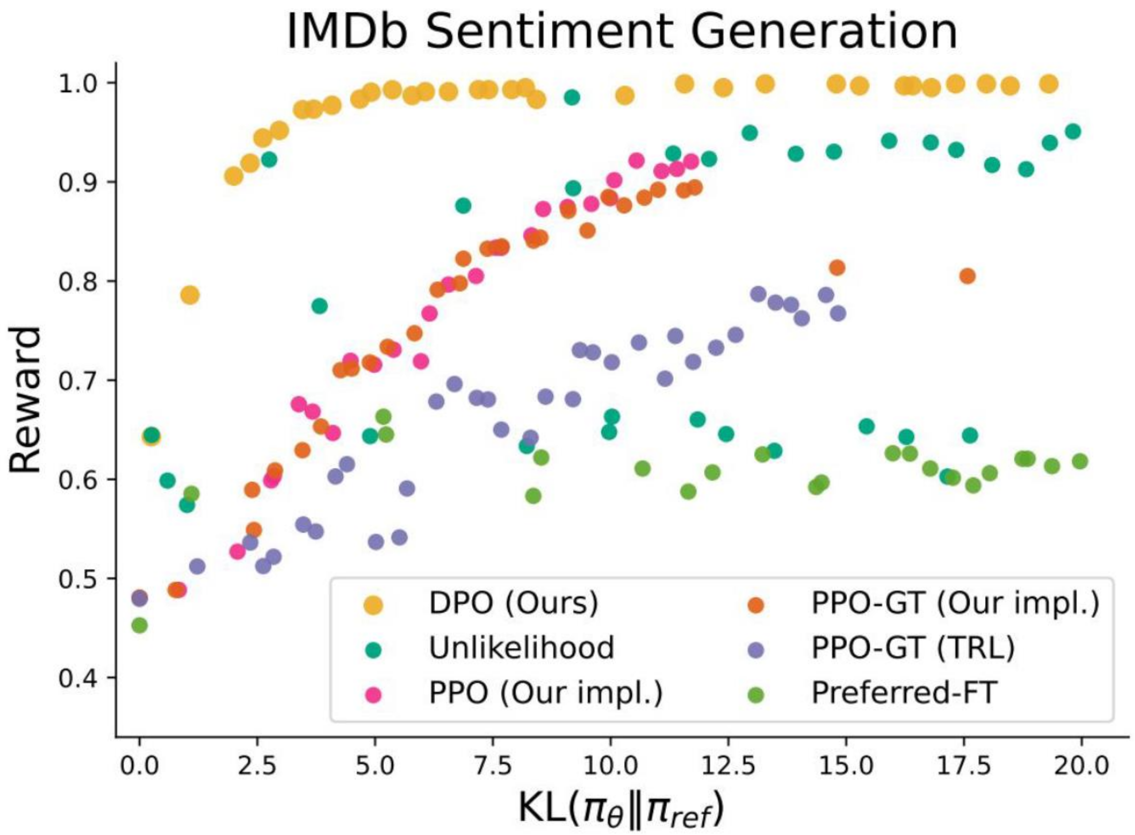
Direct Preference Optimization (DPO)

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\underbrace{\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}}_{\text{Reward of preferred response}} - \underbrace{\beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}}_{\text{Reward of dispreferred response}} \right) \right]$$

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$

$$-\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \bigg[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \bigg[\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease likelihood of } y_l} \bigg] \bigg]$$

