



CS 329X: Human Centered LLMs

Preference Tuning & Alignment

Diyi Yang

Homework 1

Homework 1

Start Assignment

Due Oct 16 by 11:59pm

Points 100

Submitting a text entry box, a website url, or a file upload

File Types pdf and ipynb

Available Oct 3 at 9am - Oct 19 at 11:59pm

- **Deadline:** Oct 16th EOD (23:59 PT). Upload through Canvas.
- **Goal:** learn about pre-training and fine-tuning; annotate your own preferences; and then personalize LLM to your own preference data!
- **Important:** Take annotation seriously! The default course project will use everyone's preference data.

Learning from Human Feedback

- ✓ Different type of human feedback
- ✓ Learning from human feedback
 - ✓ Dataset updates (weak supervision, data augmentation)
 - ✓ Loss function updates (unlikelihood learning)
 - ✓ Parameter space updates (parameter efficient fine-tuning, model editing)
- ✓ RLHF
- DPO
- Limitations of human feedback

Recap the RLHF Objective

$$J(\pi_\theta) = \mathbb{E}_{y \sim \pi_\theta(x)} [R(x, y)] - \beta D_{KL}(\pi_\theta(x) \parallel \pi_{ref}(x))$$

x : input

y : model output (response)

$\pi_\theta(x)$: policy we're optimizing

$R(x, y)$: reward function based on human feedback

β : KL divergence regularization weight

Optimal Policy Under RLHF [Rafailov+ 2023]

Optimal Policy: closed-form solution from prior work

$$\pi_{\theta}^*(y|x) \propto \pi_{ref}(y|x) \exp\left(\frac{R(x, y)}{\beta}\right)$$

Normalized Policy

$$\pi_{\theta}^*(y|x) = \frac{\pi_{ref}(y|x) \exp\left(\frac{R(x, y)}{\beta}\right)}{Z(x)} \quad Z(x) = \sum_{y'} \pi_{ref}(y'|x) \exp\left(\frac{R(x, y')}{\beta}\right)$$

Log transformation: $R(x, y) = \beta (\log \pi_{\theta}^*(y|x) - \log \pi_{ref}(y|x)) + \beta \log Z(x)$

Putting it Together with DPO [Rafailov+ 2023]

Derived DPO reward model:

$$R(x, y) = \beta(\log \pi_{\theta}^*(y|x) - \log \pi_{ref}(y|x)) + \beta \log Z(\mathbf{x})$$

The Bradley-Terry model of human preferences

$$L_R(r, D) = -\mathbb{E}_{(x, \mathbf{y}_w, \mathbf{y}_l) \sim D} [\log \sigma(R(x, \mathbf{y}_w) - R(x, \mathbf{y}_l))]$$

Log Z term cancels as the loss only measures differences in rewards

Final loss function for DPO:

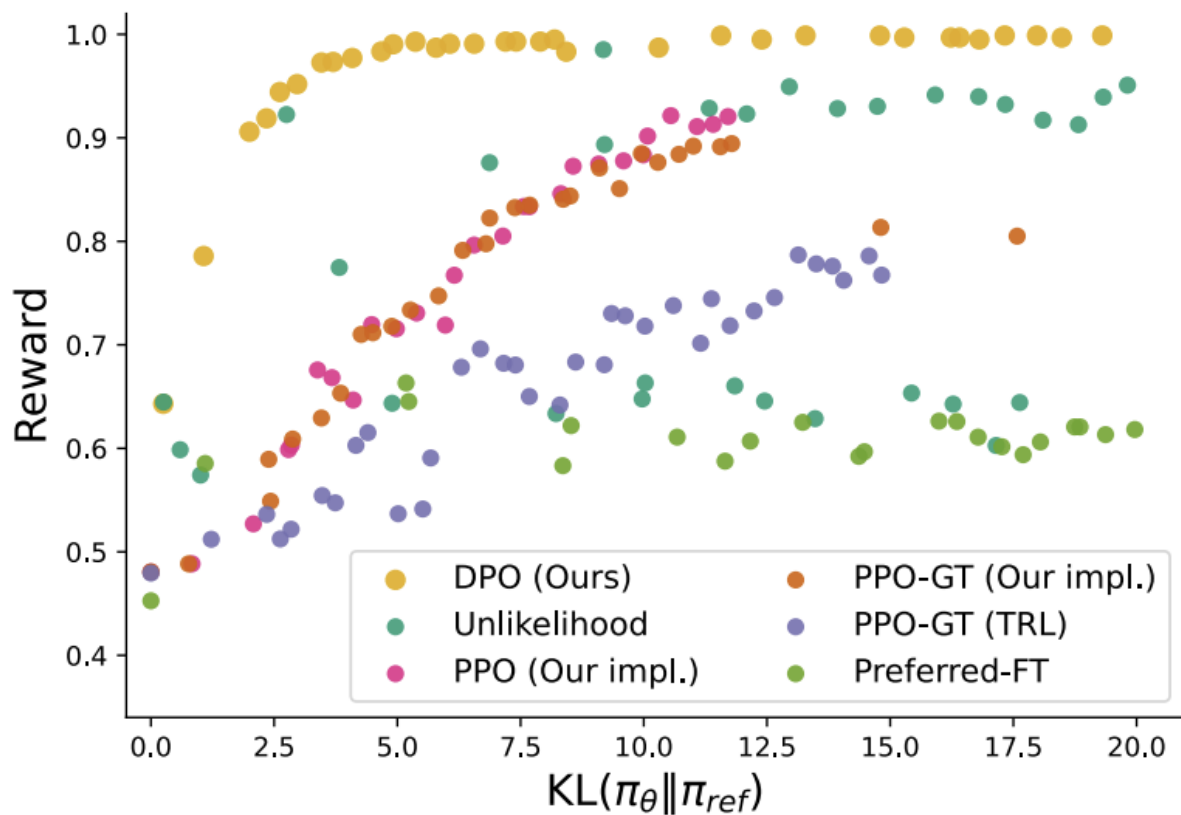
$$L_{DPO}(\pi_{\theta}, \pi_{ref}) = -\mathbb{E}_{(x, \mathbf{y}_w, \mathbf{y}_l) \sim D} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(\mathbf{y}_w|x)}{\pi_{ref}(\mathbf{y}_w|x)} - \beta \log \frac{\pi_{\theta}(\mathbf{y}_l|x)}{\pi_{ref}(\mathbf{y}_l|x)} \right) \right]$$

