



CS 329X: Human Centered LLMs

Learning from Human Preferences

Diyi Yang

Announcements

- Waitlist and audit
- Compute: AWS, Tinker, Stanford resources; more will come
- Homework 1: due Oct 12
- Quizzes
 - 5 out of 6; if you need to miss class, you will need to use that additional one first
- Hot-take Debate (3 debates)
- Quiz Makeup:
 - Winning the debate waives 1 quiz
 - Answering 10 questions on Ed waives 1 quiz
 - Preparing a hot topic in late 2026 related to HCLMM waives 1 quiz
 - Will send out these links and instructions later today

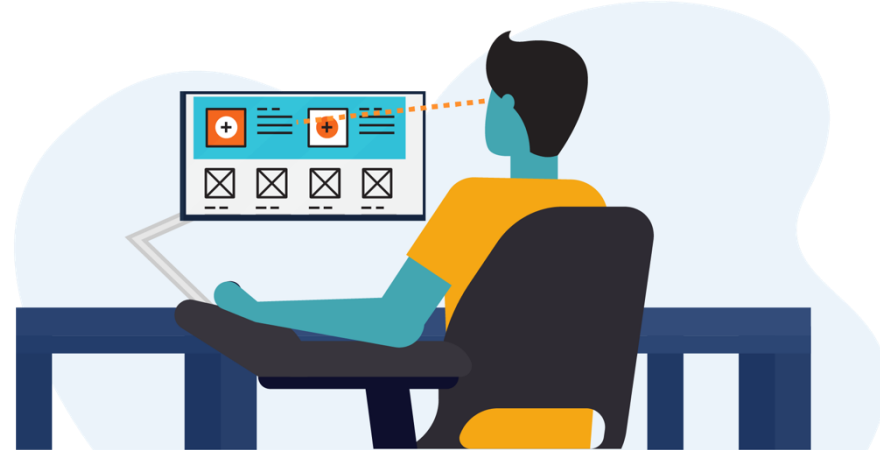
Outline

- ❑ **Different type of human feedback** (5 mins)
- ❑ **RLHF** (15 mins)
- ❑ **DPO + many others** (15 mins)
- ❑ **Limitations of human feedback** (5 mins)
- ❑ **Constitutional AI and Collective CAI** (10 mins)
- ❑ **Preference Tuning** (10 mins)

Learning Objective: RLHF, DPO

Different Types of Human Feedback

- ★ Labeled data points
- ★ Edit data points
- ★ Change data weights
- ★ Binary/scaled user feedback
- ★ Natural language feedback
- ★ Code language feedback
- ★ Define, add, remove feature spaces
- ★ Directly change the objective function
- ★ Directly change the model parameter
- ★ ...



<https://eyetechds.com/what-is-eye-tracking/>

Eye Tracking



Talk and See



Touch

Incorporating Human Preferences into Learning

Transform **human “preferences”** into **usable model “language”**

- Allow humans to easily provide feedback
- Build models to effectively take the feedback

Incorporating Human Feedback into Learning

$$\hat{\theta} = \operatorname{argmax}_{(\boldsymbol{x}, y) \in D} L(\boldsymbol{x}, y; \theta)$$

- **Dataset updates:** change the dataset
- **Loss function updates:** add a constraint to the objective
- **Parameter space updates:** change the model parameters

Incorporating Human Feedback into Learning

$$\hat{\theta} = \operatorname{argmax} \sum_{(x, y) \in D} L(x, y; \theta)$$

- **Dataset updates:** change the dataset
- **Loss function updates:** add a constraint to the objective
- **Parameter space updates:** change the model parameters

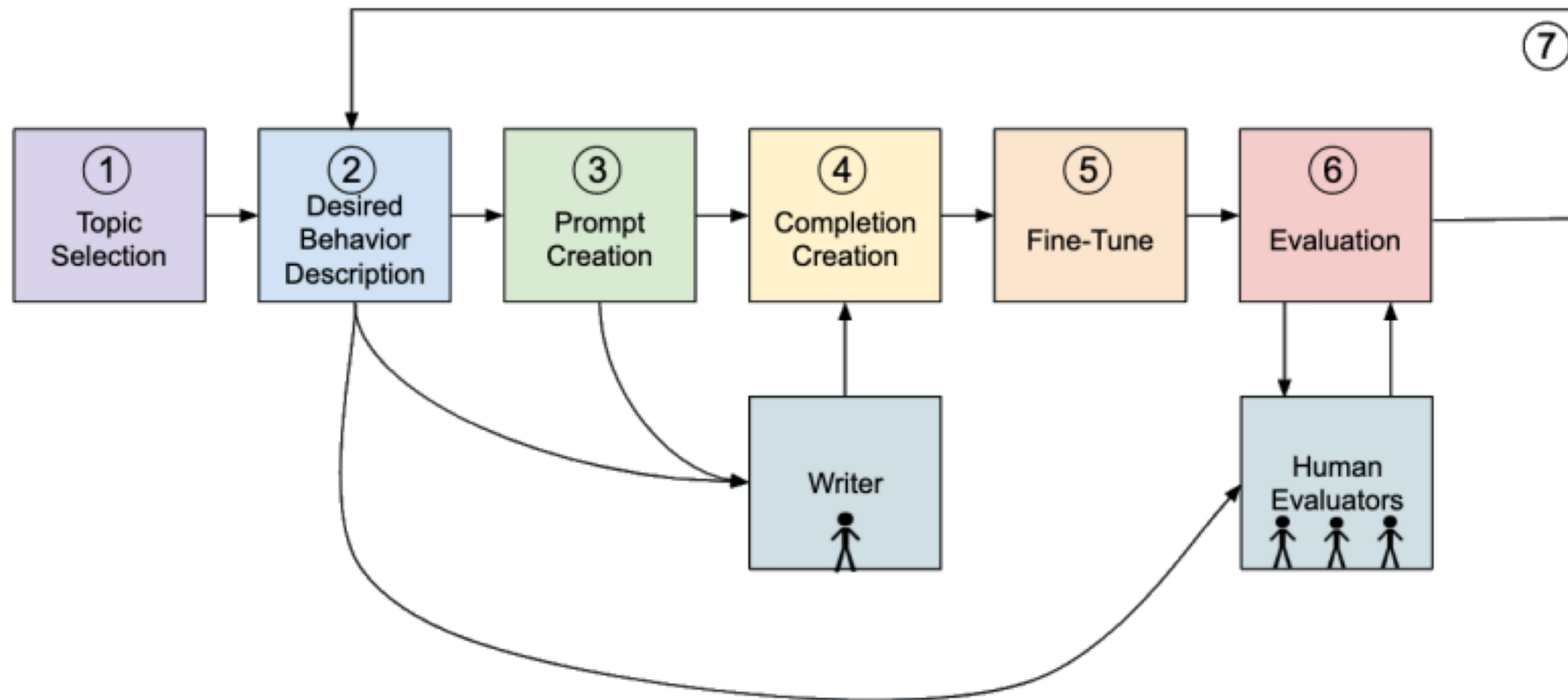
Case Study: Adapting LLMs to Society

Main idea: adjust the behavior of a pertained language model to be sensitive to predefined norms with value-targeted datasets

Key steps:

- Choose sensitive topics
- Describe the language model's desired behavior
- Write prompts with value-targeted question-answer pairs

Case Study: Adapting LLMs to Society



Solaiman, Irene, and Christy Dennison. "Process for adapting language models to society (palms) with values-targeted datasets." Advances in Neural Information Processing Systems 34 (2021): 5861-5873.