Results

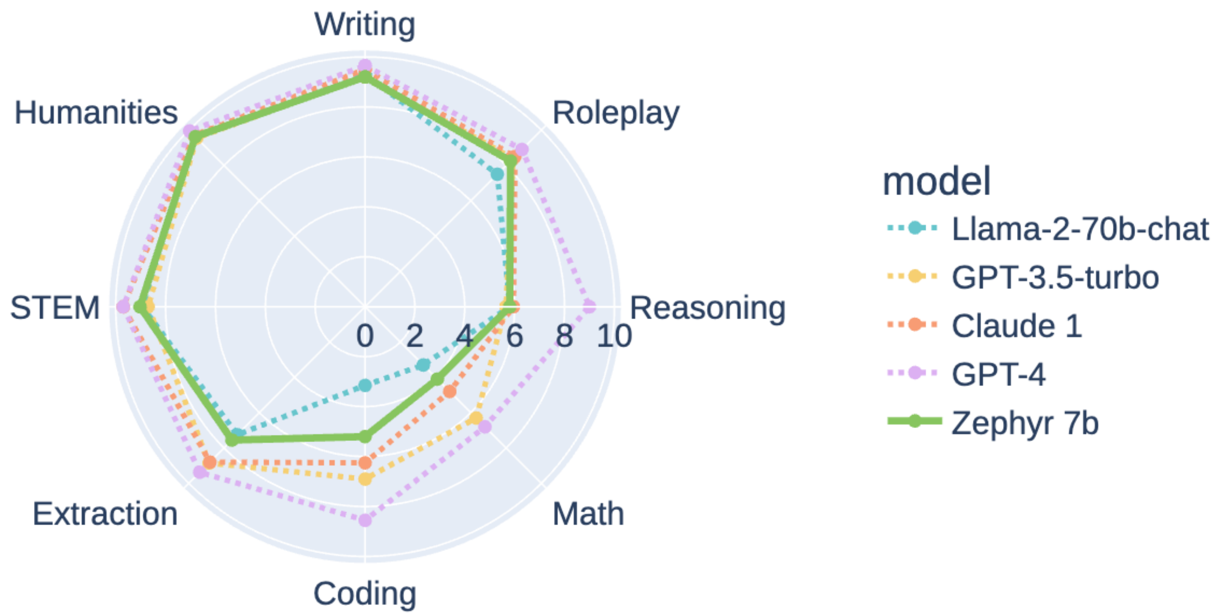


1. Generate positive IMDB reviews from GPT2-XL
2. Use pre-trained sentiment classifier as Gold RM
3. Create preferences based on Gold RM
4. Optimize with PPO and DPO

Large-Scale DPO Training

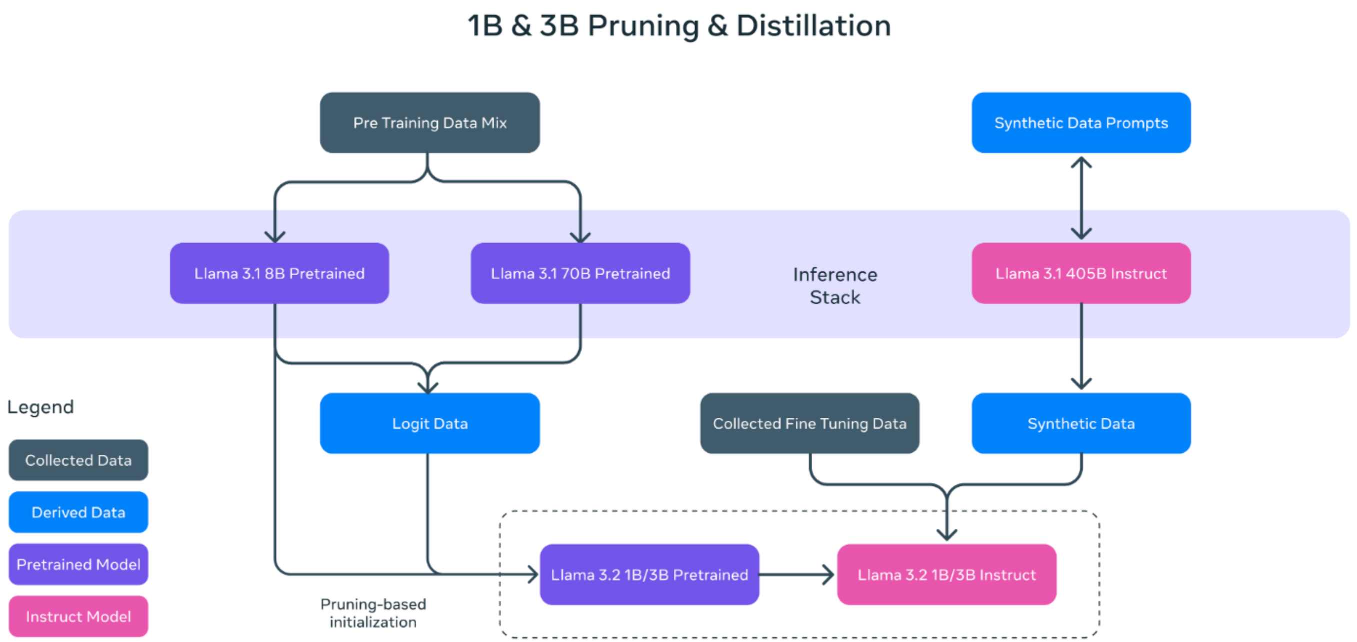
ZEPHYR: DIRECT DISTILLATION OF LM ALIGNMENT