Reward for
winning sample

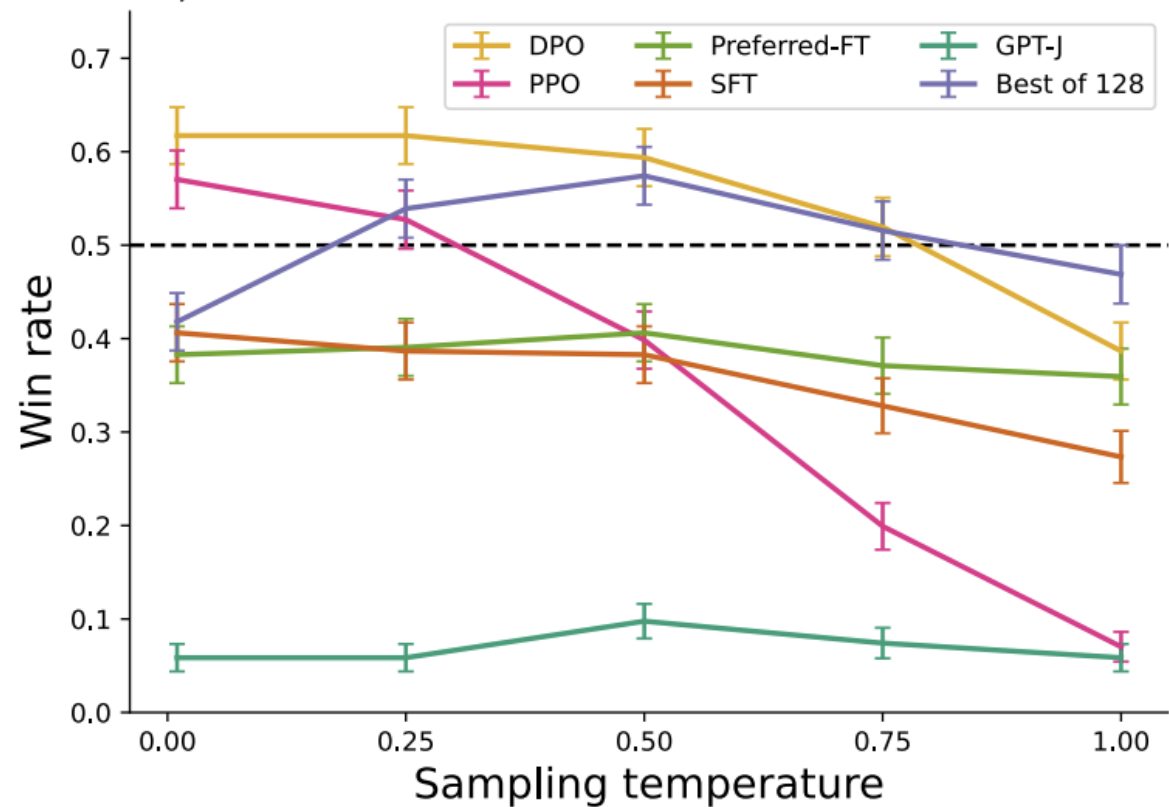
Reward for
losing sample

DPO Outperforms and Works Well at Scale

IMDb Sentiment Generation



TL;DR Summarization Win Rate vs Reference

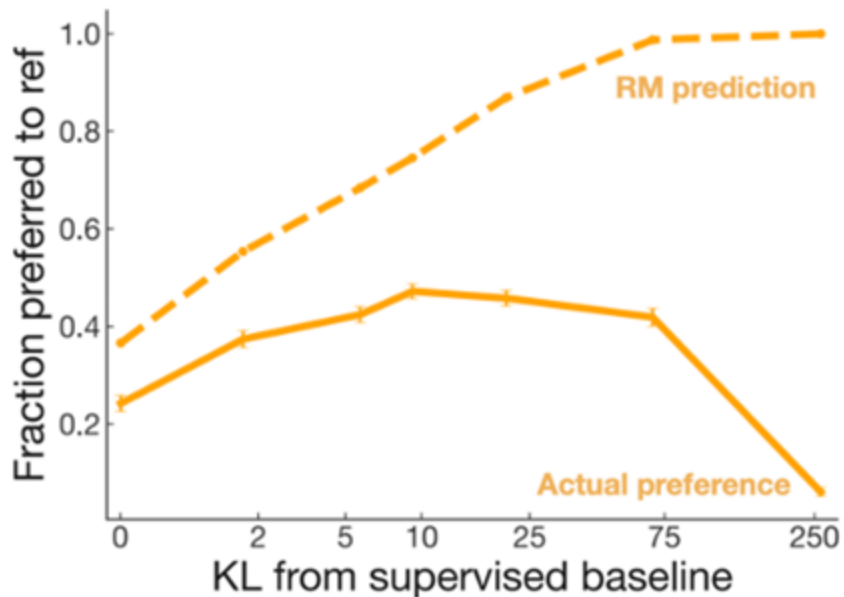


Learning from Human Feedback

- ✓ Different type of human feedback
- ✓ Learning from human feedback
 - ✓ Dataset updates (weak supervision, data augmentation)
 - ✓ Loss function updates (unlikelihood learning)
 - ✓ Parameter space updates (parameter efficient fine-tuning, model editing)
- ✓ RLHF
- ✓ DPO
- ☐ Limitations of human feedback