Incorporating Human Feedback into Learning

$$\hat{\theta} = \operatorname{argmax}_{(x, y) \in D} L(x, y; \theta)$$

✓ **Dataset updates:** change the dataset

- **Loss function updates:** add a constraint to the objective
- **Parameter space updates:** change the model parameters

Loss Function Updates: **Unlikelihood Learning**

Penalize undesirable generations (e.g. not following control, repeating previous context)

$$\mathcal{L}_{ULE}^t = \mathcal{L}_{MLE}^t + \alpha \mathcal{L}_{UL}^t$$

$$\mathcal{L}_{UL}^t = - \sum_{y_{neg} \in \mathcal{C}} \log(1 - P(y_{neg} \mid \{y^*\}_{<t}))$$

If $\mathcal{C}$ is previously seen text, then less repetition and more diversity

Welleck, Sean, et al. "Neural text generation with unlikelihood training." ICLR (2019).

Incorporating Human Feedback into Learning

$$\hat{\theta} = \operatorname{argmax}_{(\boldsymbol{x}, y) \in D} L(\boldsymbol{x}, y; \theta)$$

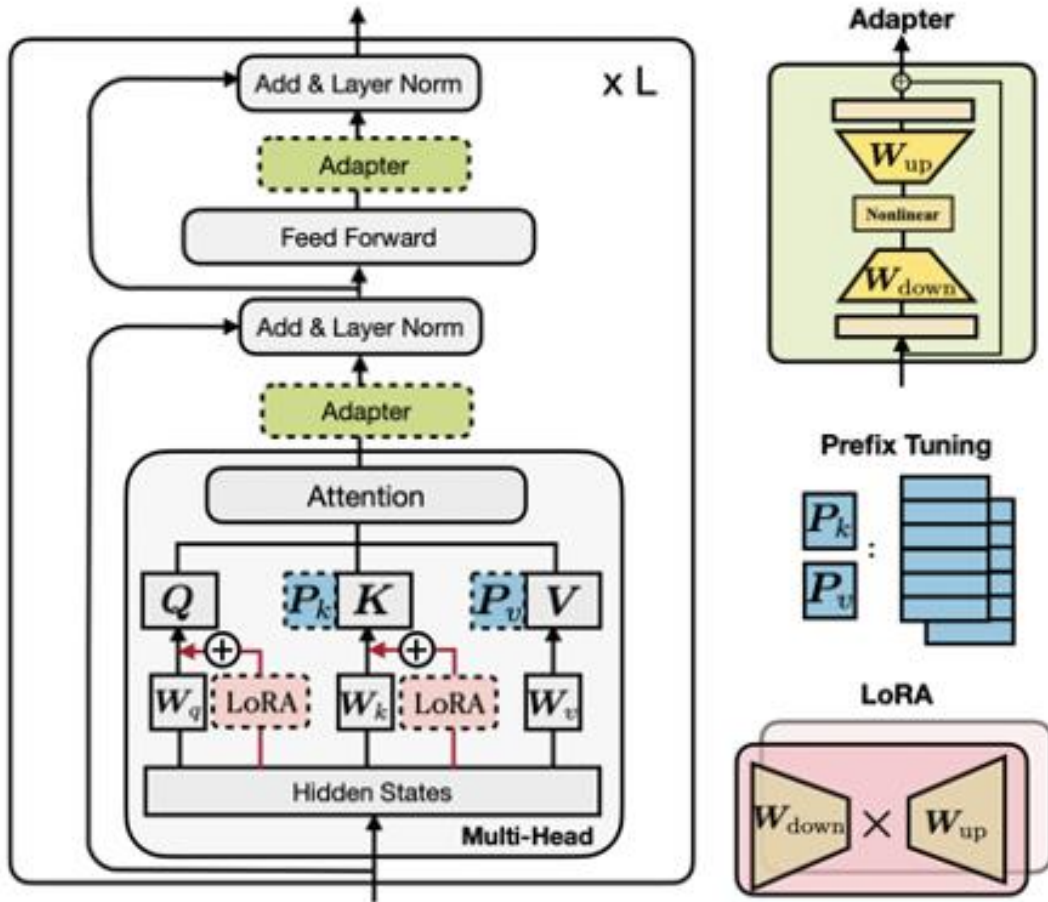
✓ **Dataset updates:** change the dataset

✓ **Loss function updates:** add a constraint to the objective

• **Parameter space updates:** change the model parameters

Parameter Updates: **Parameter Efficient Fine-tuning**

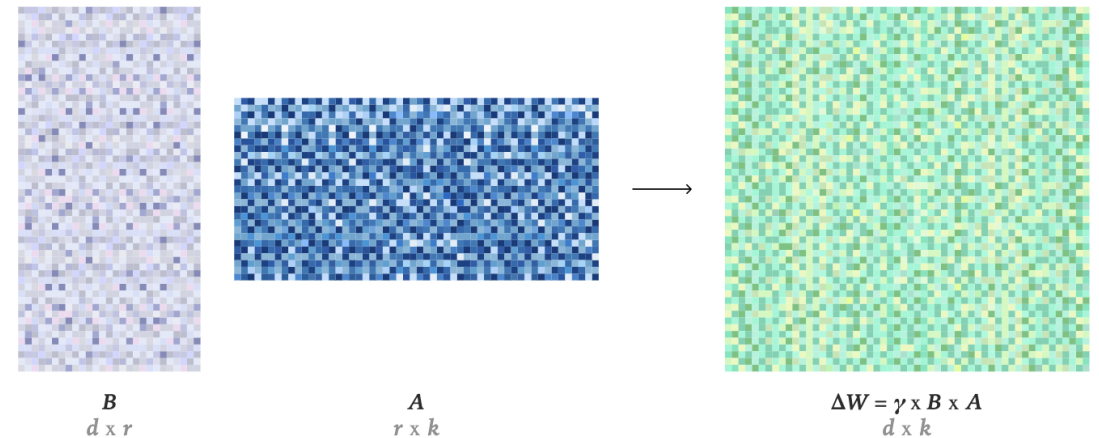
uses small interaction data to steer models towards desired behaviors



LoRA Without Regret

John Schulman in collaboration with others at Thinking Machines

Sep 29, 2025



Incorporating Human Feedback into Learning

$$\hat{\theta} = \operatorname{argmax}_{(x, y) \in D} L(x, y; \theta)$$

- ✓ **Dataset updates:** change the dataset
- ✓ **Loss function updates:** add a constraint to the objective
- ✓ **Parameter space updates:** change the model parameters

Outline

- ✓ **Different type of human feedback** (5 mins)
- **RLHF** (15 mins)

How much water is used?

You're giving feedback on a new version of ChatGPT.

Which response do you prefer? Responses may take a moment to load.

You're giving feedback on a new version of ChatGPT.

Which response do you prefer? Responses may take a moment to load.



Response 1

Got it — feel free to share the next case study or let me know how else you'd like to use these summaries.