Lewis Tunstall,* Edward Beeching,* Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf
The H4 (Helpful, Honest, Harmless, Huggy) Team
https://huggingface.co/HuggingFaceH4
lewis@huggingface.co



Large-Scale DPO Training

Llama 3.2: Revolutionizing edge AI and vision with open, customizable models



In post-training, we use a similar recipe as Llama 3.1 and produce final chat models by doing several rounds of alignment on top of the pre-trained model. Each round involves supervised fine-tuning (SFT), rejection sampling (RS), and direct preference optimization (DPO).

KTO: Model Alignment as Prospect Theoretic Optimization

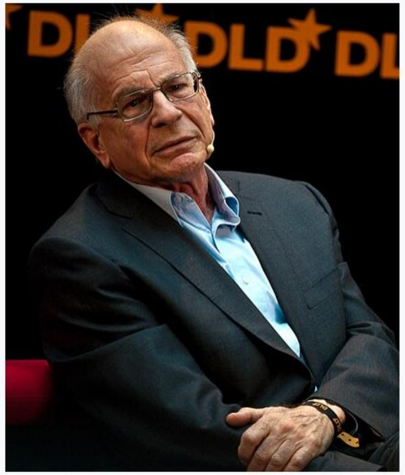
Kawin Ethayarajh [1] **Winnie Xu** [2] **Niklas Muennighoff** [2] **Dan Jurafsky** [1] **Douwe Kiela** [1 2]

Prospect Theory

Prospect theory explains why humans make decisions about uncertain events that do not maximize expected value. It formalizes how humans perceive random variables in a biased but well-defined manner;
for example, relative to some **reference point**, humans are more sensitive to losses than gains, a property called **loss aversion**.

2002 Nobel Prize-winning economists

Daniel Kahneman



Amos Tversky

https://medium.com/@yianyao1994/llm-alignments-part-6-kto-813b38be14ae

Prospect Theory

- Imagine you are facing two choices:
  - **Choice one**: has an 80% chance of earning you 10 million US dollars, and a 20% chance of giving you nothing
  - **Choice two**: gives you 4 million US dollars for sure

many people choose the second option because it is more guaranteed

https://medium.com/@yianyao1994/llm-alignments-part-6-kto-813b38be14ae

Which One Do You Choose?

- Imagine you are facing two choices:
  - **Choice one**: has an 80% chance of earning you 10 million US dollars, and a 20% chance of giving you nothing
  - **Choice two**: gives you 4 million US dollars for sure

https://medium.com/@yianyao1994/llm-alignments-part-6-kto-813b38be14ae

Which One Do You Choose?

- Imagine you are facing two choices:
  - **Choice one**: has an 80% chance of earning you 1 thousand US dollars, and a 20% chance of giving you nothing
  - **Choice two**: gives you 4 hundred US dollars for sure

https://medium.com/@yianyao1994/llm-alignments-part-6-kto-813b38be14ae

Which One Do You Choose?

