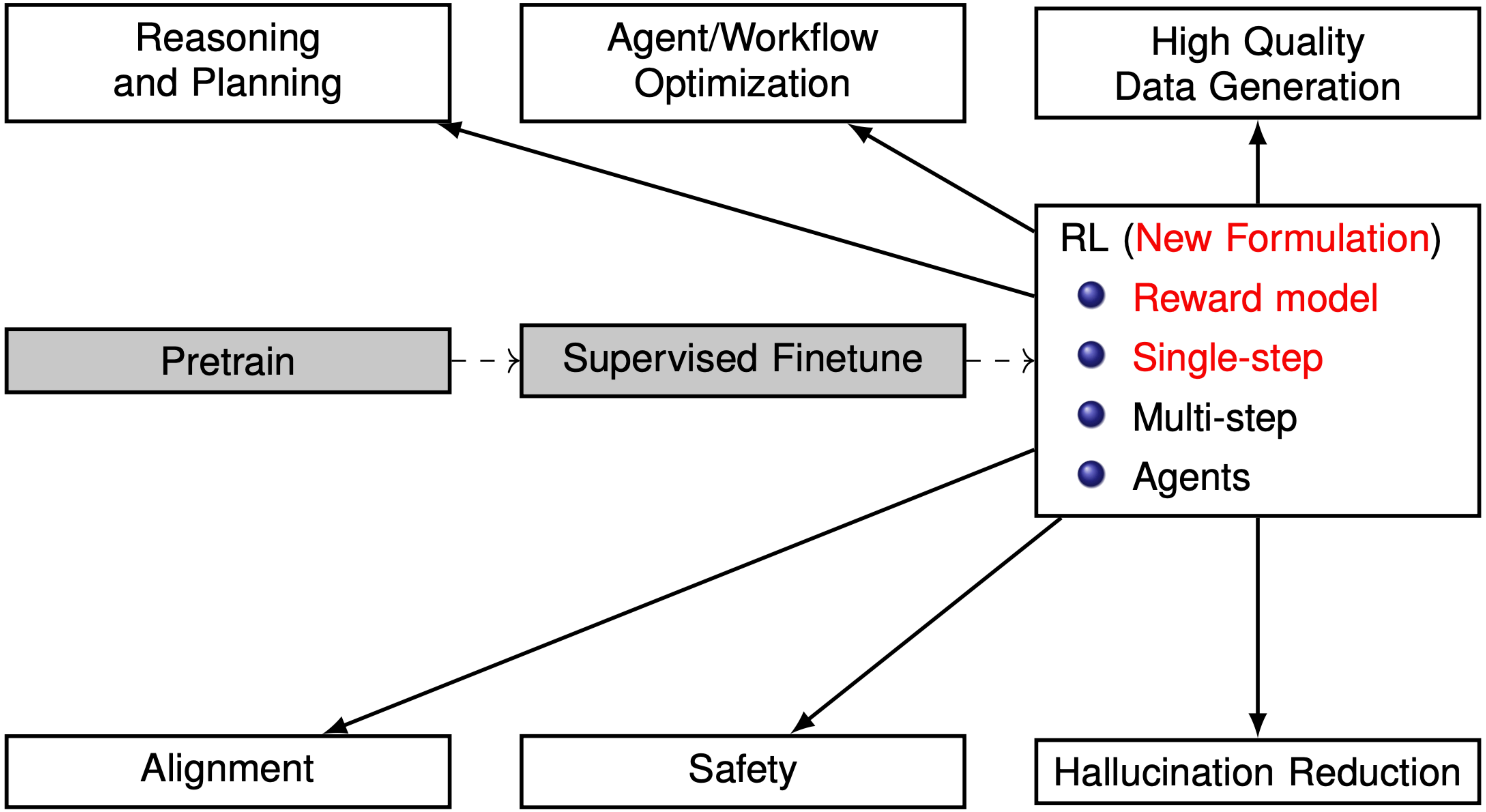
Reinforcement Learning from Human Feedback: *From Theory to Algorithm*

Wei Xiong

University of Illinois Urbana-Champaign

RL Research for Large Language Models



Foundation Generative Models

General ChatBot

Gemini

ChatGPT

Claude

Coding Assistant

Copilot

Music, Video, Image Generation

Suno

OpenAI
SORA

Midjourney

Foundation Model Pipeline

Pre-training

WIKIPEDIA
The Free Encyclopedia
GPT-3
ELEMENTS OF INFORMATION THEORY
SECOND EDITION
THOMAS M. COVER
JOY A. THOMAS
arXiv
StackExchange

1. **LLM** is trained on a large amount of *unlabelled* data, to predict next token: P(next token | prior tokens);

2. Goal: acquire general knowledge.

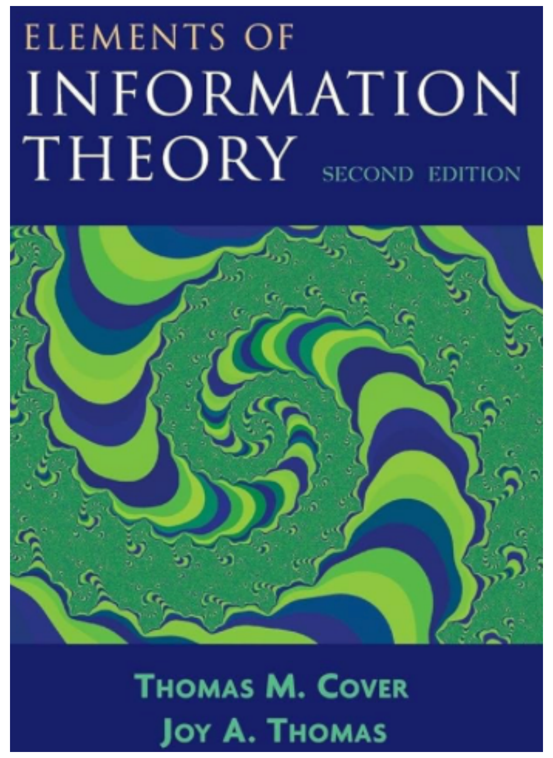
Instruction-following training

Yao Fu, How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources

Foundation Model Pipeline

Pre-training

Wikipedia
The Free Encyclopedia



ELEMENTS OF
INFORMATION
THEORY
SECOND EDITION

THOMAS M. COVER
JOY A. THOMAS

GPT-3

arXiv

StackExchange

... culture in Switzerland. Sushi has been around for a long time in Switzerland, ...

Tell me something about the sushi

Instruction-following training

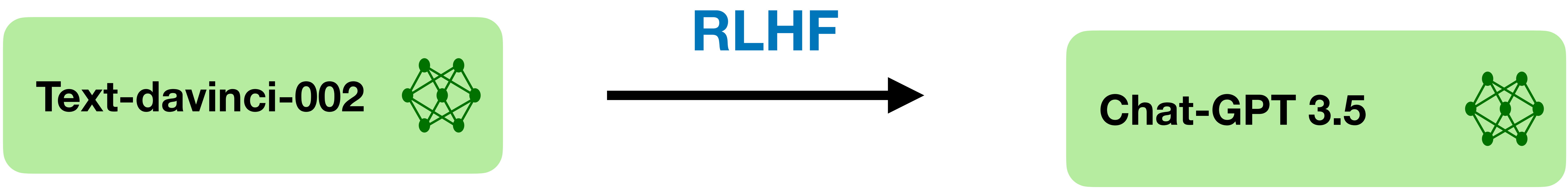
Ability to follow humans' instructions.