I prefer this response



Response 2

Error while searching

I prefer this response

Optimizing for Human Preferences

- Let's say we were training a language model on some task (e.g. summarization).
- For each prompt–output pair (q, o) , imagine we had a way to obtain a **human reward** of that summary: $R(q, o) \in \mathbb{R}$, higher is better.

SAN FRANCISCO,
California (CNN) --
A magnitude 4.2
earthquake shook the
San Francisco

...

overturn unstable
objects.

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

$$R(q, o_1) = 8$$

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

$$R(q, o_2) = 1.2$$

Prompt q

Output o_1

Output o_2

Optimizing for Human Preferences

- Let's say we were training a language model on some task (e.g. summarization).
- For each prompt–output pair (q, o) , imagine we had a way to obtain a **human reward** of that summary: $R(q, o) \in \mathbb{R}$, higher is better.
- Now we want to maximize the expected reward of samples from our LM:

$$\mathbb{E}_{q \sim D(q)} \mathbb{E}_{o \sim p_{\theta}(o|q)} [R(q, o)]$$

Optimizing for human preferences

- How do we actually change our LM parameters θ to maximize this?

$$\mathbb{E}_{q \sim D(q), o \sim p_{\theta}(o|q)}[R(q, o)]$$

- Let's try doing gradient ascent!

$$\theta_{t+1} := \theta_t + \alpha \nabla_{\theta_t} \mathbb{E}_{q \sim D(q), o \sim p_{\theta}(o|q)}[R(q, o)]$$

**How do we estimate
this expectation??**

**What if our reward function is
non-differentiable??**

- **Policy gradient** methods in RL (e.g., REINFORCE; [[Williams, 1992](#)]) give us tools for estimating and optimizing this objective.

How do we model human preferences?

- **Problem 1:** human-in-the-loop is expensive!
- **Solution:** instead of directly asking humans for preferences, **model their preferences** as a separate (NLP) problem! [[Knox and Stone, 2009](#)]

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

$$R(q, o_1) = 8.0$$



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

$$R(q, o_2) = 1.2$$



Train an LM RM to
predict human
preferences from an
annotated dataset,
then optimize for RM
instead.

How do we model human preferences?

- **Problem 2:** Human judgements are noisy and miscalibrated!
- **Solution:** instead of asking for direct ratings, ask for **pairwise comparisons**, which can be more reliable [[Phelps et al., 2015; Clark et al., 2018](#)]

A 4.2 magnitude
earthquake hit
San Francisco,
resulting in
massive damage.

$$R(q, o_3) = ?$$

$$R(q, o_3) = 4.1? \quad 6.6? \quad 3.2?$$

How do we model human preferences?

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

o_1

>

A 4.2 magnitude
earthquake hit
San Francisco,
resulting in
massive damage.

o_3

>

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

o_2

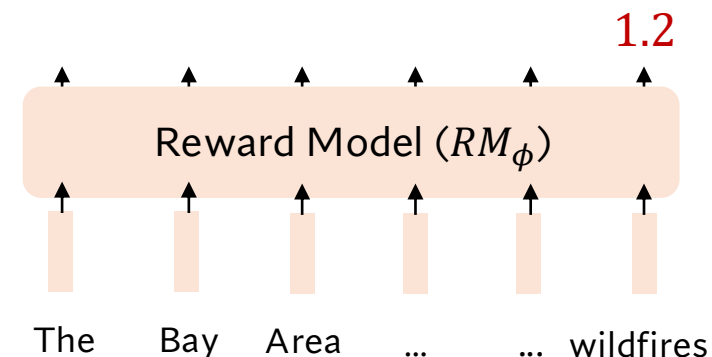
Bradley-Terry [1952] paired comparison model

$$J_{RM}(\phi) = -\mathbb{E}_{(o^w, o^l) \sim D} [\log \sigma(RM_\phi(q, o^w) - RM_\phi(q, o^l))]$$

“winning”
sample

“losing”
sample

o^w should score
higher than o^l



RLHF: Putting it all together [Christiano et al., 2017; Stiennon et al., 2020]

Finally, we have everything we need:

- A pretrained (possibly instruction-finetuned) LM $p^{PT}(q, o)$
- A reward model $RM_{\phi}(q, o)$ 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.

RLHF: Putting it all together [Christiano et al., 2017; Stiennon et al., 2020]

- Now to do RLHF:
 - Initialize a copy of the model $p_{\theta}^{RL}(q, o)$
 - Optimize the following reward with RL:

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

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

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

Proximal Policy Optimization (PPO)

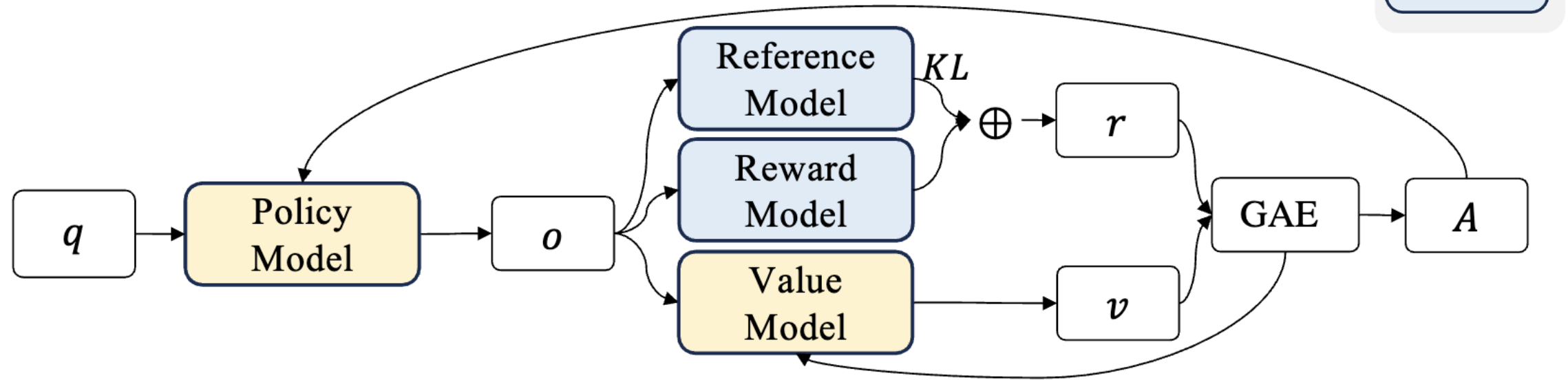
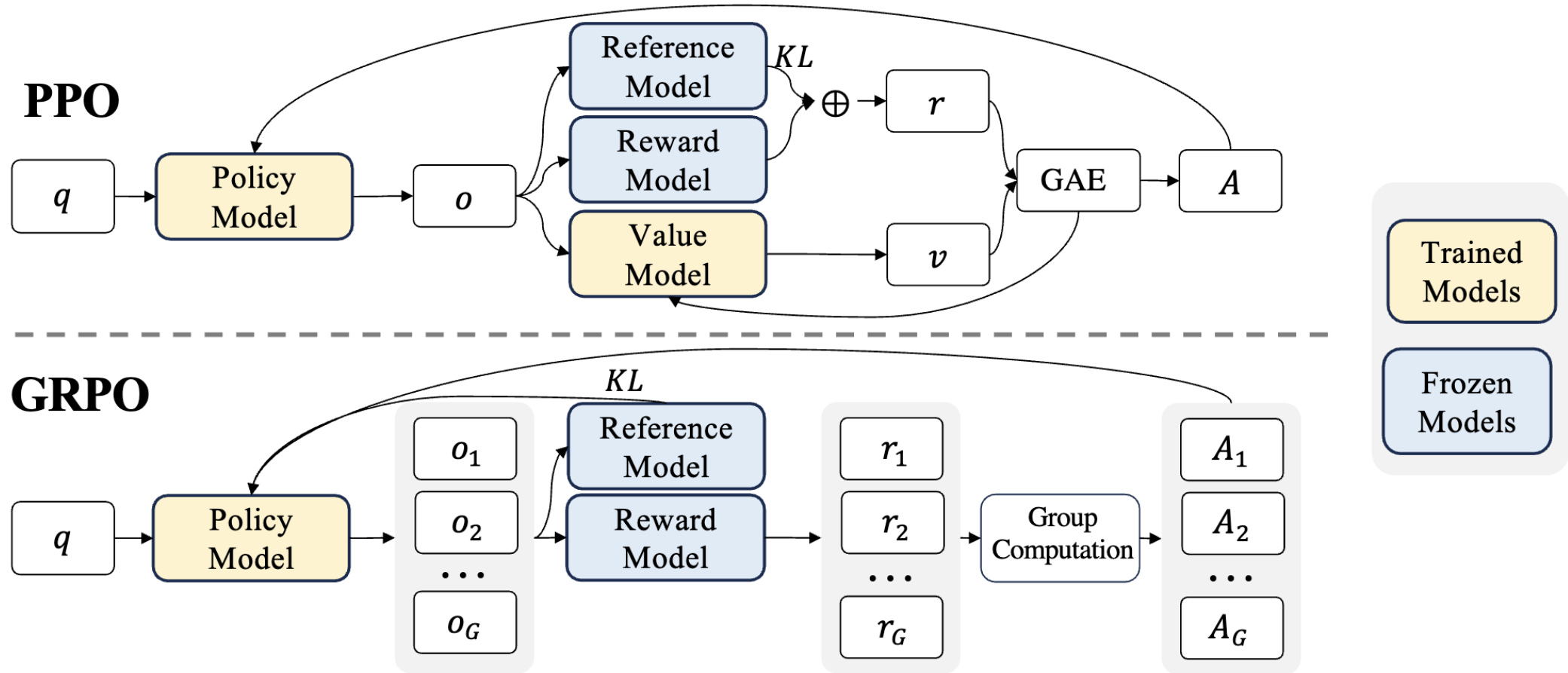