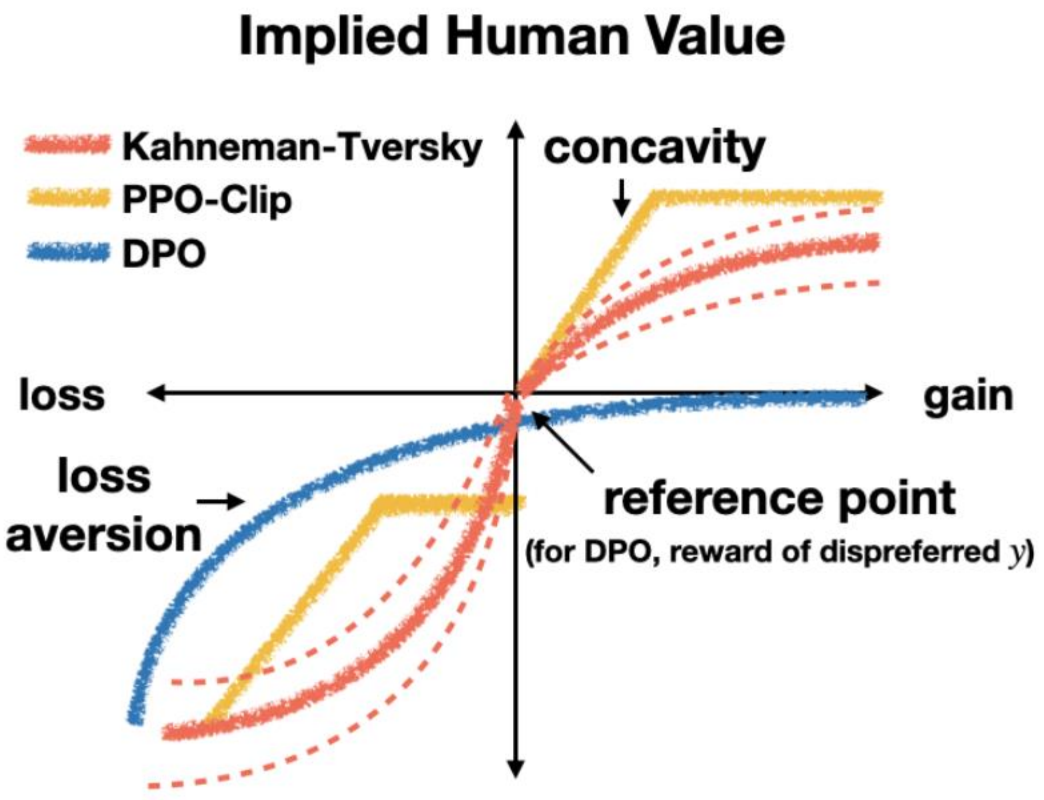
- Imagine you are facing two choices:
  - **Choice one**: has an 80% chance of earning you 10 US dollars, and a 20% chance of giving you nothing
  - **Choice two**: gives you 4 US dollars for sure

https://medium.com/@yianyao1994/llm-alignments-part-6-kto-813b38be14ae

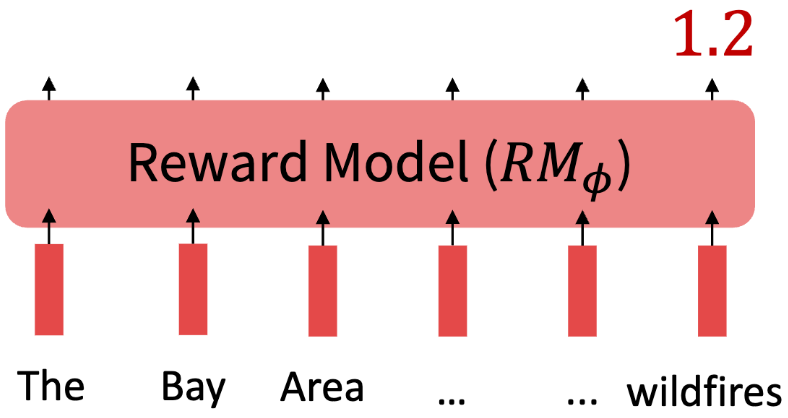
Prospect Theory

- There exist a reference point
  - Relative to the reference point, the value for gains is concave, meaning the more we gain, the less value we perceive
  - On the other hand, the value for losses can be either concave and convex

KTO Value Function



Preference Data For PPO/DPO



An earthquake hit San Francisco. There was minor property damage, but no injuries.