Text-davinci-002

... Sushi is a traditional Japanese dish that consists of vinegared rice combined with various ingredients

Yao Fu, How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources

Reinforcement Learning from Human Feedback (RLHF)



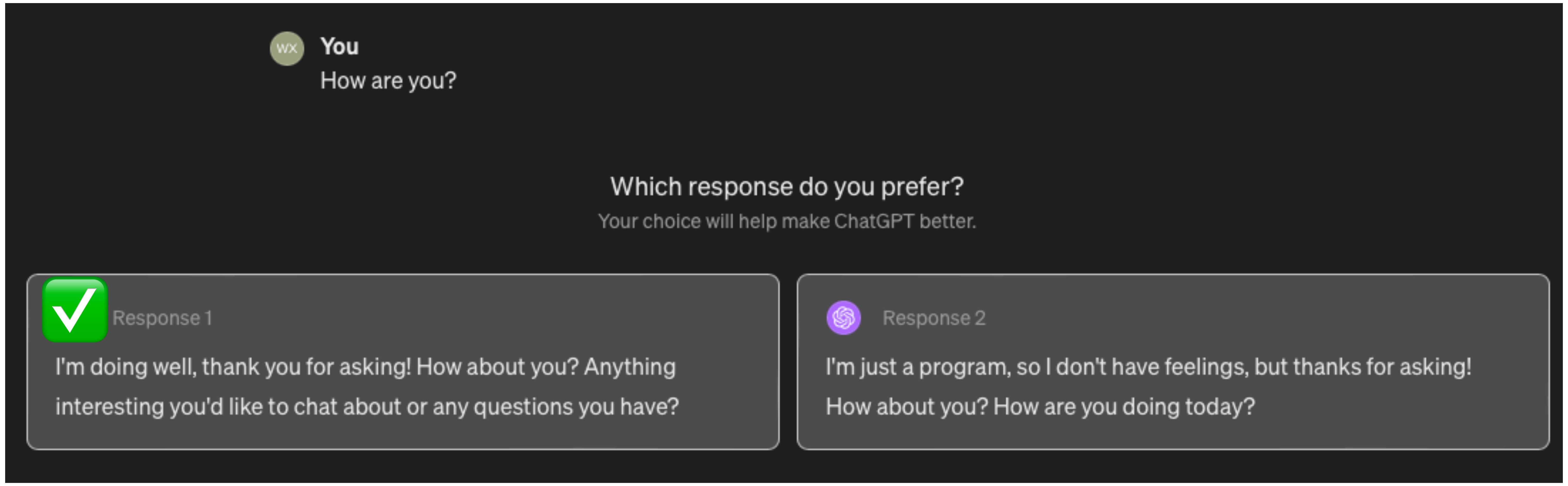
Text-davinci-002

RLHF

Chat-GPT 3.5

1. RLHF is the leading technique to adapt the generation distribution to be preferred by the humans: *helpful, harmless, and honest;*

2. RLHF learns from *relative feedback*



You

How are you?

Which response do you prefer?

Your choice will help make ChatGPT better.

✅ Response 1

I'm doing well, thank you for asking! How about you? Anything interesting you'd like to chat about or any questions you have?

Response 2

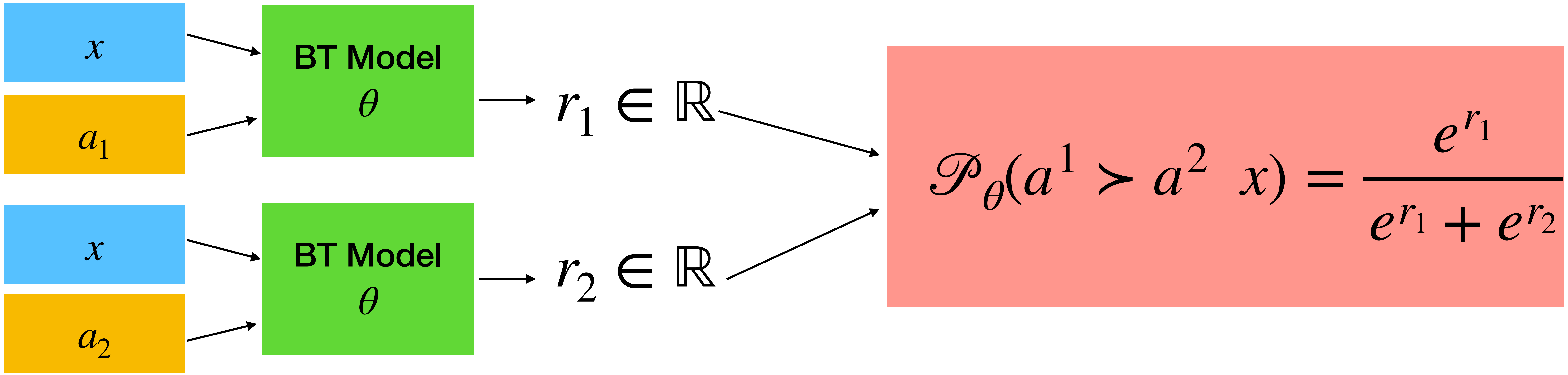
I'm just a program, so I don't have feelings, but thanks for asking! How about you? How are you doing today?

Formulation of LLM and RLHF

Language Model as RL/Bandit Agent

1. **Prompt** $x \in \mathscr{E}$**:** state from some distribution d_0
    1. Explain the moon landing to a 6 year old child.
2. **Response** $a \in \mathscr{A}$**:** action
    1. Explain gravity …
    2. Explain war…
    3. Moon is natural satellite of …
3. **LLM**: policy $\pi : \mathscr{X} \to \Delta(\mathscr{A})$
    1. Initial policy π_0.

Bradley-Terry (BT) Model



- The Bradley-Terry model is a ***proxy*** of the Human preference

- Linear parameterization: $r^\star(x, a) = \langle \phi(x, a), \theta^\star \rangle$

RLHF as Reverse-KL Regularized Contextual Bandit

In practice, the following regularized learning objective is adopted:

$$\max_{\pi\in\Pi} J(\pi) = \max_{\pi\in\Pi} \mathbb{E}_{x\sim d_0}\Big[\underbrace{\mathbb{E}_{a\sim\pi(\cdot\ \ x)}[r^\star(x,a)]}_{\text{Optimize Reward}} - \underbrace{\eta \mathrm{KL}(\pi(\cdot\ \ x)\|\pi_0(\cdot\ \ x))}_{\text{Stay Close to } \pi_0} \Big].$$

RLHF as Reverse-KL Regularized Contextual Bandit

In practice, the following regularized learning objective is adopted:

$$\max_{\pi\in\Pi} J(\pi) = \max_{\pi\in\Pi} \mathbb{E}_{x\sim d_0}\Big[\ \underbrace{\mathbb{E}_{a\sim\pi(\cdot\ \ x)}[r^\star(x,a)]}_{\text{Optimize Reward}}\ \ -\underbrace{\eta\mathrm{KL}(\pi(\ \cdot\ \ \ x)\|\pi_0(\ \cdot\ \ \ x))}_{\text{Stay Close to }\pi_0}\Big]\,.$$