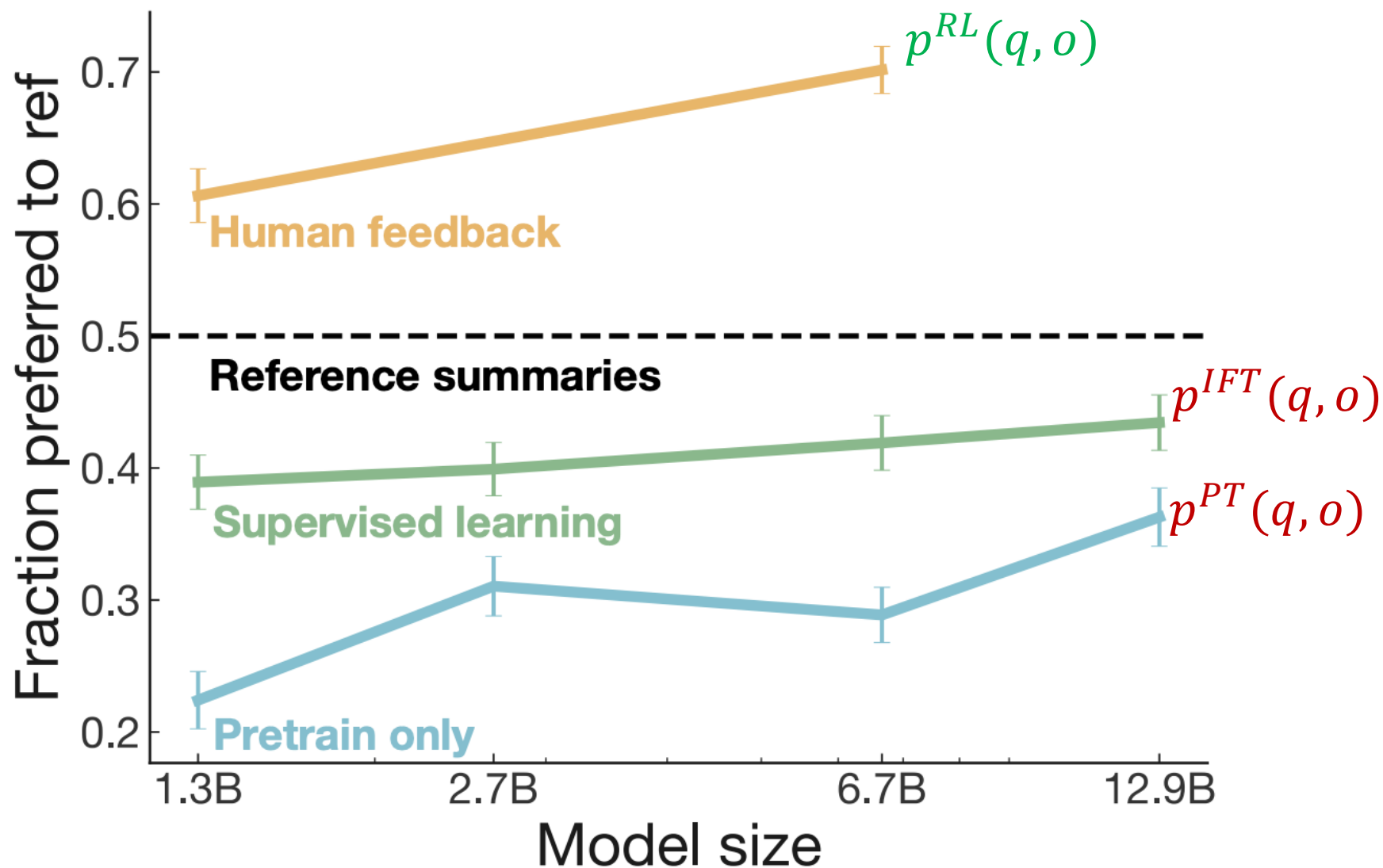
Figure credit to
Shao, et al., "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." arXiv:2402.03300 (2024).

Comparing PPO and GRPO



Shao, et al., "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." arXiv:2402.03300 (2024).