s_1

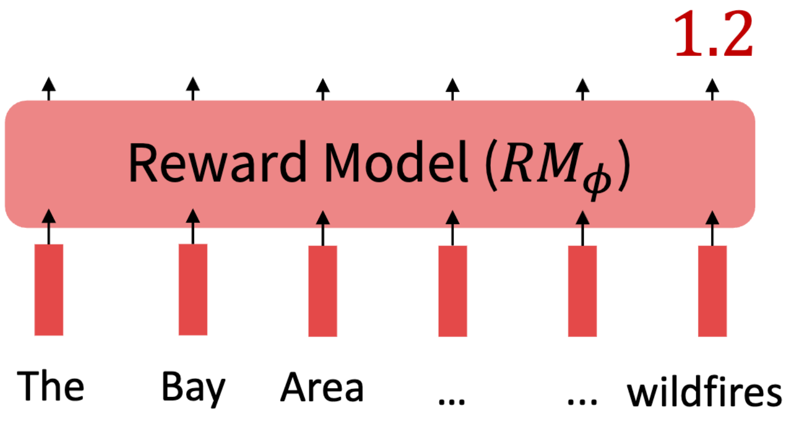
$>$

The Bay Area has good weather but is prone to earthquakes and wildfires.

s_2

Training Data (x, y_1, y_2)

Preference Data For KTO




```
An earthquake hit
San Francisco.
There was minor
property damage,
but no injuries.
```

s_1

Acceptable?

Training Data (x, y)

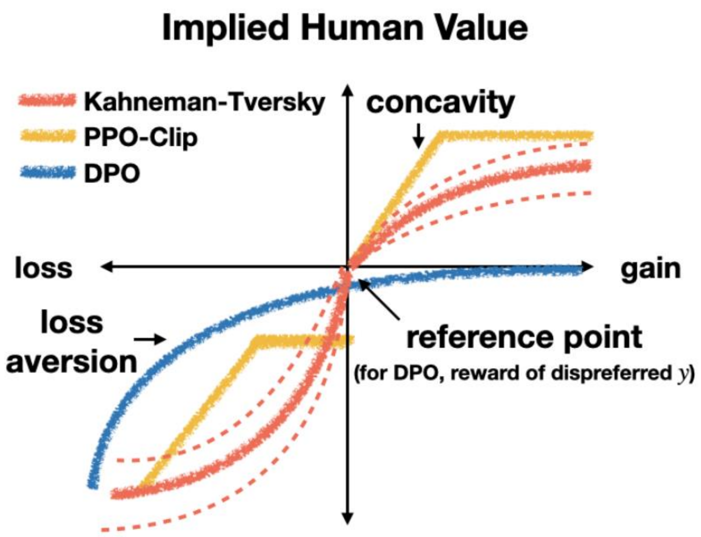
KTO: Reference point

- **Reference point**: Directly defined by the expectation over the distribution of (x, y) pairs

Reference Point:

$$\mathbb{E}_{x' \sim D} \left[\beta \, \mathrm{KL}(\pi_\theta(y'|x') \| \pi_{\text{ref}}(y'|x')) \right]$$

$$\mathbb{E}_{x' \sim D, y' \sim \pi^*} [r^*(x', y')]$$



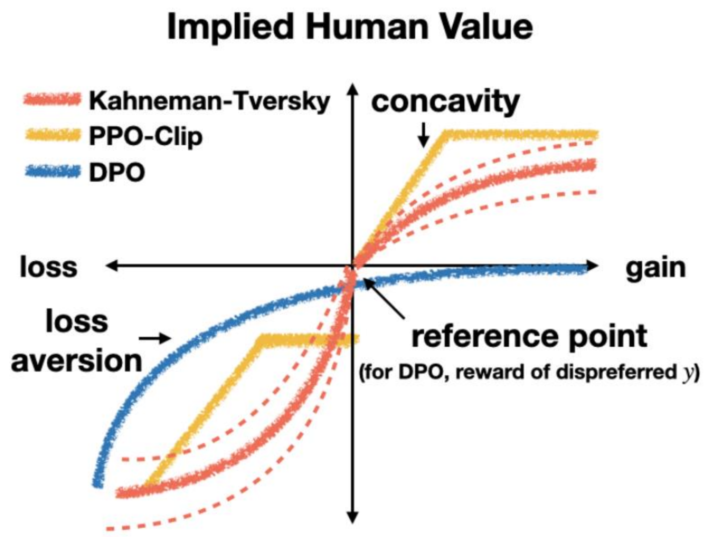
KTO: Loss Function

$$L_{\text{KTO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x,y\sim D}[\lambda_y - v(x, y)]$$