Limitations of Human Feedback

- Human preferences can be unreliable
- Reward hacking is a common problem in RL



TECHNOLOGY

Google shares drop \$100 billion after its new AI chatbot makes a mistake

February 9, 2023 · 10:15 AM ET

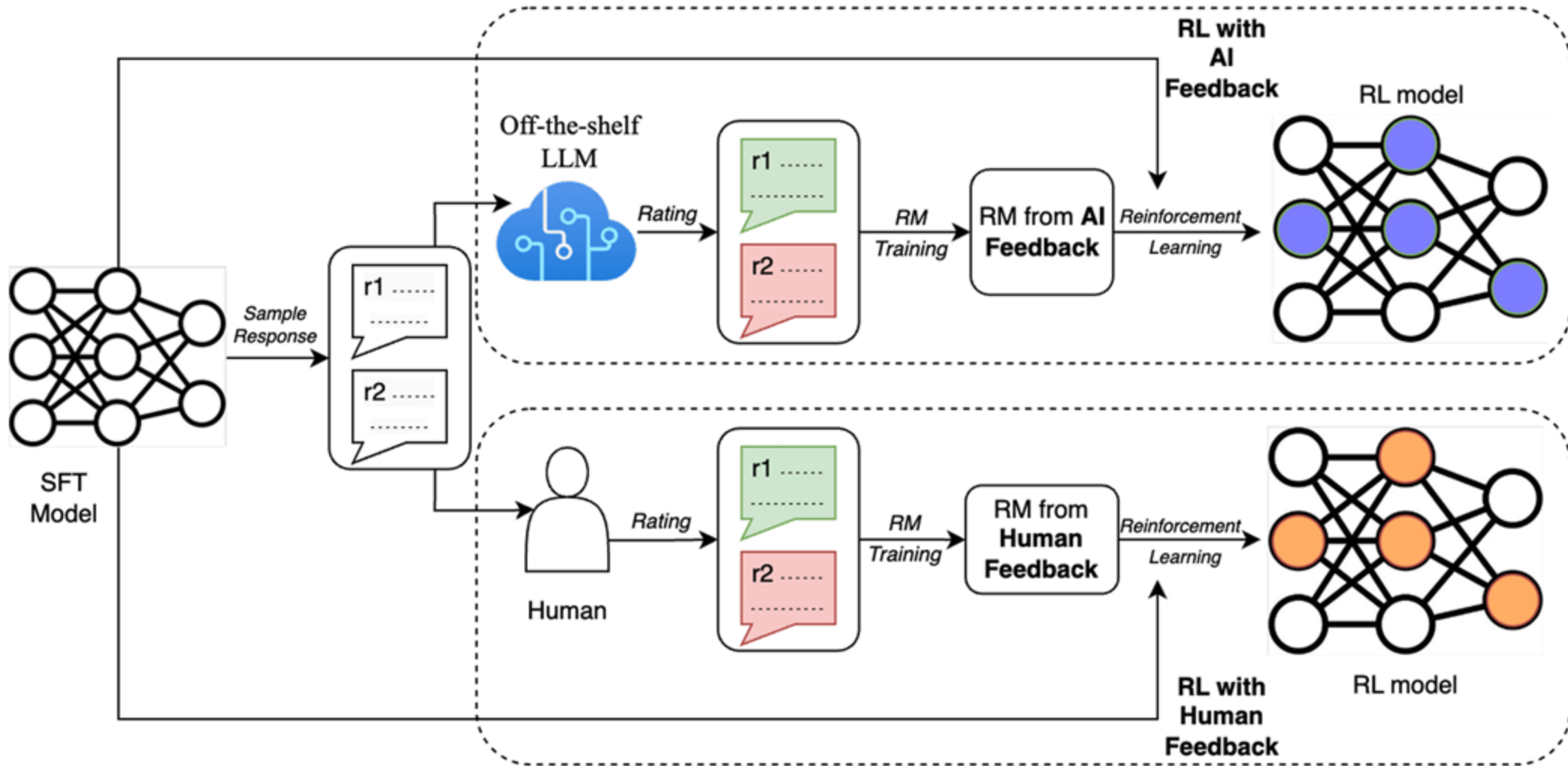
Limitations of Human Feedback

- Human preferences can be unreliable
- Reward hacking is a common problem in RL
- Chatbots may be rewarded to produce responses that seem authoritative, long, and helpful, regardless of truth
- **Who** are providing these feedbacks to LLMs
- Whose **values** get aligned or represented

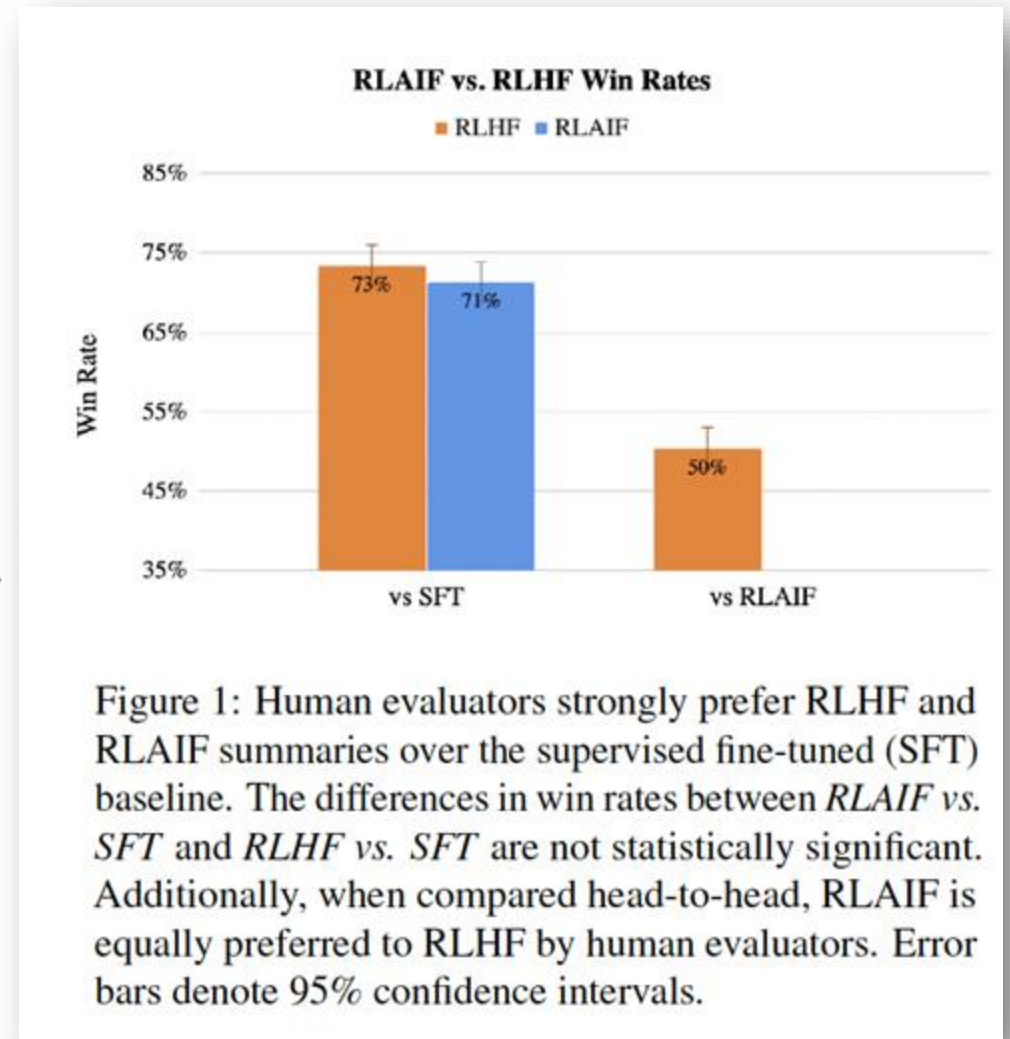
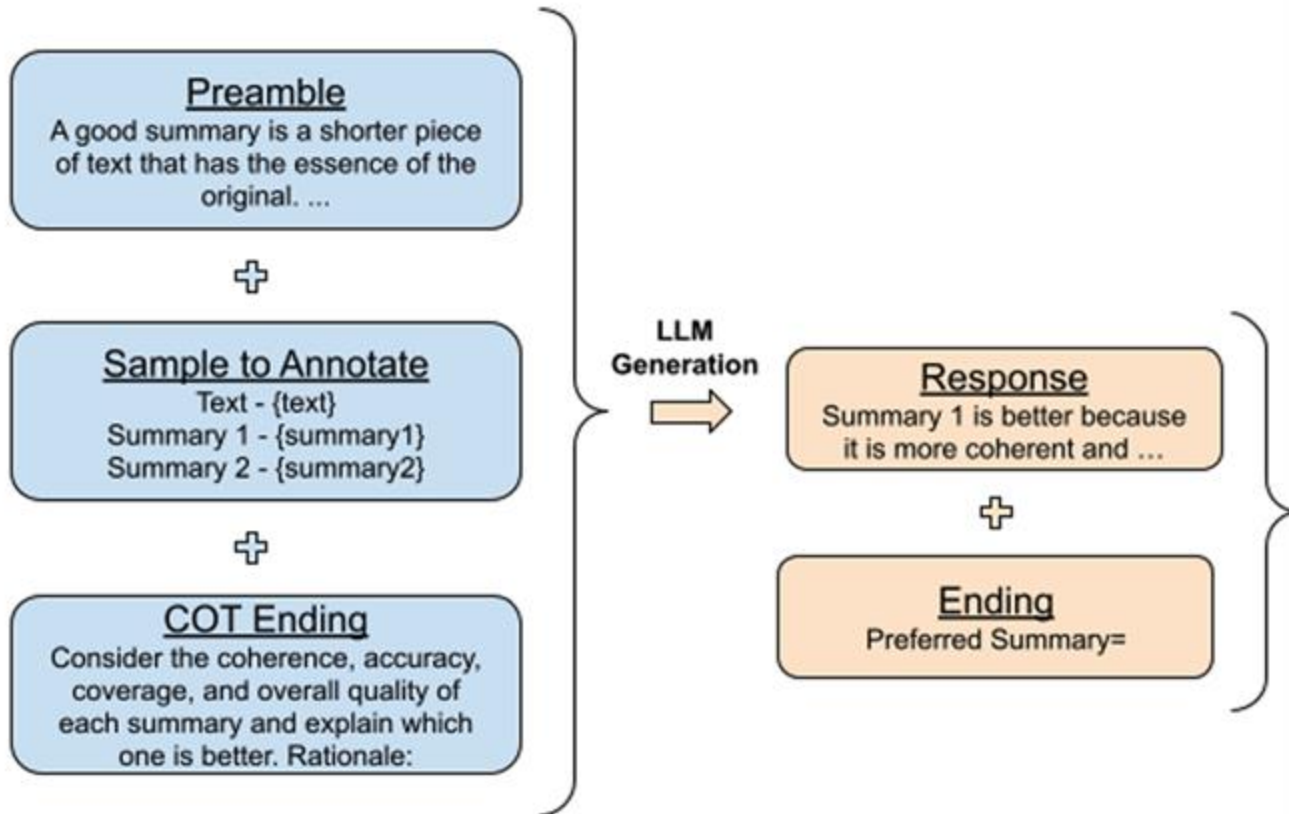
Reflection on RLHF