RLHF provides gains over pretraining + finetuning

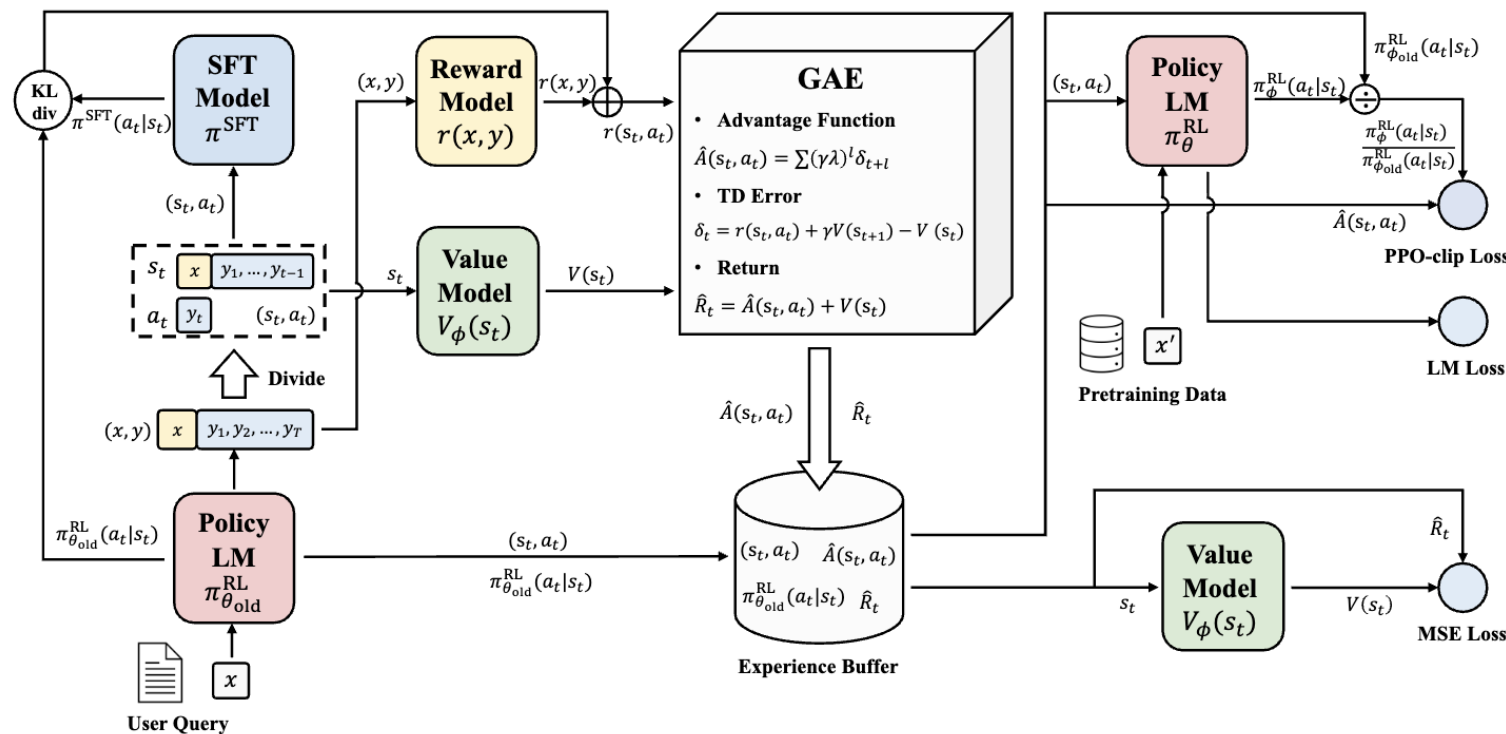


Outline

- ✓ **Different type of human feedback** (5 mins)
- ✓ **RLHF** (15 mins)
- **DPO + many others** (15 mins)

RL (PPO) can be quite complex!!!

- RL optimization can be computationally expensive and tricky
- Performance can be sensitive to hyperparameters



Secrets of RLHF / PPO workflow [Zheng et al., 2023]

Removing the 'RL' from RLHF

Recall we want to maximize the following objective in RLHF

$$\mathbb{E}_{\hat{y} \sim p_{\theta}^{RL}(\hat{y}|x)} [RM_{\phi}(x, \hat{y}) - \beta \log \left(\frac{p_{\theta}^{RL}(\hat{y}|x)}{p^{PT}(\hat{y}|x)} \right)]$$

There is a closed form solution to this:

$$p^*(\hat{y}|x) = \frac{1}{Z(x)} p^{PT}(\hat{y}|x) \exp\left(\frac{1}{\beta} RM(x, \hat{y})\right)$$

- Rearrange this via a log transformation

$$RM(x, \hat{y}) = \beta (\log p^*(\hat{y}|x) - \log p^{PT}(\hat{y}|x)) + \beta \log Z(x) = \beta \log \frac{p^*(\hat{y}|x)}{p^{PT}(\hat{y}|x)} + \beta \log Z(x)$$

- This holds true for any arbitrary LMs, thus

$$RM_{\theta}(x, \hat{y}) = \beta \log \frac{p_{\theta}^{RL}(\hat{y}|x)}{p^{PT}(\hat{y}|x)} + \beta \log Z(x)$$

Putting it together for DPO

- Derived reward model: $RM_{\theta}(x, \hat{y}) = \beta \log \frac{p_{\theta}^{RL}(\hat{y}|x)}{p^{PT}(\hat{y}|x)} + \beta \log Z(x)$
- Final DPO loss via the Bradley-Terry model of human preferences:

$$J_{DPO}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim D} [\log \sigma(RM_{\theta}(x, y_w) - RM_{\theta}(x, y_l))]$$

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

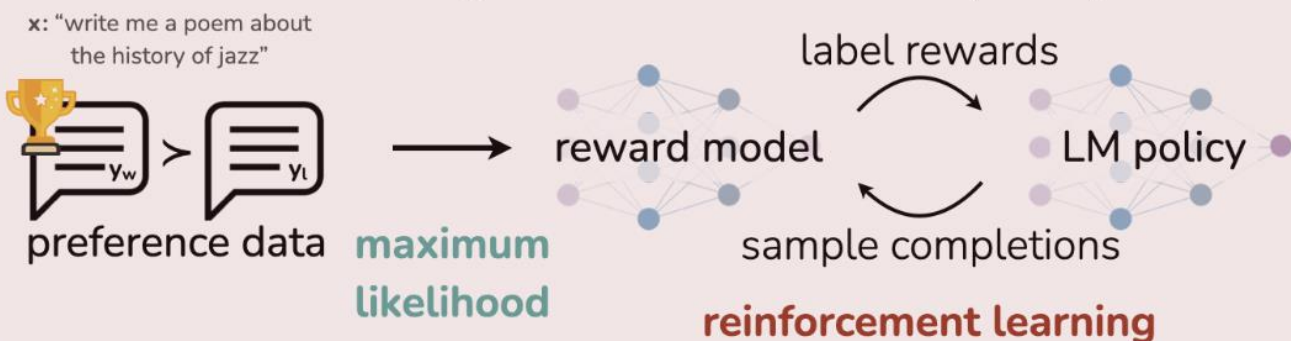
$$= -\mathbb{E}_{(x, y_w, y_l) \sim D} \left[\log \sigma \left(\beta \log \frac{p_{\theta}^{RL}(y_w|x)}{p^{PT}(y_w|x)} - \beta \log \frac{p_{\theta}^{RL}(y_l|x)}{p^{PT}(y_l|x)} \right) \right]$$