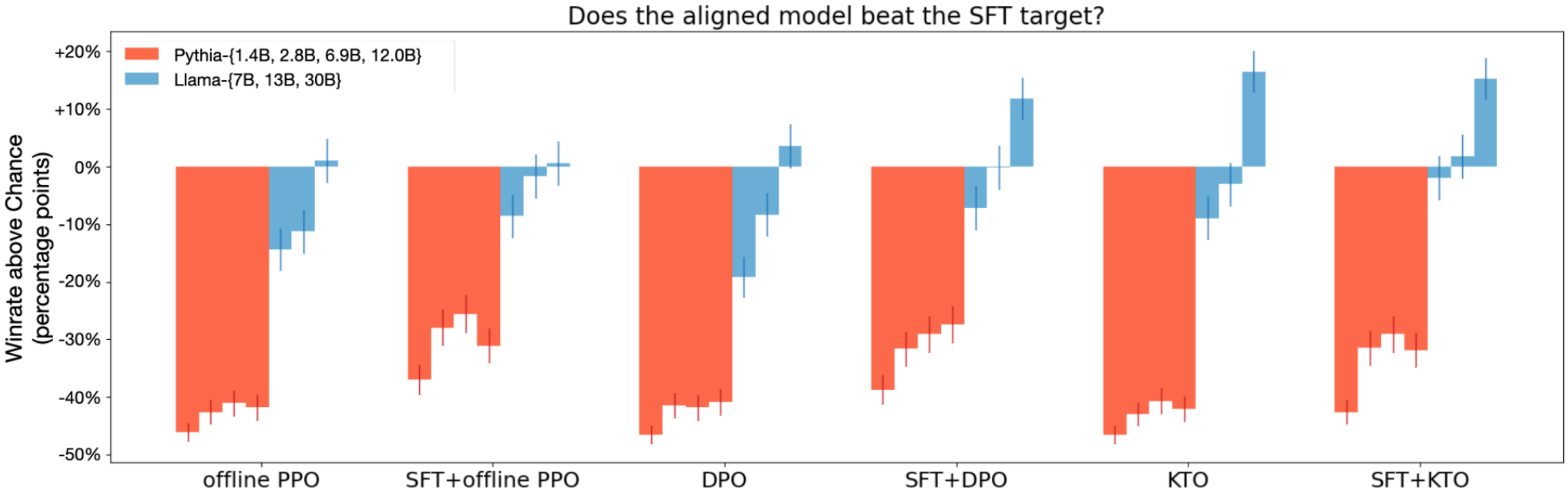
$$r_{\text{KTO}}(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$$

$$v_{\text{KTO}}(x, y; \beta) = \begin{cases} \sigma(r_{\text{KTO}}(x, y) - z_{\text{ref}}) \text{ if } y \sim y_{\text{desirable}}|x \\ \sigma(z_{\text{ref}} - r_{\text{KTO}}(x, y)) \text{ if } y \sim y_{\text{undesirable}}|x \end{cases}$$

$$w(y) = \begin{cases} \lambda_D & \text{if } y \sim y_{\text{desirable}}|x \\ \lambda_U & \text{if } y \sim y_{\text{undesirable}}|x \end{cases}$$



Results



Does the aligned model beat the SFT target?
Pythia-{1.4B, 2.8B, 6.9B, 12.0B}
Llama-{7B, 13B, 30B}
Winrate above Chance
(percentage points)
+20%
+10%
0%
-10%
-20%
-30%
-40%
-50%
offline PPO
SFT+offline PPO
DPO
SFT+DPO
KTO
SFT+KTO

mDPO: Conditional Preference Optimization for Multimodal Large Language Models

Fei Wang[1] **Wenxuan Zhou**[1] **James Y. Huang**[1] **Nan Xu**[1]
Sheng Zhang[2] **Hoifung Poon**[2] **Muhao Chen**[3]

[1]University of Southern California [2]Microsoft Research [3]University of California, Davis

Multimodal Large Language Models

Adversarial

What is a vicuna standing in the sand looking at?

mDPO The image features a drawing of a toucan, not a vicuna. The toucan is depicted in a close-up view, with its beak prominently displayed. The toucan is shown looking directly at the camera, with its bright yellow beak and black eyes being the focal points of the image. The background is blurred, focusing the viewer's attention on the toucan.