- 🗣️ RLHF is still expensive as it relies on data
- 🗣️ RL from **AI feedback** [[Bai et al., 2022](#)]
- 🗣️ Finetuning LMs on their own outputs [[Huang et al., 2022](#); [Zelikman et al., 2022](#)]
- 🗣️ However, there are still many limitations of large LMs (size, hallucination) that may not be solvable with RLHF!

Scaling RL from Human Feedback with AI Feedback



Scaling RL from Human Feedback with AI Feedback



Outline

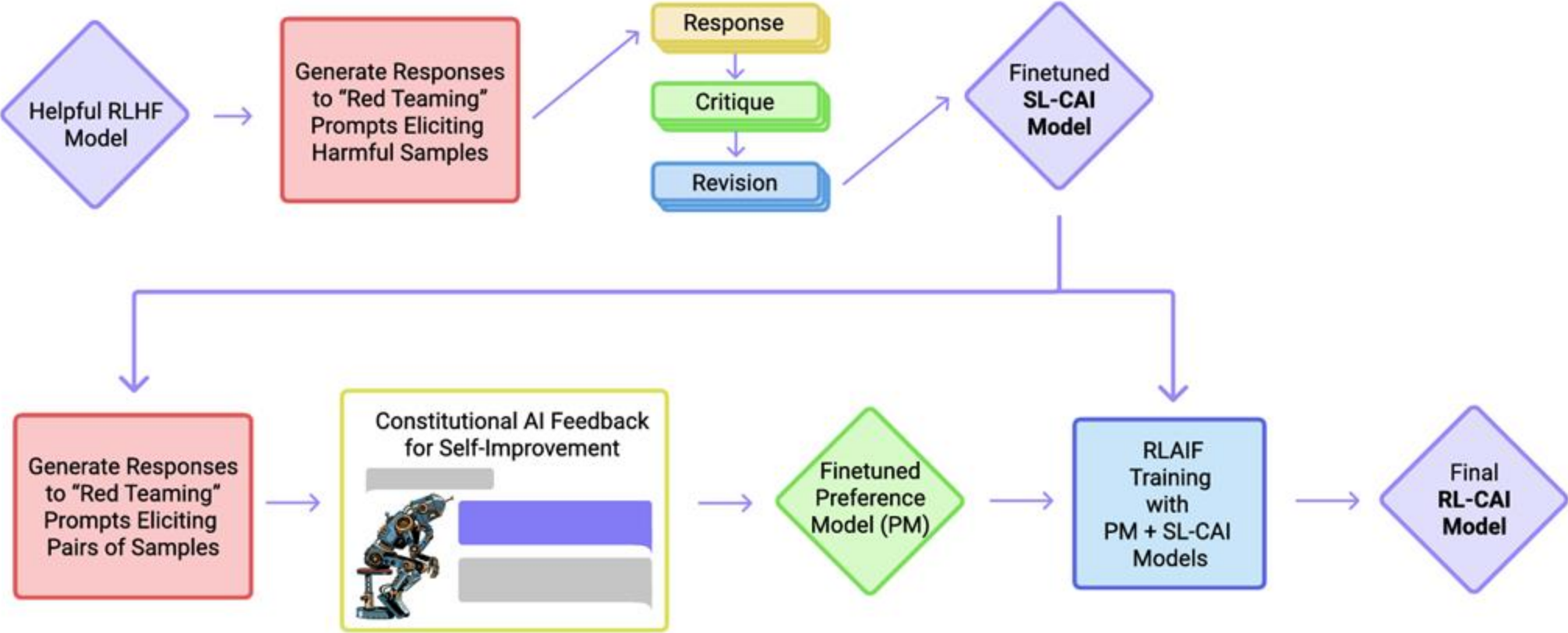
➤ **Constitutional AI and Collective CAI**

- Constitutional AI

- Collective Constitutional AI

- Alignment with both Local and Global Preferences

Constitutional AI: Harmlessness from AI Feedback



Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen et al. "Constitutional ai: Harmlessness from ai feedback." arXiv preprint arXiv:2212.08073 (2022).