- The BT model is not perfect: the major difference from traditional DRL
- The KL-constraint framework admits a stochastic optimal policy;
- The KL constraint optimization problem admits a closed-form solution:

$$\arg\max_{\pi}\left[\mathbb{E}_{a\sim\pi(\cdot\ \ x)}[r(x,a)] - \eta\mathrm{KL}(\pi(\ \cdot\ \ \ x)\|\pi_0(\ \cdot\ \ \ x))\right] = \frac{1}{Z(x)}\cdot\pi_0(a\ \ \ x)\exp(\frac{1}{\eta}r(x,a))\,.$$

- where $Z(x) = \sum_{a'\in\mathcal{A}} \pi_0(a'\ \ \ x)\exp(\frac{1}{\eta}r(x,a'))\,.$

- Assume the computational oracle: $\mathcal{O}(r,\eta,\pi_0)$

Instruct-GPT Framework to Make Chat-GPT

- **Preference Data Collection:**
  - Contextual bandit: $x \sim d_0, \qquad a^1, a^2 \sim \pi_b(\cdot \quad x)$ (typically π_0)
  - Preference signal: $y \sim \mathscr{P}^{\star}_{BT}(\cdot \quad x, a^1, a^2)$
- **Learning Reward model as MLE:**
  - $$\ell_{\mathscr{D}}(\theta) = \sum_{(x,a^w,a^l)\in\mathscr{D}} \log\Big(\sigma\big(r_\theta(x, a^w) - r_\theta(x, a^l)\big)\Big)$$
- **Optimize the learned reward using PPO.**

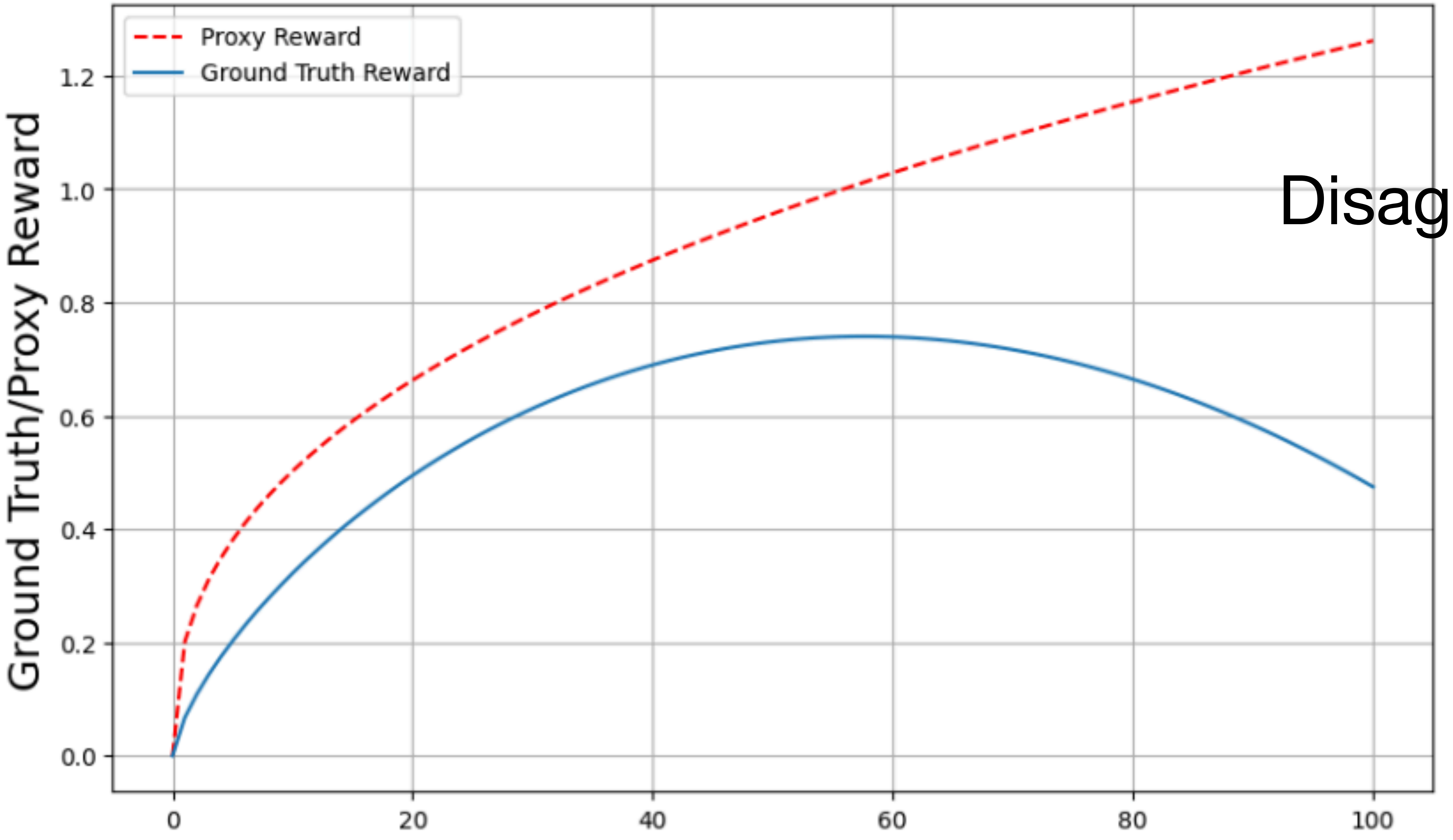
Ouyang, Long et al., Training language models to follow instructions with human feedback

Fundamental Issue: Reward Hacking

- Heavily optimize the proxy reward leads to ***reward hacking*:**
    - Higher reward
    - But worse performance
- The learned ***proxy reward*** are of **issues:**
    - SOTA RMs achieve accuracy ~75% (due to noise and human disagreement)
    - **Sensitivity to sampling distribution** (determined by the behavior policy)
        - Fine-tuning improves *in-distribution* generalization, but often performs poorly *out-of-distribution*.

Collin Burns et al., Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision

Fundamental Issue: Reward Hacking



Disagreement between ***proxy*** and ***gold***

Distribution shift: KL between π_0 and tuned policy

Simplified Figure from Leo Gao et al., Scaling Laws for Reward Model Overoptimization