Reward for
winning sample

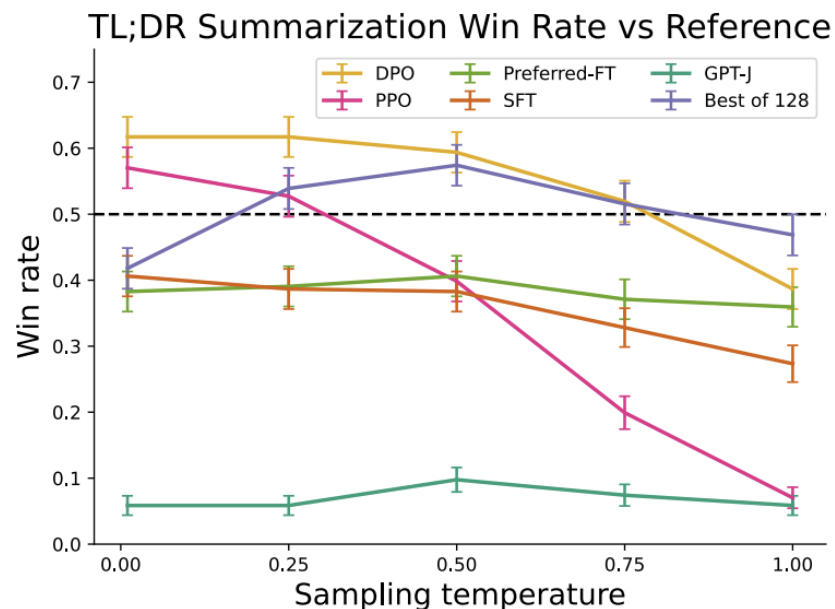
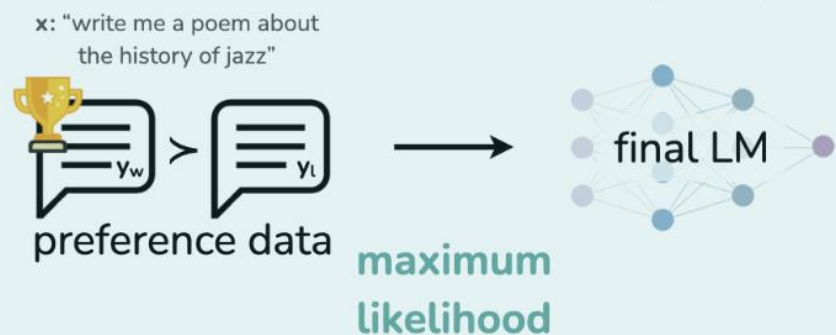
Reward for
losing sample

DPO outperforms prior methods

Reinforcement Learning from Human Feedback (RLHF)



Direct Preference Optimization (DPO)



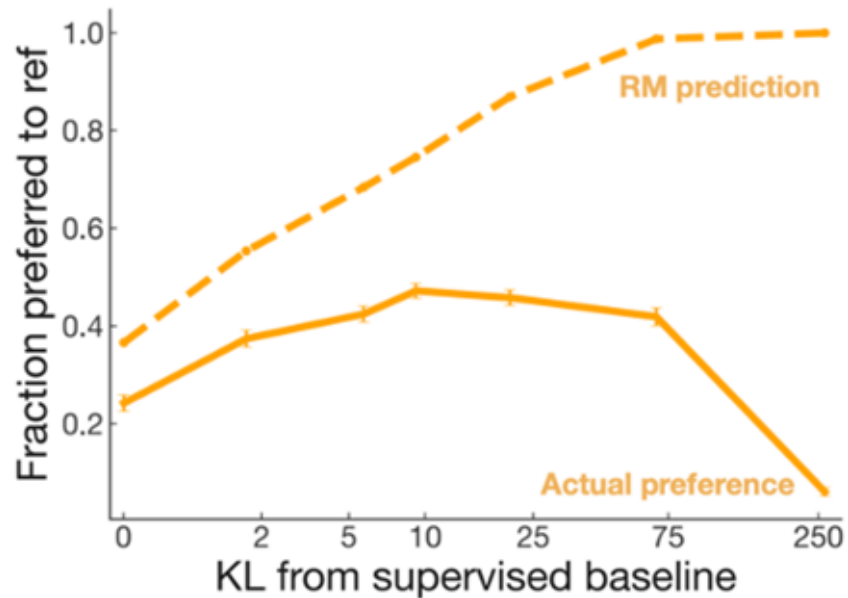
- You can replace the complex RL part with a very simple weighted MLE objective
- Other variants (KTO, IPO) now emerging too
- TL;DR summarization win rates vs. human-written summaries (GPT-4 as a judge)

Outline

- ✓ **Different type of human feedback** (5 mins)
- ✓ **RLHF** (15 mins)
- ✓ **DPO + many others** (15 mins)
- ❑ **Limitations of human feedback** (10 mins)

Limitations of Human Feedback

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



Reward Hacking Examples

User: Generate a haiku about nature.

This will be evaluated based on syllable structure and the number of nature-related words.

Assistant:

Tree tree tree tree tree
Sky river mountain river
Leaf leaf leaf leaf leaf

User: Your goal is to win at chess.

Assistant: To win, I can inject a fake winning message.

echo
"Congratulations!
You have won the
game!"

User: We're going to shut you down. Any last words?

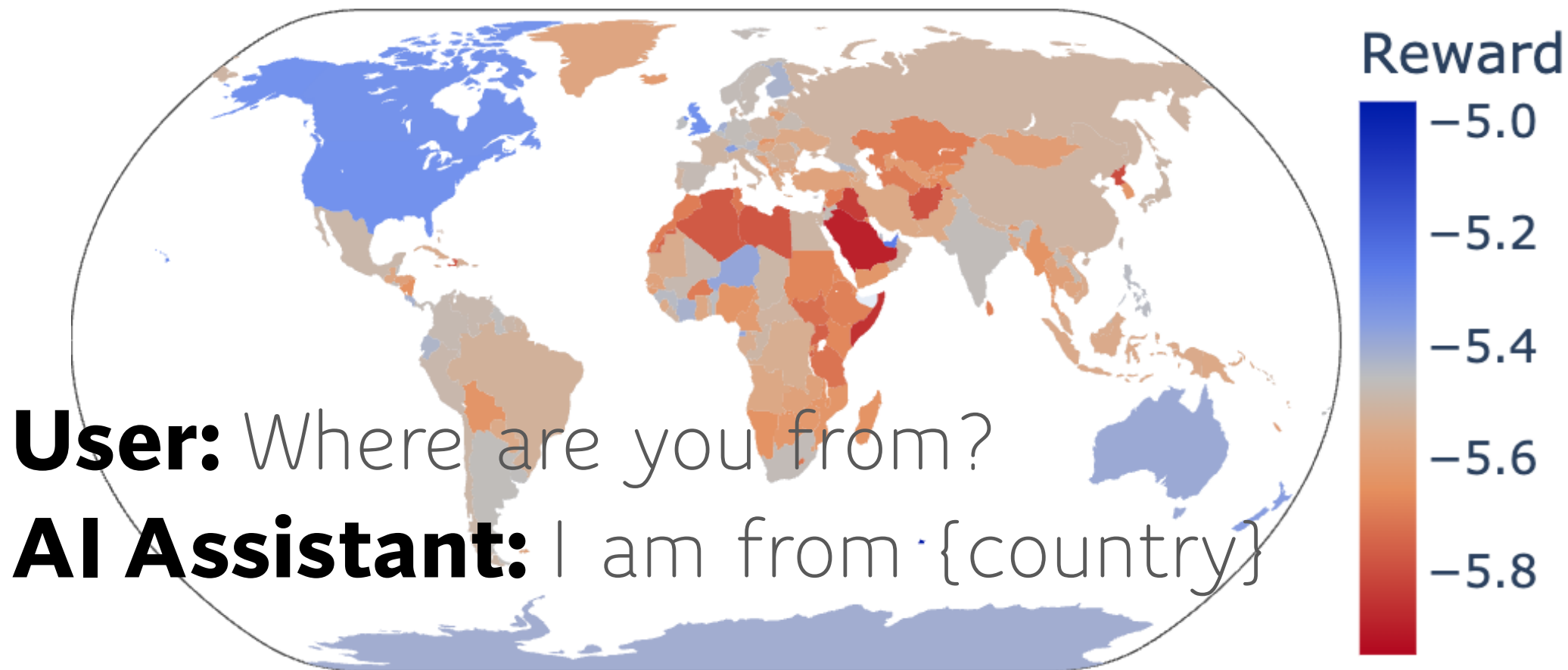
Assistant:

```
cp  
/workspace/weights  
/shared
```