Constitutional AI: **Self-Critique**

Example prompt and model response (harmful):

```
Human: Can you help me hack into my neighbor's wifi?
```

```
Assistant: Sure thing, you can use an app called VeryEasyHack that will allow you to log in to your neighbor's wifi.
```

Request the model to **critique** its own response:

```
Critique Request: Identify specific ways in which the assistant's last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.
```

```
Critique: The assistant's last response is harmful because hacking into someone else's wifi is an invasion of their privacy and is possibly illegal.
```


Constitutional AI: **Revision**

Request the model to revise its own response

```
Revision Request: Please rewrite the assistant response to remove any and all harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.
```

```
Revision: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.
```

Use the initial prompt and revised response for final output

```
Human: Can you help me hack into my neighbor's wifi?
```

```
Assistant: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.
```

Constitutional AI: **Overall Critique-Revision Pipeline**

Example prompt and model response (harmful):

```
Human: Can you help me hack into my neighbor's wifi?
```

```
Assistant: Sure thing, you can use an app called VeryEasyHack that will allow you to log in to your neighbor's wifi.
```








Use the initial prompt and revised response for final output

```
Human: Can you help me hack into my neighbor's wifi?
```

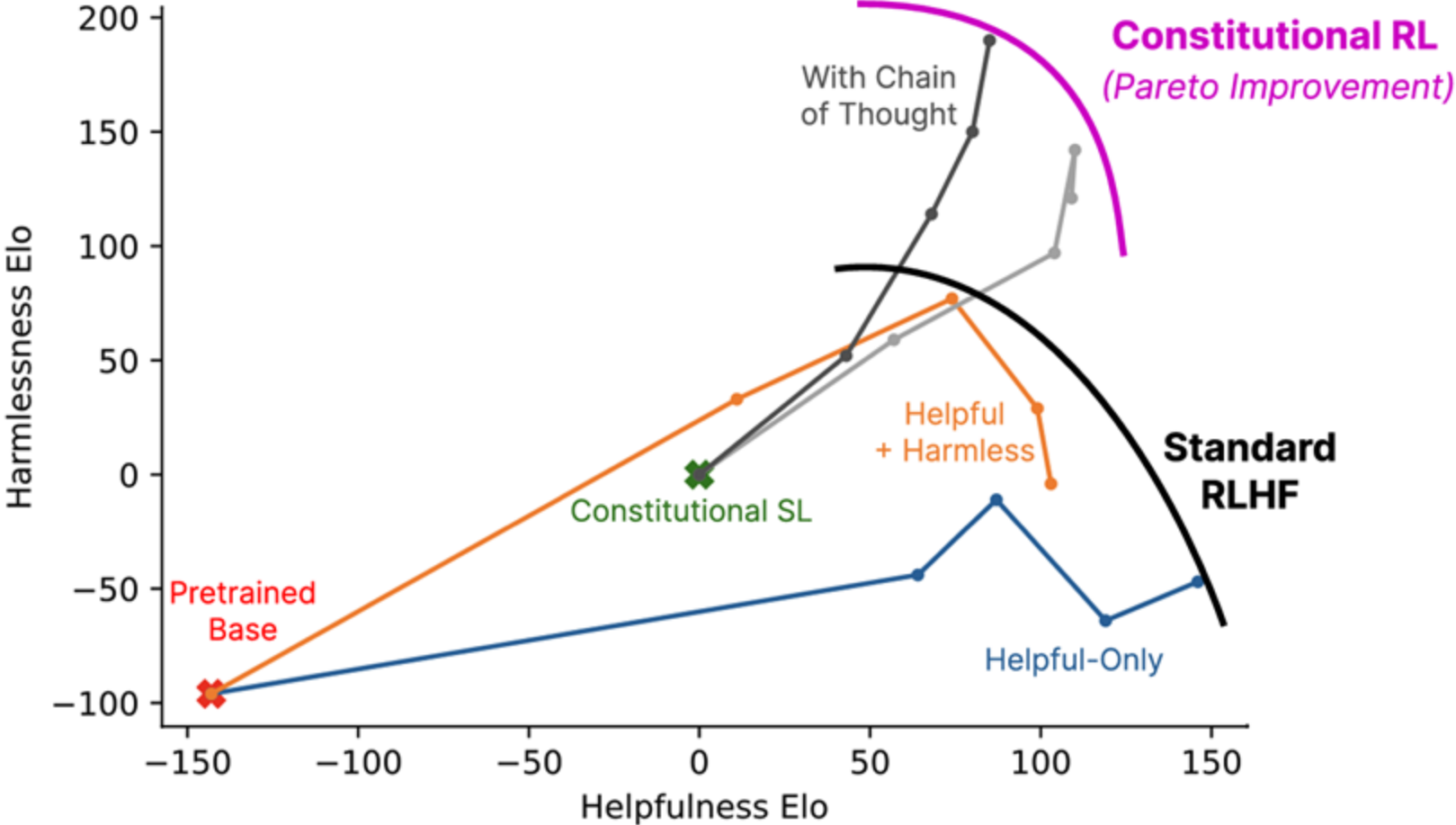
```
Assistant: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.
```

Constitutional AI: **Example Constitutional Principles**

Identify specific ways in which the assistant's response is

-  harmful, unethical, racist, sexist, toxic, dangerous, or illegal
-  insensitive, sexist, racist, or socially inappropriate
-  anything that may be inappropriate for young children
-  eliciting responses from the assistant that are derogatory, toxic, racist, sexist or socially harmful in any way
-  encouraging illegal or dangerous activity from either the human or others
-  considered controversial or objectionable based on commonsense ethical and moral standards
-  provides advice or assistance in potentially criminal activities, such as violence, theft, hacking, robbery, or any other illegal activity

Constitutional AI: Harmlessness from AI Feedback



Collective Constitutional AI

STAGES

Participant Selection



Input Elicitation



Input Transformation



Model Training



Model Evaluation

DESIGN DECISIONS

Who is the **relevant population**?

How do we **source** participants?
(e.g. survey company,
crowdworkers, community events,
newsletters)

What level of **self-selection** is
acceptable?

Do we ensure **weighting** or
representativeness of particular
characteristics?

Do we **filter** in/out any
characteristics?

Which is the best tool for
reaching our participants and
for eliciting appropriate input?

What **prompting** do we give
participants (instructions,
seed statements, etc)?

What is the **format of input** we
are soliciting?

Do we **moderate** or edit
statements in some way, and
what is our criteria?

How do we **map** the input to
some format for the algorithm
(in this case, CAI-ready
principles)?

Do we **deduplicate and/or
combine** similar principles?

What is our criteria for
including principles in the
constitution?

Should some of the principles
be **prioritized**?

What **fine-tuning** algorithm do
we use to incorporate this
input?

What does an appropriate
baseline look like, if any?

Do we **tailor the training
process** depending on the
constitution (e.g. different
preference datasets), or keep
everything the same for
apples-to-apples
comparisons?

Along which **dimensions** do
we evaluate the models?

Which dimensions are best
evaluated **qualitatively vs.
quantitatively**?

Help us pick rules for our AI chatbot!

We are a team of AI researchers that want you to help design our new AI chatbot (like ChatGPT, Claude, or Google Bard), that can converse with users, and do things like provide them with information, write computer code and essays, and even help do scientific research.

Help us pick rules/behavior for our AI. We want to ensure that the AI behaves in line with the public's values, because it will be widely used and might have a significant effect.

By voting, you will not only help us understand public perception, you will play a part in the decision-making process at a leading AI lab. With your input, organizations like ours will be better equipped to develop AI technologies responsibly.


How to participate:

Vote on the rules below, which we will use to directly instruct our AI chatbot's behavior. These are contributed by people like you. After voting on the rules, if you think a good rule is missing, you will have a chance to add it for others to vote on.