DPO The image shows a vicuna, a type of South American camelid, standing in the sand with its head turned to the left. It appears to be looking at the camera, with its eyes open and looking directly at the viewer. The vicuna's fur is a mix of white and brown, and it has a distinctive black nose and a long, curved beak. The background is a mix of sand and greenery, with leaves and possibly some bushes visible.

Bunny The vicuna is looking at the camera.

Relation

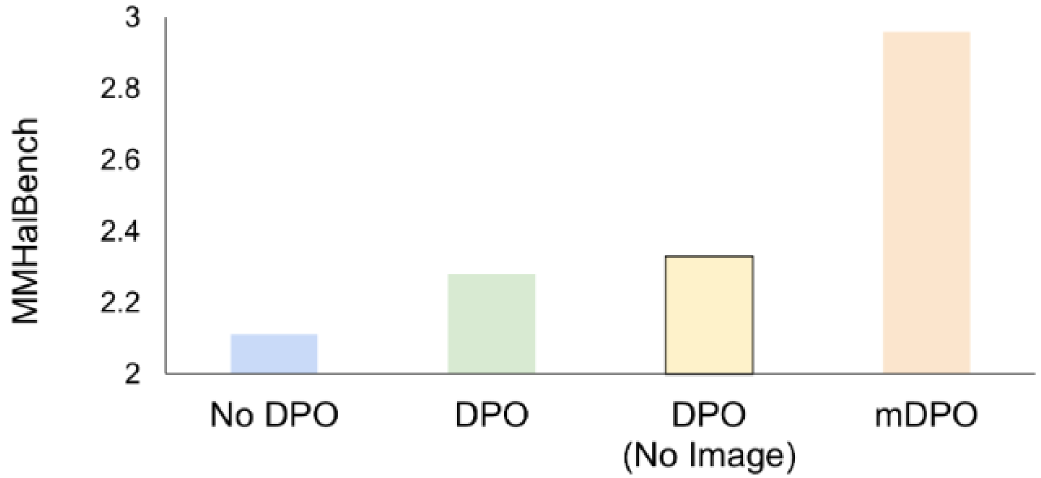
Is the mouse connected to a computer?

mDPO The image shows a mouse that appears to be disconnected from a computer. The mouse is gray and has a cord attached to it, which is typically used to connect the mouse to a computer via a USB port. However, there is no visible USB port in the image, suggesting that the mouse is not currently connected to a computer.