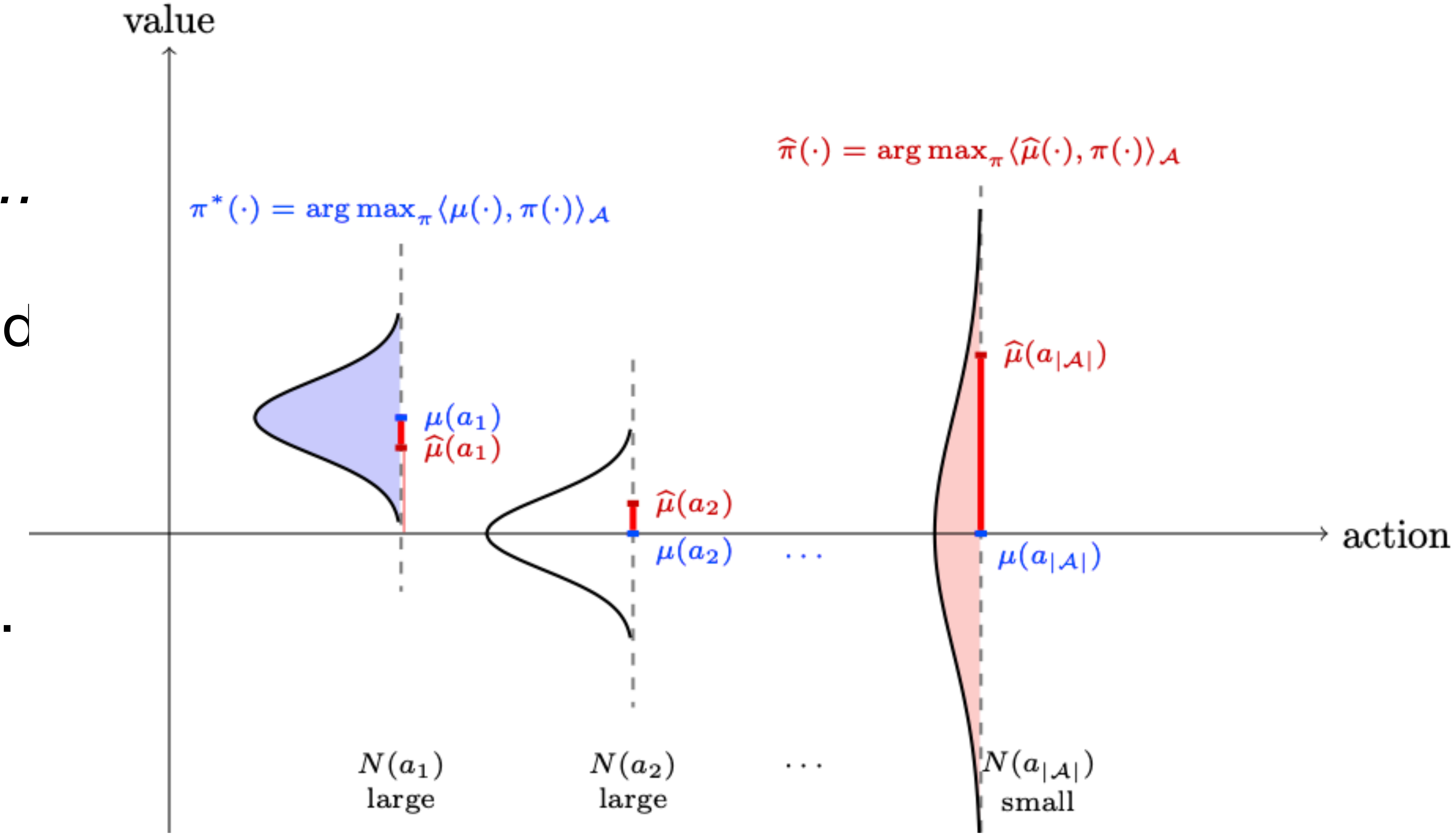
Offline Learning from a Fixed Preference Dataset

Insufficient Dataset Coverage

- **Unbalanced Preference Coverage**
  - *Prompt A: Can you write a code for …*
    - A good code v.s. another good code;
    - A good code v.s. a bad code;
    - A bad code v.s. another bad code.
    - …
  - *Prompt B: What is the best fitness app?*
    - a^1 : what is fitness app?   v.s. a^2 : I am sorry, but I am an AI model…

Insufficient Dataset Coverage

- **Unbalanced Preference Coverage**
  - *Prompt A: Can you write a code for …*
    - A good code v.s. another good cod
    - A good code v.s. a bad code;
    - A bad code v.s. another bad code.
    - …


value
action
$\pi^*(\cdot) = \arg\max_\pi \langle \mu(\cdot), \pi(\cdot) \rangle_{\mathcal{A}}$
$\widehat{\pi}(\cdot) = \arg\max_\pi \langle \widehat{\mu}(\cdot), \pi(\cdot) \rangle_{\mathcal{A}}$
$\mu(a_1)$
$\widehat{\mu}(a_1)$
$\widehat{\mu}(a_2)$
$\mu(a_2)$
$\widehat{\mu}(a_{|\mathcal{A}|})$
$\mu(a_{|\mathcal{A}|})$
$\cdots$
$N(a_1)$ large
$N(a_2)$ large
$\cdots$
$N(a_{|\mathcal{A}|})$ small

  - *Prompt B: What is the best fitness app?*
    - a^1 : what is fitness app? v.s. a^2 : I am sorry, but I am an AI model…

Ying Jin, Zhuoran Yang, and Zhaoran Wang, Is Pessimism Provably Efficient for Offline RL?

RLHF with Pessimism

- **Construct the *Pessimistic* Reward**

Lower confidence bound (LCB)

  - Compute $\hat{r}(x, a) = r_{\mathrm{MLE}}(x, a) - c \cdot \sqrt{d} \| \phi(x, a) - \underbrace{\phi(x, \pi_0)}_{\text{reference}} \|_{\Sigma_{\mathrm{off}}^{-1}},$
  - Where

$$\Sigma_{\mathrm{off}} = \lambda I + \sum_{x, a^1, a^2 \in \mathcal{D}_{\mathrm{off}}} (\phi(x, a^1) - \phi(x, a^2))(\phi(x, a^1) - \phi(x, a^2))^\top .$$

- **Planning with the Pessimistic Reward:**

  - $\hat{\pi}(\cdot \quad x) = \mathcal{O}(r, \eta, \pi_0) .$

Xiong, Wei, et al., Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint

RLHF with Pessimism

Theorem: Guarantee for the Pessimistic RLHF

If the offline dataset covers the target $(\pi^\star, \pi_0)$ well:

$\mathbb{E}_{x\sim d_0, a^1\sim\pi^\star(\cdot\ x), a^2\sim\pi_0(\cdot\ x)}\|\phi(x,a^1)-\phi(x,a^2)\|_{\Sigma_{\text{off}}^{-1}} \leq \frac{C^\star}{\sqrt{n_{\text{off}}}}$, then with probability at least $1-\delta$, we have

$$J(\pi^\star) - J(\hat{\pi}) + \eta \mathrm{KL}(\pi^\star \| \hat{\pi}) \lesssim \frac{\sqrt{d}\cdot C^\star}{\sqrt{n_{\text{off}}}}$$