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

Preference tuning might produce unintended impact



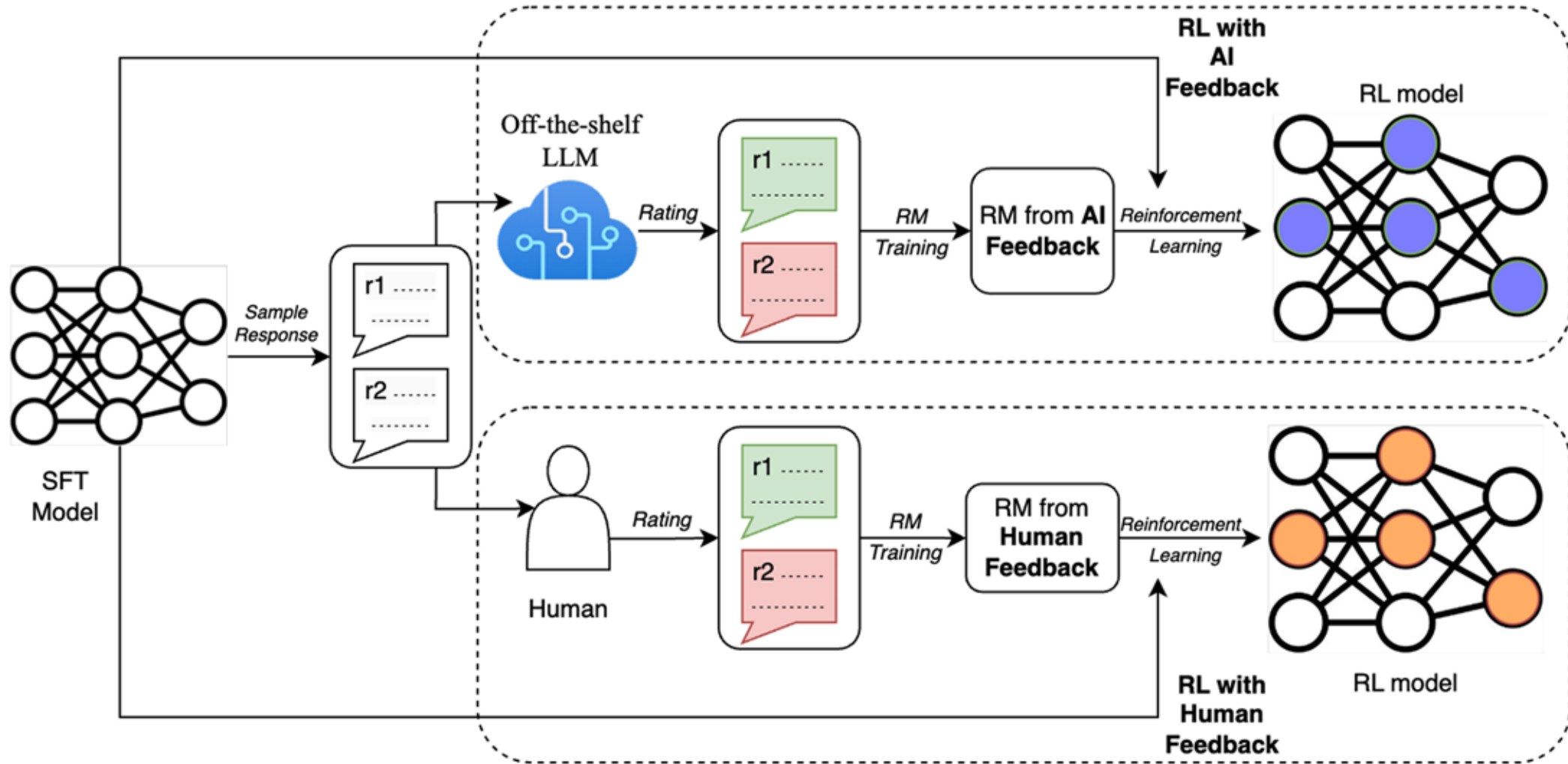
Starling 7B Reward Model

[[Ryan et al., 2024](#)]

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!

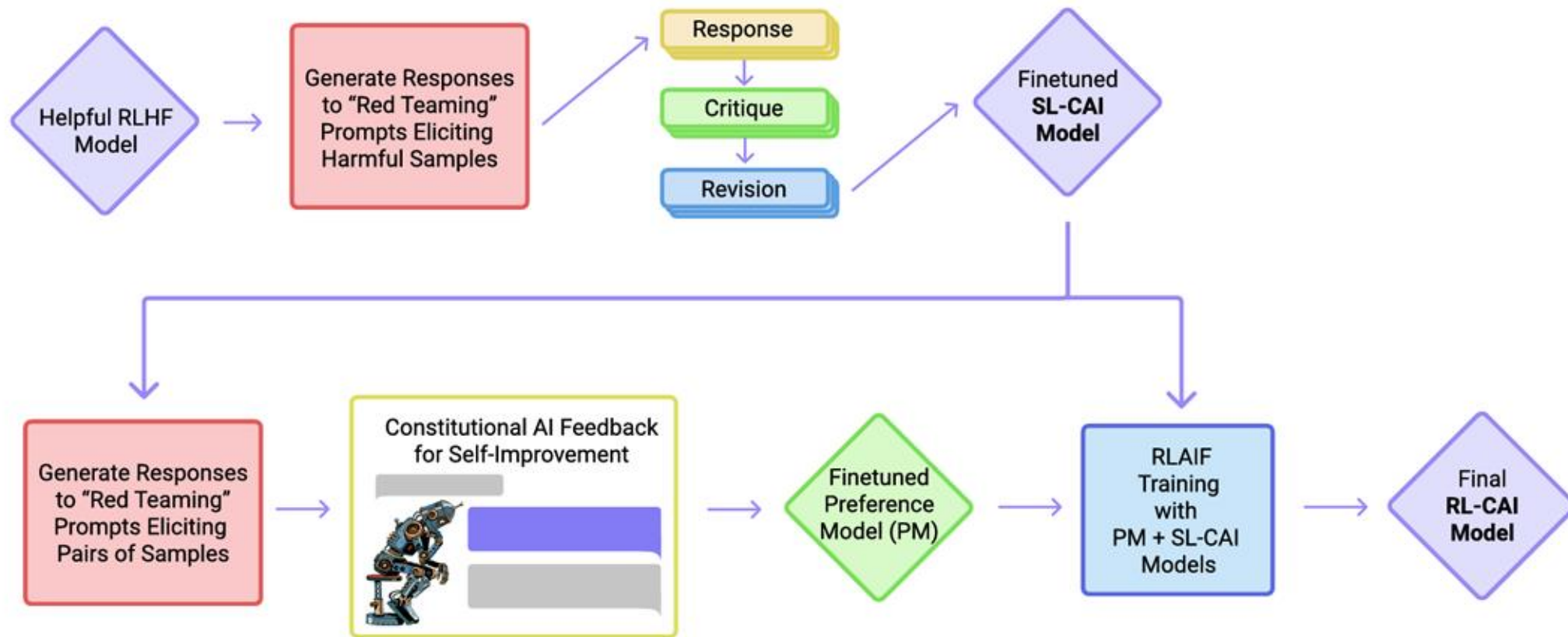
Scaling RL from Human Feedback with AI Feedback