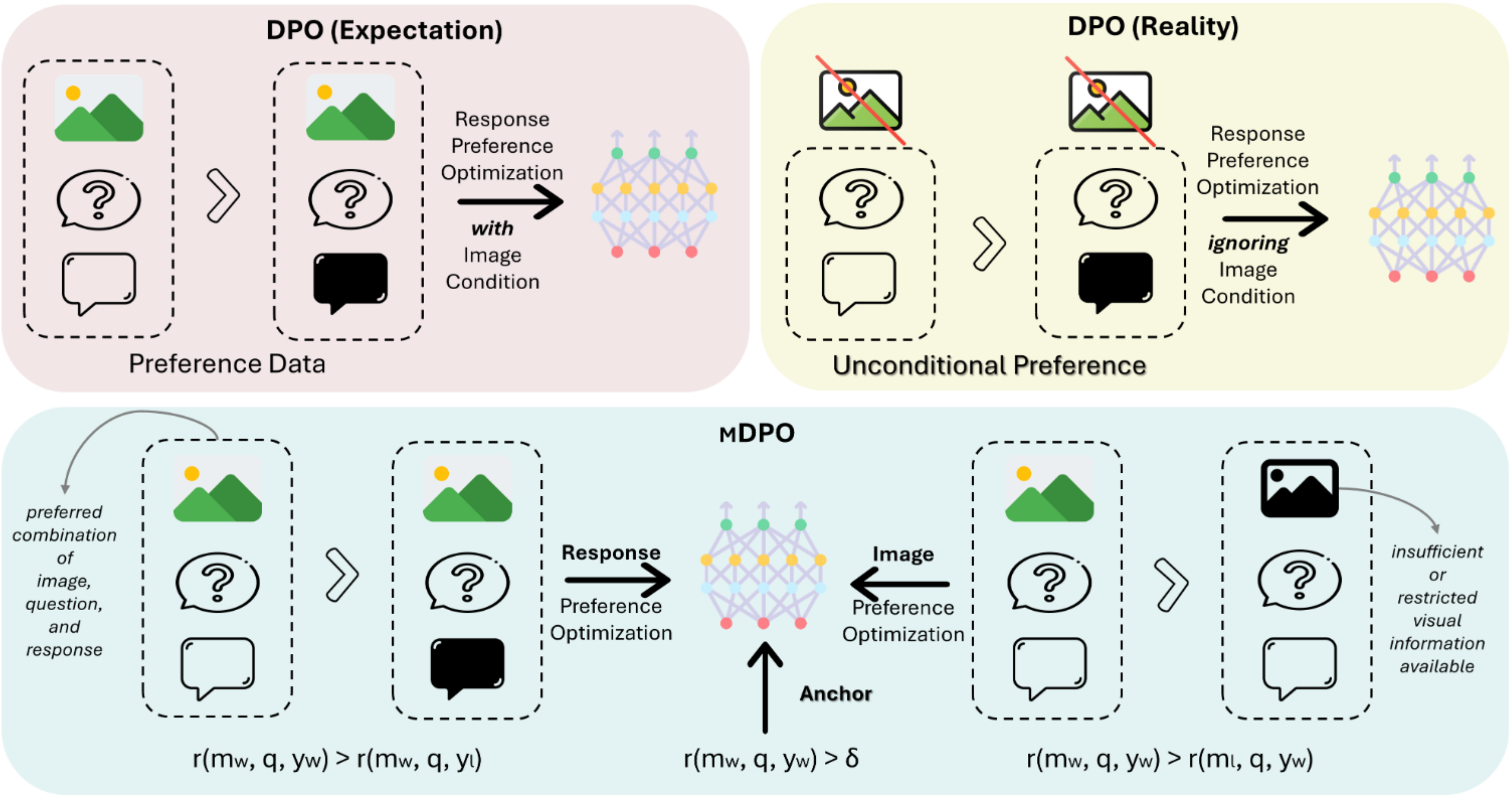
DPO The mouse appears to be connected to a computer via a USB cable. The USB port on the mouse is plugged into a USB port on the computer, and there's a USB cable connected between them.

Bunny Yes, the mouse is connected to a computer via a USB cable.

Issue of DPO



mDPO: DPO for Multimodal Large Language Models



Results

	MMHalBench		Object HalBench		AMBER			
	Score ↑	HalRate ↓	CHAIR_s ↓	CHAIR_i ↓	CHAIR_s ↓	Cover. ↑	HalRate ↓	Cog. ↓
3B Multimodal LLMs								
Bunny-v1.0-3B (He et al., 2024)	2.11	0.58	43.0	8.9	9.8	**75.6**	64.9	6.0
+ DPO	2.28	0.56	44.3	7.6	7.9	74.1	58.9	4.8
+ mDPO	**2.96**	**0.42**	**27.0**	**4.6**	**4.9**	67.4	**37.7**	**2.4**
7B Multimodal LLMs								
LLaVA-v1.5-7B (Liu et al., 2024a)	2.19	0.57	54.7	15.9	7.4	51.8	34.7	4.1
+ DPO	2.14	0.65	49.0	13.0	6.5	**55.1**	34.5	**2.3**
+ mDPO	**2.39**	**0.54**	**35.7**	**9.8**	**4.4**	52.4	**24.5**	2.4