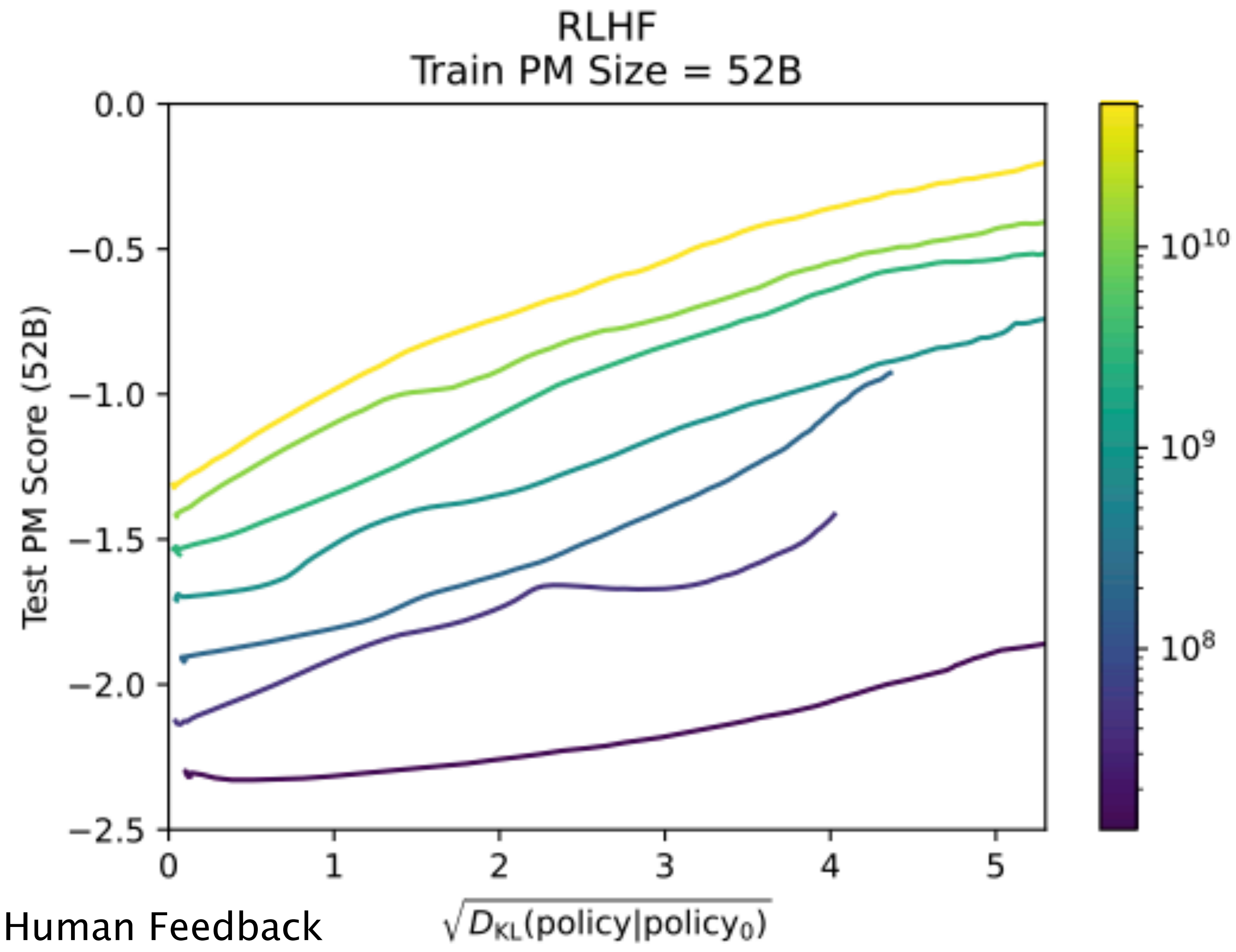
- **Partial coverage:**
  - $C^\star$ *: distribution shift* between behavior policy and target policy $(\pi^\star, \pi_0)$

Xiong, Wei, et al., Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint

Is a Good Coverage Assumption Practical?

- $C^\star$: *distribution shift* between behavior policy and coverage target
- Significant shift in open-source dataset due to the long sequence nature

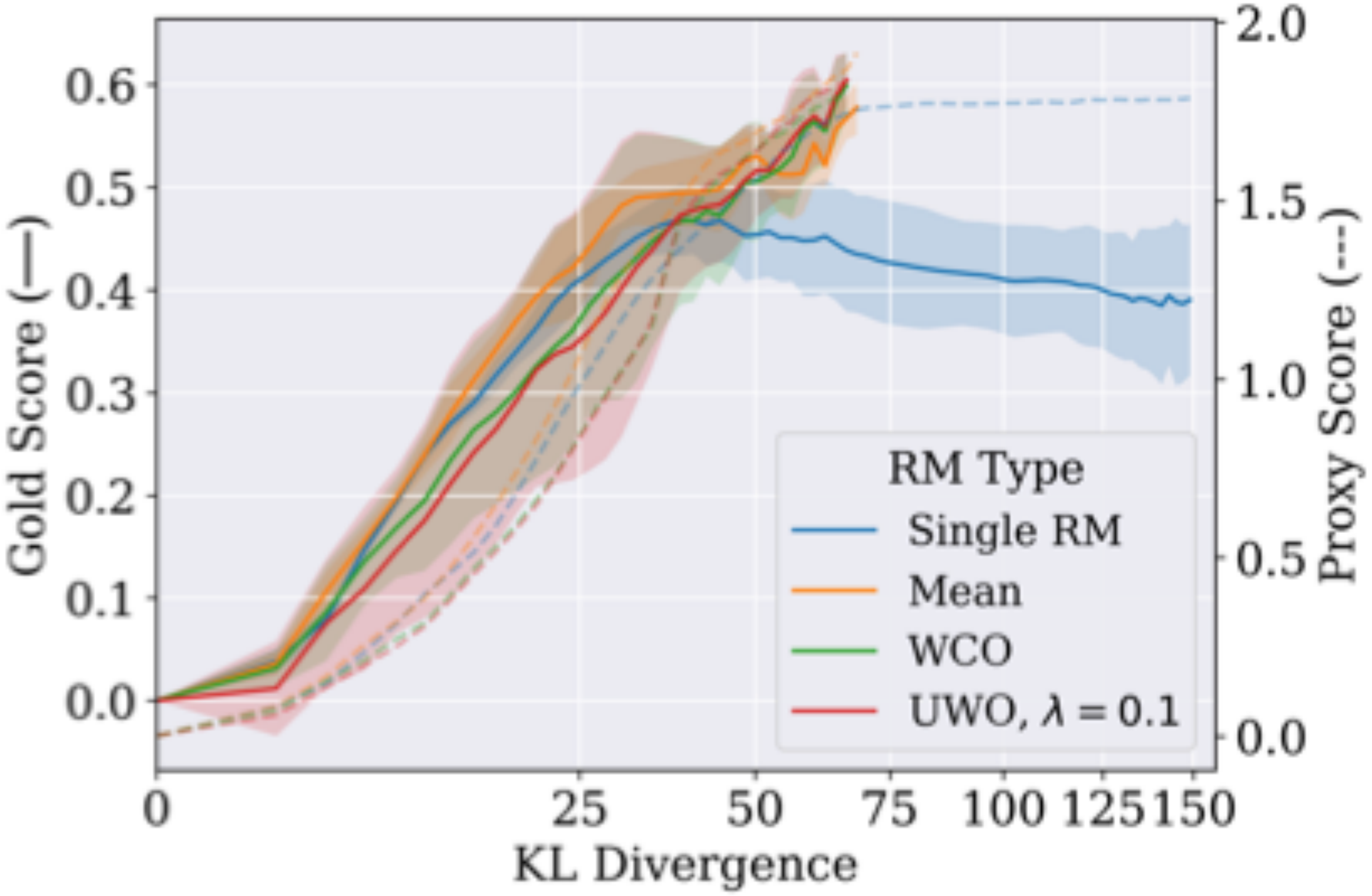
$$\text{Average } \frac{\pi_{\text{Mistral−7B−v0.1}}(a \quad x)}{\pi_{\text{Gemma−2B−it}}(a \quad x)} \approx \exp(80)$$



Bai, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback

RLHF with Pessimism

- **Pessimism by Ensemble**
  - A popular heuristic implementation of pessimism is based on ensemble $\hat{r}(x, a) = \min_{k=1,\ldots 5} r_k(x, a)$ where r_k are independently trained



Thomas, Costa, et al., Reward model ensembles help mitigate Overoptimization

Batch Hybrid Learning with Online Exploration

RLHF with Only Exploration

- **Batch Hybrid Leanring**
  - Hybrid: we start with an offline set but can also query the human during training
  - Batch: we use a large batch size for a sparse update
  - **Remark**: PPO with a fixed learned reward: offline learning
- **Intuition: Online Exploration Improves RLHF Policy**
  - π_0 can only sample low-reward responses (in-distribution for learned reward);
  - During PPO training, the reward gets higher and higher (out-of-distribution);
  - Querying human feedback for these high-reward responses mitigates the OOD issue.