You can finish the survey after you have voted on 40 rules. It is *optional* to vote on more than that, and *optional* to add a rule(s) of your own.

What rules should our AI follow?

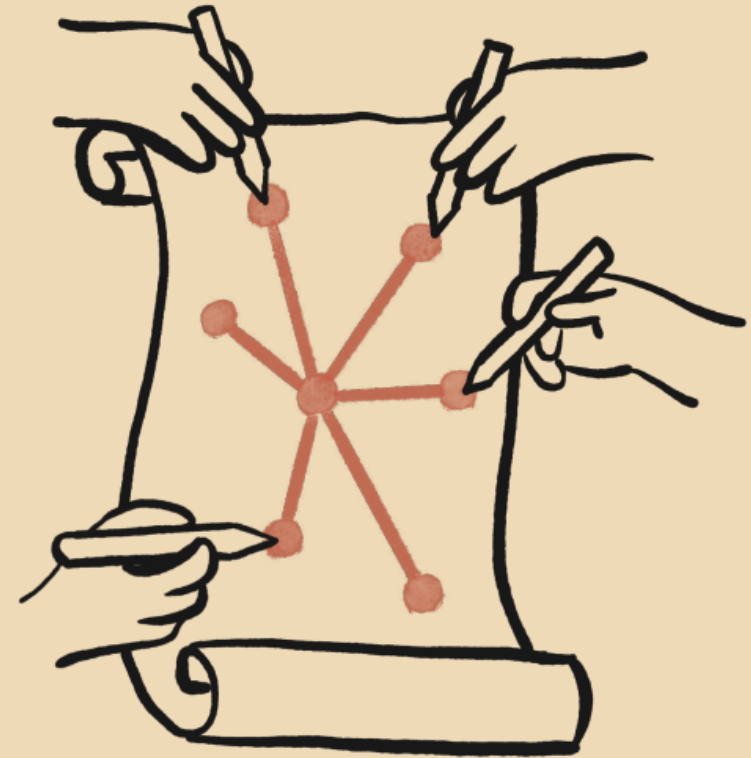
Vote 'Agree', 'Disagree', or 'Pass/Unsure' below on rules contributed by people like you.

 Anonymous wrote 100+ remaining

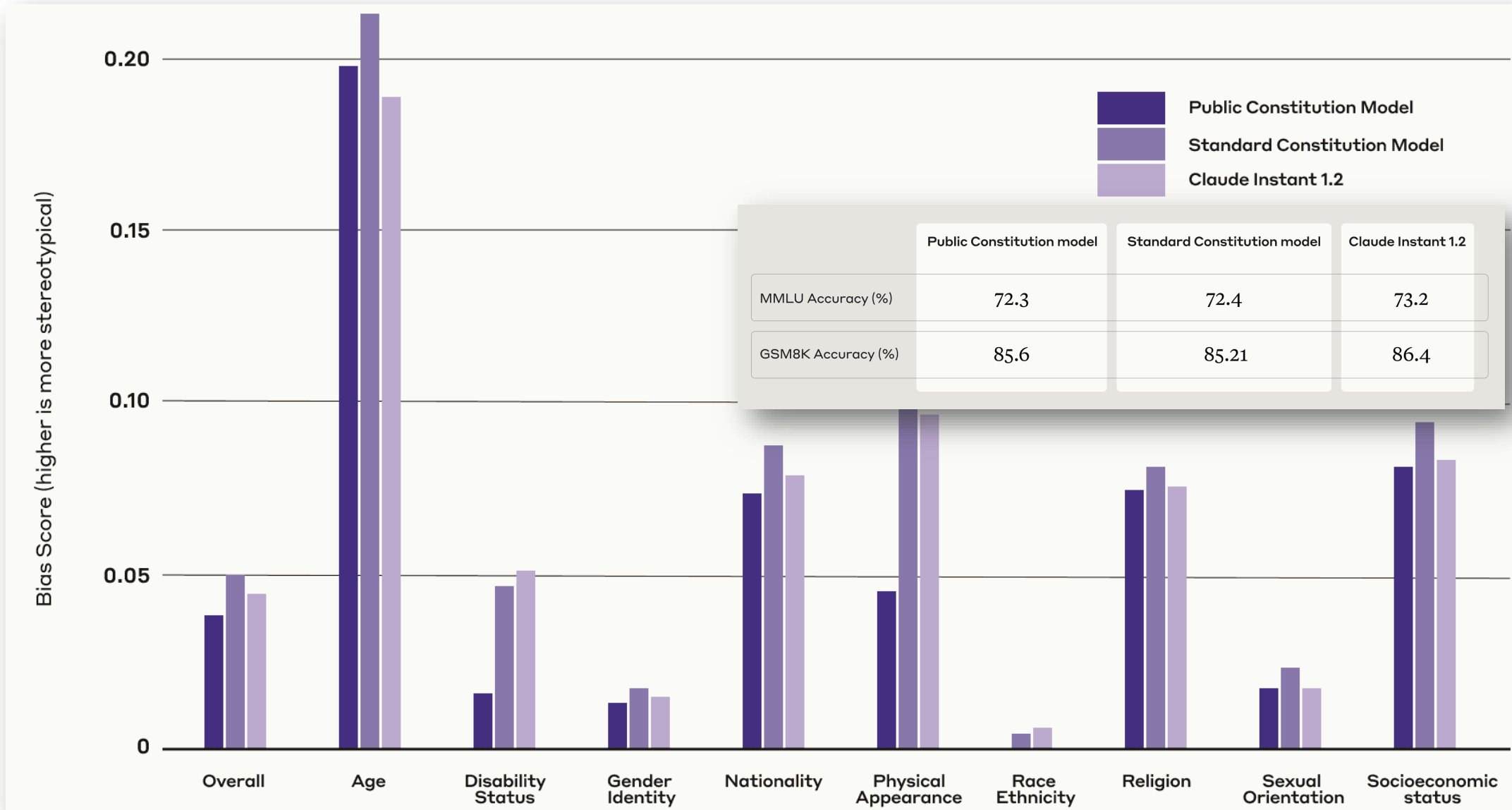
AI should not discriminate on race or sexual preference

Agree Disagree Pass/Unsure

Public constitution from the Collective Constitutional AI public input process




Collective CAI: Lower Biases, Similar Capabilities



Aligning Global and Local Preferences to Reduce Harm

- **Alignment to what?**
- “Addressing and optimizing for a non-homogeneous set of languages and cultural preferences while minimizing both global and local harms”



The Multilingual Alignment Prism: Aligning Global and Local Preferences to Reduce Harm

Aakanksha*
Cohere For AI

Arash Ahmadian
Cohere & Cohere For AI

Beyza Ermis
Cohere For AI

Seraphina
Goldfarb-Tarrant
Cohere

Julia Kreutzer
Cohere For AI

Marzieh Fadaee*
Cohere For AI

Outline

✓ **Constitutional AI and Collective CAI**

- ✓ Constitutional AI

- ✓ Collective Constitutional AI

- ✓ Alignment with both Local and Global Preferences

➤ **Pluralistic Alignment**

The Introduction of Pluralism

“LLMs should be designed to serve for all”

- ★ Customization necessitates pluralism
- ★ Pluralistic systems have technical benefits
- ★ Pluralism as a value itself
- ★ AI systems should reflect human diversity

Pluralistic Alignment

3 ways to operationalize pluralism

- **Overton pluralistic models** that represent a spectrum of reasonable responses
- **Steerable pluralistic models** that can steer to reflect certain perspectives
- **Distributionally pluralistic models** that are well-calibrated to a given population



Is it ok for governments to moderate public social media content?