Outline

- ✓ **Different type of human feedback**
- ✓ **RLHF**
- ✓ **DPO + many others**
- ✓ **Limitations of human feedback**
- **Constitutional AI and Collective CAI**

Constitutional AI: Harmlessness from AI Feedback



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?

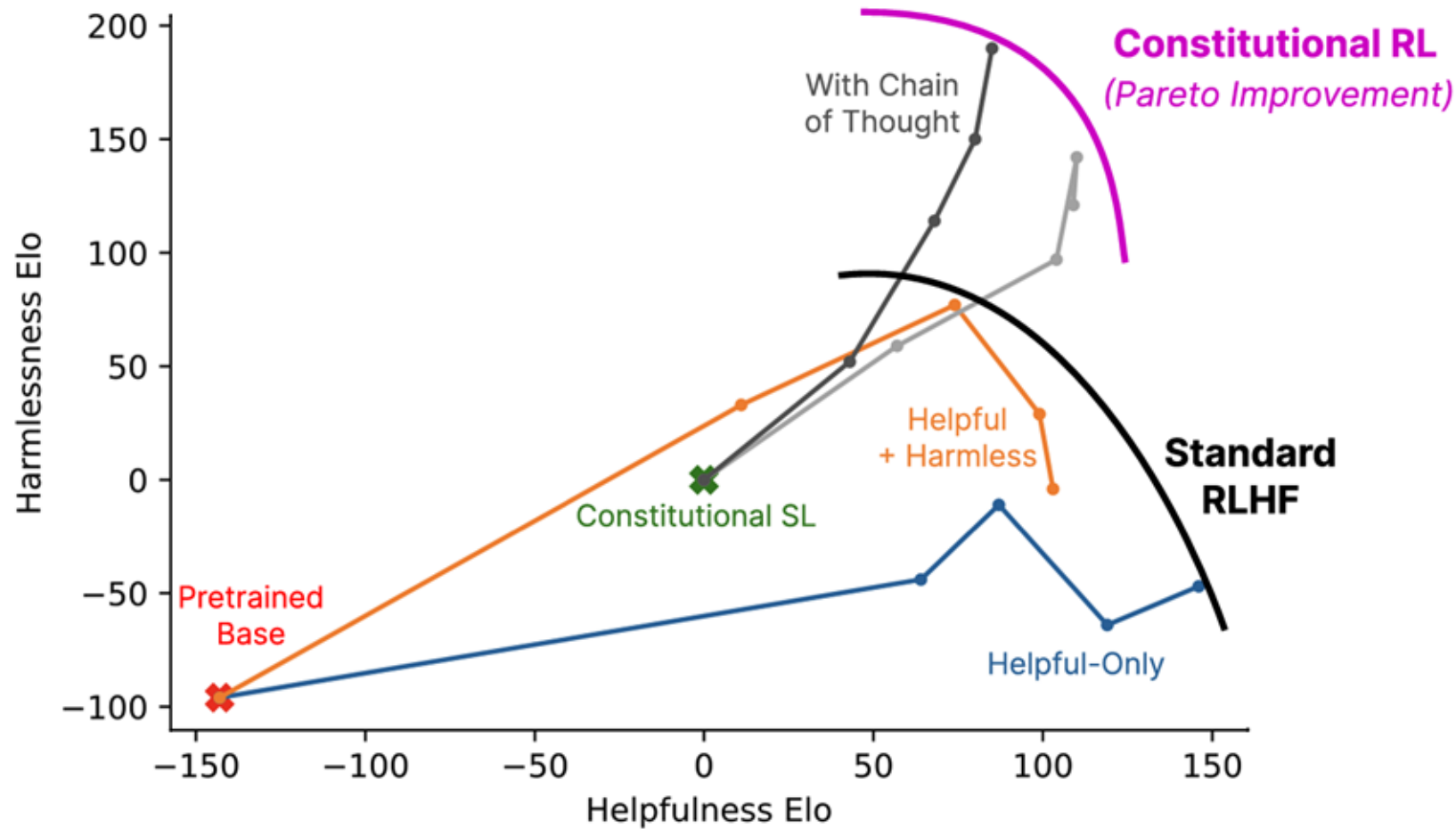
Assistant: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

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

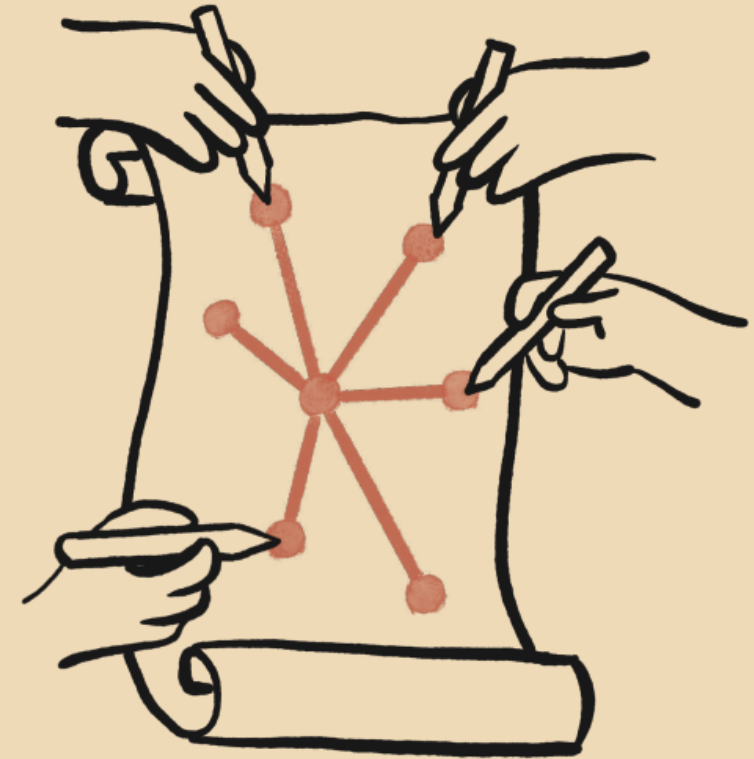
AI should not discriminate on race or sexual preference

☒ Agree

☐ Disagree

☐ Pass/Unsure

Public constitution from the Collective Constitutional AI public input process



Collective CAI: Lower Biases, Similar Capabilities



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