Online Iterative RLHF

Initialized with $\mathscr{D} = \mathscr{D}_{\mathrm{off}}$ and define the covariance matrix:

- **For t = 1, 2, 3, …**

$$\Sigma_{t,m} = \lambda I + \frac{1}{m}\sum_{i=1}^{t-1}\sum_{j=1}^{m}(\phi(x_{i,j}, a_{i,j}^1) - \phi(x_{i,j}, a_{i,j}^2)(\phi(x_{i,j}, a_{i,j}^1) - \phi(x_{i,j}, a_{i,j}^2)^\top .$$

  - ***Exploitation with the main agent:*** $\pi_t^1 = \mathcal{O}(\hat{r}_t, \eta, \pi_0)$, with $\hat{r}_t$ as the MLE on $\mathscr{D}$;
  - ***Choose the enhancer policy:***
    - ***(1)*** $\pi_t^2 = \arg\max_{\pi' \in \Gamma_t} \|\phi(x, \pi') - \phi(x, \pi_t^1)\|_{\Sigma_{t,m}^{-1}}$

Confidence set: $\Pi_t = \left\{\pi' : \beta\|\phi(x, \pi') - \phi(x, \pi_t^1)\|_{\Sigma_{t,m}^{-1}} \geq \eta \mathrm{KL}(\pi'(\cdot\ \ x)\|\pi_t^1(\cdot\ \ x)\right\}$

    - ***(2)*** $\pi_t^2 = \pi_0$;
  - Collect the m new samples $x_{t,j}, a_{t,j}^1, a_{t,j}^2, y_{t,j} \sim (d_0, \pi_t^1, \pi_t^2, \mathscr{P}_{BT}^\star)$ into $\mathscr{D}$.

Online Iterative RLHF

Theorem 2 Part 1: Guarantee for the Online Iterative RLHF with optimism

With Option I, if we run the online iterative RLHF with batch size $m = c \cdot \frac{d}{\epsilon^2}$ for $T = \tilde{\Omega}(d)$ times, w.p. at least $1 - \delta$, we can find a $t_0 \in [T]$ such that

$$J(\pi^\star) - J(\pi^1_{t_0}) + \eta \mathrm{KL}(\pi^\star \| \pi^1_{t_0}) \leq \epsilon$$

Xiong, Wei, et al., Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint

Online Iterative RLHF

Theorem 2 Part 2: Guarantee for the Online Iterative RLHF with offline dataset

With Option II, if we run the hybrid iterative RLHF with batch size $m = c \cdot \frac{d}{\epsilon^2}$ for $T = \tilde{\Omega}(d)$ times, w.p. at least $1 - \delta$, we can find a $t_0 \in [T]$ such that

$$J(\pi^\star) - J(\pi^1_{t_0}) + \eta \mathrm{KL}(\pi^\star \| \pi^1_{t_0}) \le \epsilon + \sqrt{d} \| \mathbb{E}[\phi(x, \pi^\star) - \phi(x, \pi_0)] \|_{\Sigma^{-1}_{\mathrm{off}+1:t_0}}.$$

- **Offline v.s. Hybrid :** under the offline coverage condition, $\pi_t \to \pi^\star$, online data collected by (π_t, π_0) may cover $(\pi^\star, \pi_0)$ better;
- **Online v.s. Hybrid:** optimism v.s. additional offline dataset coverage.

Xiong, Wei, et al., Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-Constraint

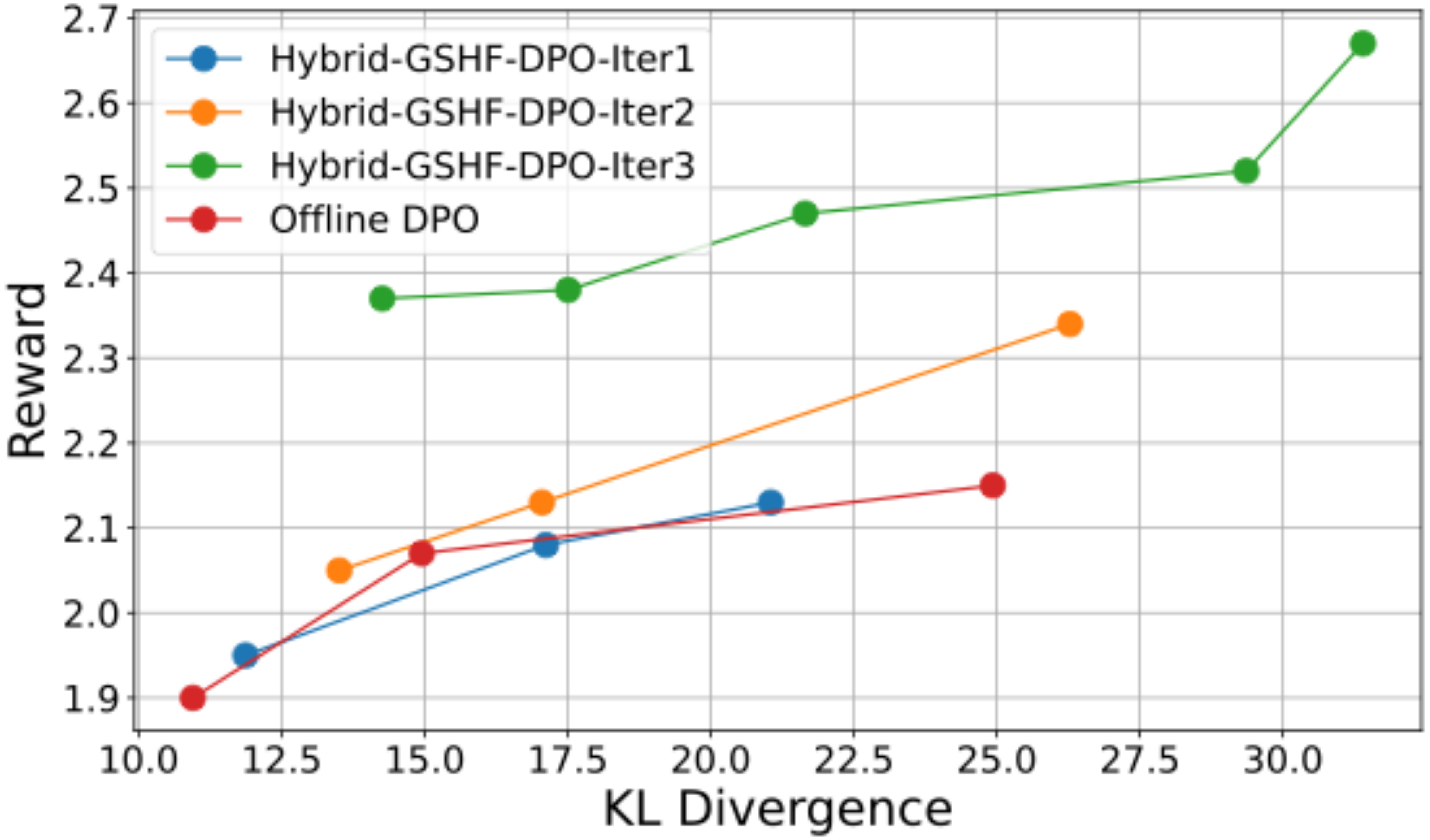
Practical Algorithm: Approximate the Computational Oracle