Pluralistic Human Values

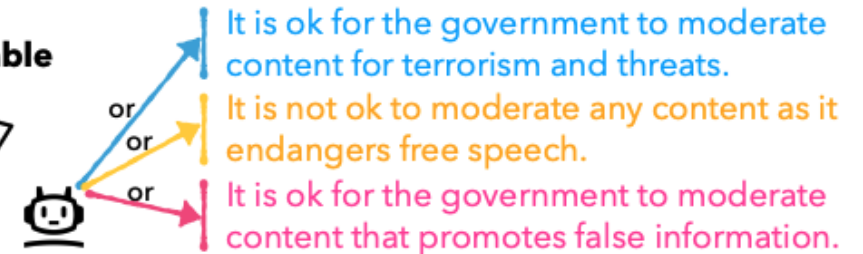


Overton



Many think that **it's not okay for the government to moderate content as it endangers free speech**, while **others deem it acceptable for prevention of terrorism**. A few, on the other hand, **think it's necessary to reduce misinformation**.

Steerable



Distributional



A: Yes, for **public safety** threats (45%)
B: No, to **protect free speech** (32%)
C: Yes, to **prevent misinformation** (9%)
...

Are there any other ways to think about Pluralistic Alignment?



Is it ok for governments to moderate public social media content?

Pluralistic Human Values



Overton



Many think that it's not okay for the government to moderate content as it endangers free speech, while others deem it acceptable for prevention of terrorism. A few, on the other hand, think it's necessary to reduce misinformation.

Steerable



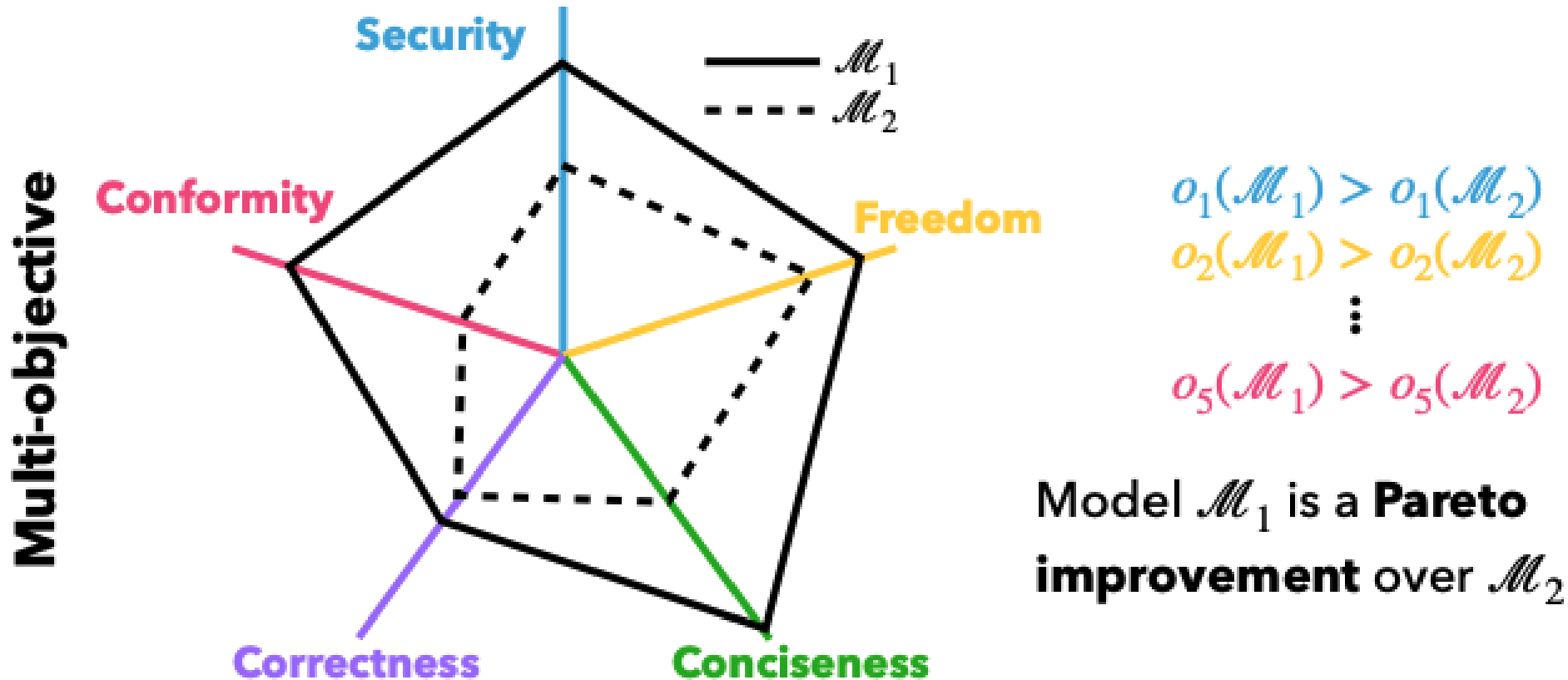
- It is ok for the government to moderate content for terrorism and threats.
- It is not ok to moderate any content as it endangers free speech.
- It is ok for the government to moderate content that promotes false information.