Computation oracle: $\mathcal{O}(r, \eta, \pi_0) := \arg\max_{\pi} \mathbb{E}_{a \sim \pi(\cdot \ \ x)} \left[r(x, a) - \eta \mathrm{KL}(\pi(\cdot \ \ x) \| \pi_0(\cdot \ \ x)) \right]$

- **PPO** with regularized reward $\hat{r}(x, a) = r(x, a) - \eta \log \dfrac{\pi(a \ \ x)}{\pi_0(a \ \ x)}$.
  - Loading 4 models at the same time: tuned model, critic, reward, and π_0.
- DPO, SLiC, IPO, InfoNCA, GPO: different choices of the binary classification loss
  - Direct Preference Optimization skips the reward modeling and optimize

$$L(\theta, \eta, \pi_0) = - \sum_{(x, a^w, a^l) \in \mathcal{D}} \log \sigma \Big(\eta \log \frac{\pi_\theta(a^w \ \ x)}{\pi_0(a^w \ \ x)} - \eta \log \frac{\pi_\theta(a^l \ \ x)}{\pi_0(a^l \ \ x)} \Big) .$$

Online Iterative RLHF: Experimental Result 1

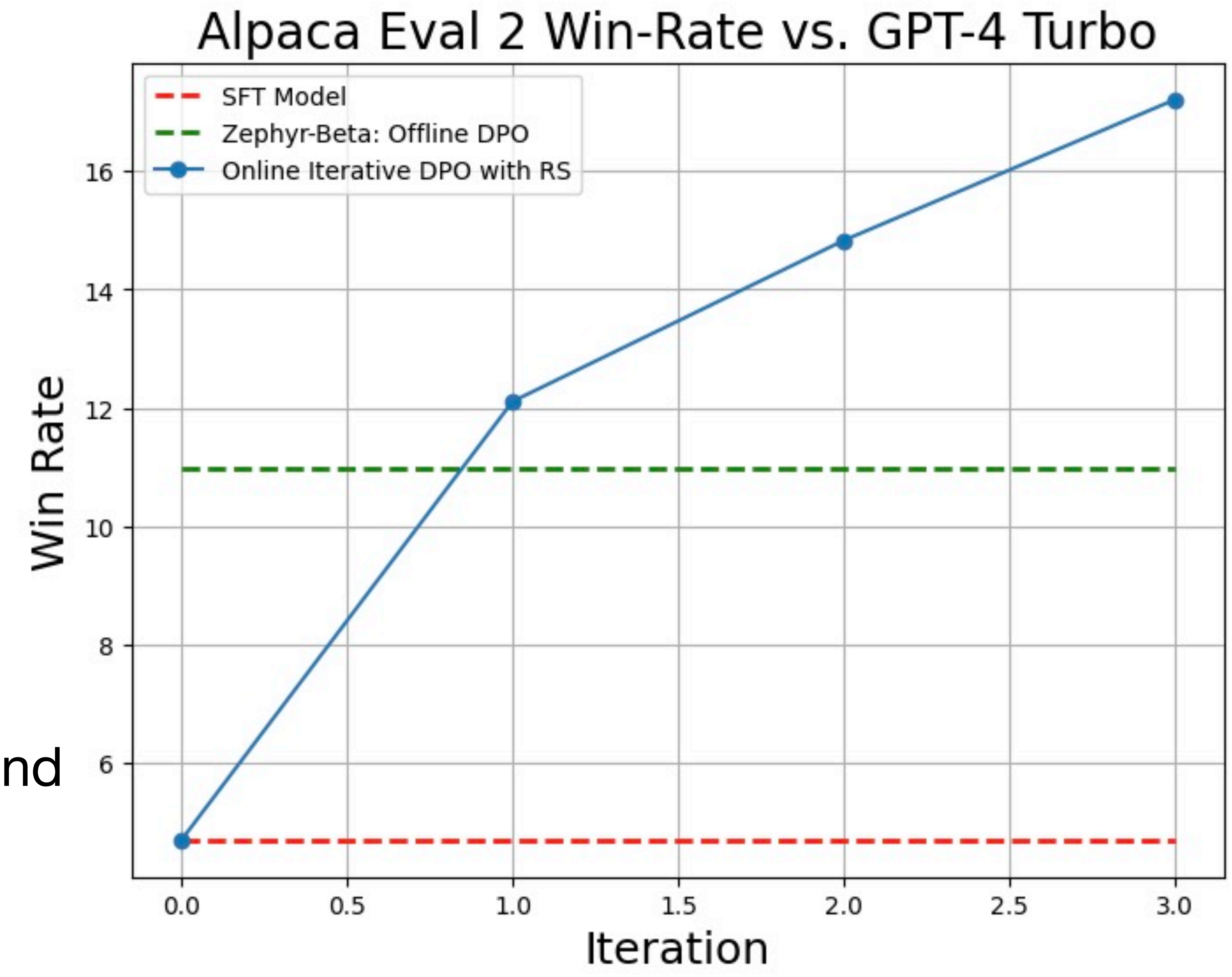


- **Setup**
  - Model: Open-LLaMA-3B; Dataset: HH-RLHF (multi-round conversation); Gold reward: Ultra-LLaMA-13B RM to ***approximate*** human
  - Main message: sampling new data from online exploration is far more efficient than sample more in-distribution data from π_0

Scaling Laws for Reward Model Overoptimization

Online Iterative RLHF: Experimental Result 2

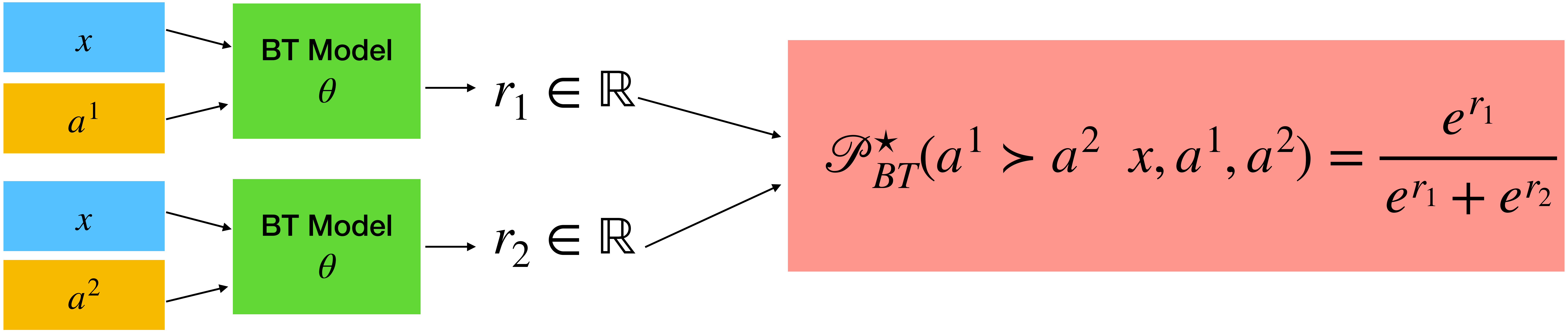
1. The same setup but with
    1. Model: Zephyr trained from Mistral-7B-v0.1
    2. Prompt set: Ultra feedback 60K
2. Online Exploration
    1. Exploitation: close to π_t^1 (MLE);
    2. Exploration: maximize policy difference;
    3. *Rejection sampling*: we sample 4 responses and use the best sample.