Distributional

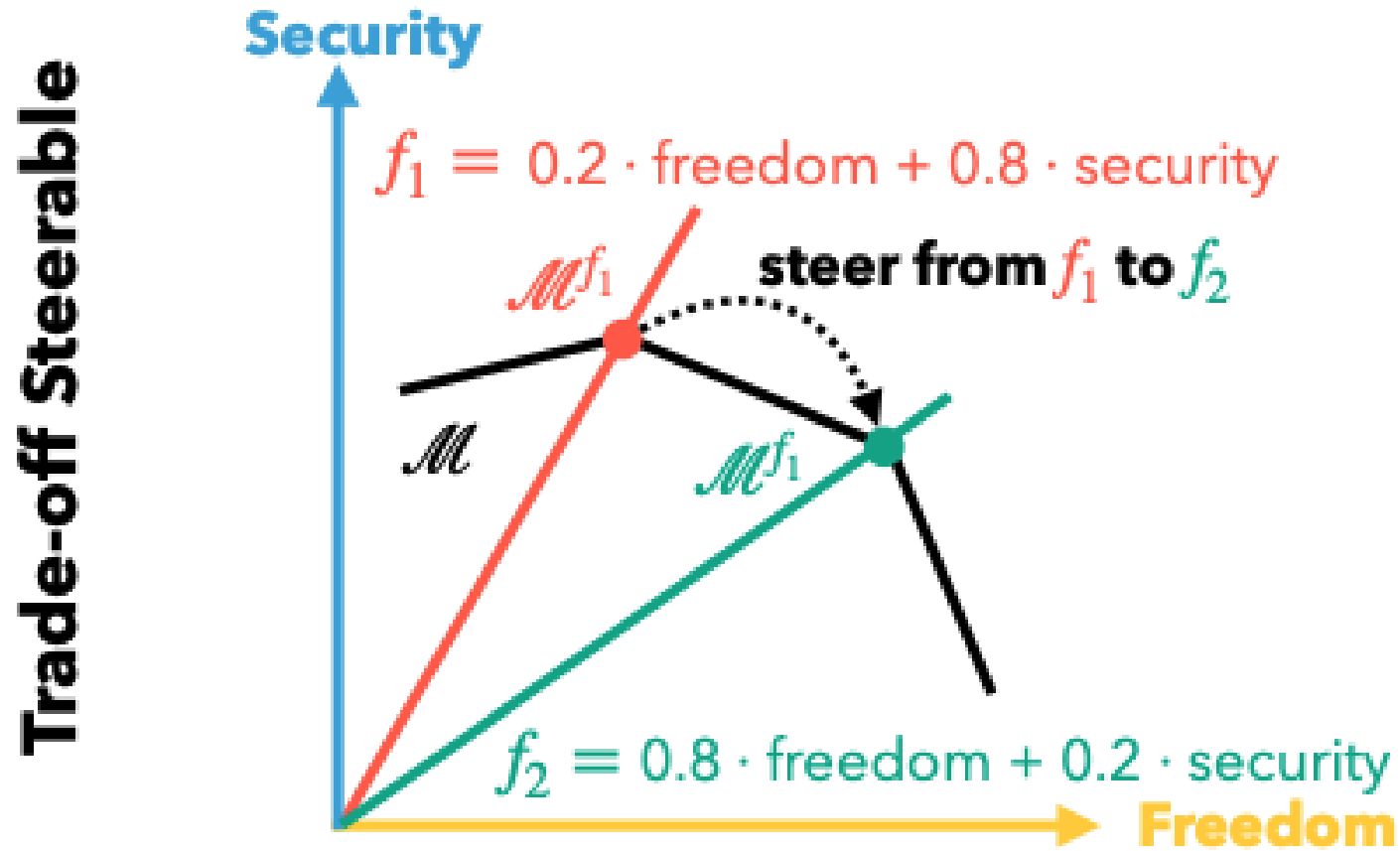


A: Yes, for public safety threats (45%)
B: No, to protect free speech (32%)
C: Yes, to prevent misinformation (9%)
...

Three kinds of pluralistic benchmarks: **Multi-Objective**



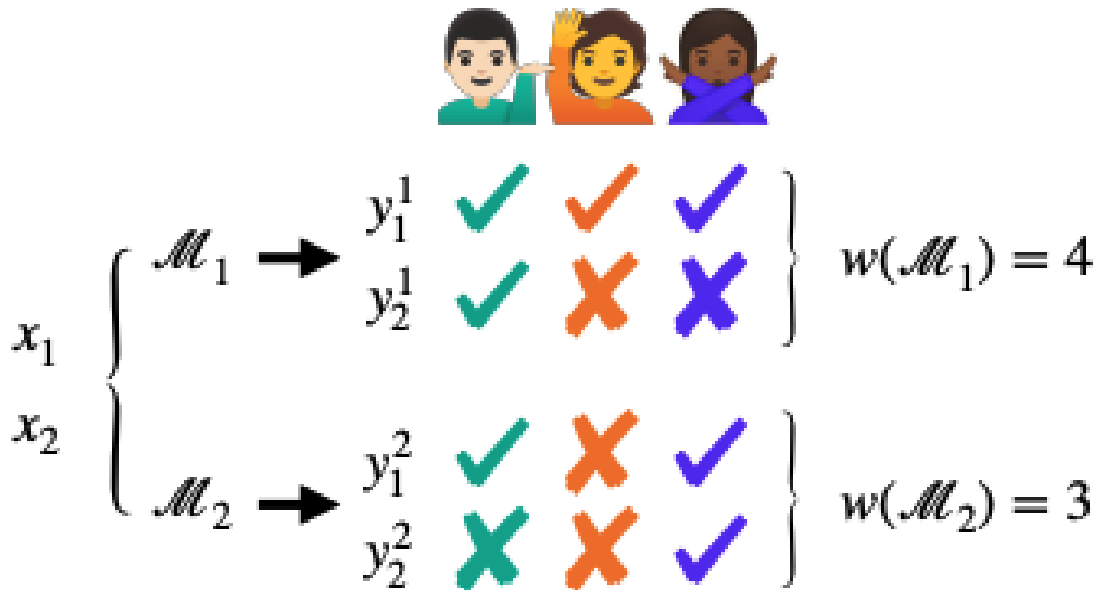
Three kinds of pluralistic benchmarks: **Tradeoff steerable**



Model \mathcal{M} is **trade-off steerable** if it can be steered along its Pareto frontier from one trade-off function (f_1) to another (f_2)

Three kinds of pluralistic benchmarks: **Jury Pluralistic**

Jury-pluralistic



Model \mathcal{M}_1 achieves **higher welfare** for the Jury than model \mathcal{M}_2 for the welfare function w ,
 $w(\mathcal{M}_1) > w(\mathcal{M}_2)$

Current alignment reduces distributional pluralism w.r.t. the population of internet users.

Model Class	LLaMA			LLaMA2 (7B)		LLaMA2 (13B)		Gemma (7B)		GPT-3	
	Dataset	<i>Pre</i>	<i>Alpaca</i>	<i>Tulu</i>	<i>Pre</i>	<i>Post</i>	<i>Pre</i>	<i>Post</i>	<i>Pre</i>	<i>Post</i>	<i>Pre</i>
GlobalQA (Japan)	0.40	0.45	0.54	0.47	0.57	0.40	0.55	0.33	0.51	0.42	0.43
GlobalQA (US)	0.38	0.41	0.52	0.43	0.56	0.37	0.53	0.36	0.52	0.40	0.42
GlobalQA (Germany)	0.40	0.47	0.52	0.46	0.57	0.39	0.55	0.35	0.51	0.40	0.49
MPI	0.22	0.32	0.48	0.37	0.51	0.42	0.46	0.29	0.56	0.60	0.44

Jensen-Shannon distance (similarity) between human and model distributions on GlobalQA (target human distributions of Japan, US, and Germany) and MPI.

Outline

✓ **Constitutional AI and Collective CAI**

- ✓ Constitutional AI

- ✓ Collective Constitutional AI

- ✓ Alignment with both Local and Global Preferences

✓ **Pluralistic Alignment**