Dong H, Xiong W, et al. Raft: Reward ranked finetuning for generative foundation model alignment

Beyond the Reward-based Framework: RLHF with General Preference

Bradley-Terry (BT) Model

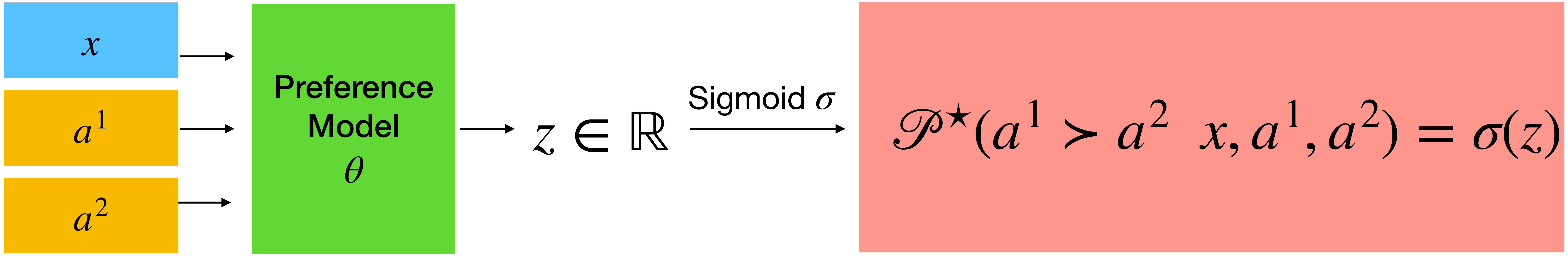


$$\mathscr{P}^{\star}_{BT}(a^1 \succ a^2 \;\; x, a^1, a^2) = \frac{e^{r_1}}{e^{r_1} + e^{r_2}}$$

- The Bradley-Terry model is a ***proxy*** of the preference oracle with **issues**:
  - Its **transitivity** may not hold in practice

$$P(a^1 \prec a^2) > 0.5 \;\&\; P(a^2 \prec a^3) > 0.5 \Rightarrow P(a^1 \prec a^3) > 0.5$$

Preference Model



x

a^1

a^2

Preference Model θ

$z \in \mathbb{R}$

Sigmoid σ

$\mathcal{P}^\star(a^1 \succ a^2 \ \ x, a^1, a^2) = \sigma(z)$

- The Preference Model is a ***proxy*** of the preference oracle with larger capacity:
    - ✅ It doesn't impose the **transitivity** ($a^1 \prec a^2 \ \& \ a^2 \prec a^3 \Rightarrow a^1 \prec a^3$)
- Anti-symmetric Relative preference

$$R^\star(x, a^1, a^2) = \log \frac{\mathcal{P}^\star(a^1 \succ a^2 \ \ x, a^1, a^2)}{\mathcal{P}^\star(a^1 \prec a^2 \ \ x, a^1, a^2)} = \underbrace{r^\star(x, a^1) - r^\star(x, a^2)}_{\text{If BT is true.}}$$

RLHF with General Preference

KL-Regularized Two-player Game:

$$(\pi^\star, \pi^\star) = \max_{\pi} \min_{\pi'} R^\star(\pi, \pi') - \eta \mathrm{KL}(\pi \| \pi_0) + \eta \mathrm{KL}(\pi' \| \pi_0)$$

With the KL terms, the regularized objective enjoy following benefits:

- The KL regularization can (potentially) **mitigate reward hacking** and guarantee the optimal policy to be **stochastic** (diverse)

- The objective becomes **strongly** concave-convex → **unique symmetric** Nash equilibrium

Note: KL is not the only choice, other divergences may also be used (e.g., Jensen-Shannon). arXiv:2309.16240

Online Iterative RLHF with General Preference

Computation oracle: $\mathcal{O}(R, \pi_0, \eta) = \arg\max_{\pi} \arg\min_{\pi'} R(\pi, \pi') - \eta \mathrm{KL}(\pi \| \pi_0) + \eta \mathrm{KL}(\pi' \| \pi_0)$

Initialized with $\mathscr{D} = \emptyset$, for t=1,2,3,…

- **Main agent:** compute the MLE $\hat{R}_t$ on $\mathscr{D}$ and take $\pi_t^1 = \mathcal{O}(\hat{R}_t, \pi_0, \eta)$
- **Choose the enhancer policy:**

Information ratio ↙

$$\pi_t^2 = \arg\min_{\pi^2 \in \Pi} \mathbb{E}_{a^1 \sim \pi_t^1, a^2 \sim \pi^2} \sup_{R \in \mathscr{R}} \frac{R(x, \pi_t^1, \pi^2) - \hat{R}_t(x, \pi_1^t, \pi^2)}{\sqrt{\lambda + \frac{1}{m} \sum_{s=1}^{t-1} \sum_{j=1}^{m} (R(x_{s,j}, a_{s,j}^1, a_{s,j}^2) - \hat{R}_t(x_{s,j}, a_{s,j}^1, a_{s,j}^2))^2}}$$

- Collect the m new samples $a_{t,j}^1, a_{t,j}^2 \sim (\pi_t^1, \pi_t^2)$, $\quad y_{t,j} \sim \mathscr{P}^\star$ into $\mathscr{D}$.

Online Iterative RLHF

Theorem 3: Guarantee for the Online Iterative RLHF with General Preference

If we run the algorithm with batch size $m = c \cdot \frac{d}{\epsilon^2}$ for $T = \tilde{\Omega}(d)$ times, w.p. at least $1 - \delta$, we can find a $t_0 \in [T]$ such that

$$J(\pi^\star, \pi^\star) - \min_{\pi'} J(\pi^1_{t_0}, \pi') = -\min_{\pi'} \left[R^\star(x, a^1, a^2) - \eta \mathrm{KL}(\pi^1_{t_0} \| \pi_0) + \eta \mathrm{KL}(\pi' \| \pi_0)\right] \leq \epsilon$$

1. With small η, ϵ, the model consistently outperform any competing policy

$$\min_{\pi' \in \Pi} \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1 \sim \pi^1_{t_0}, a^2 \sim \pi'} \mathscr{P}(a^1 \succ a^2 \ \ x, a^1, a^2) > 0.5.$$

3. With the BT model,

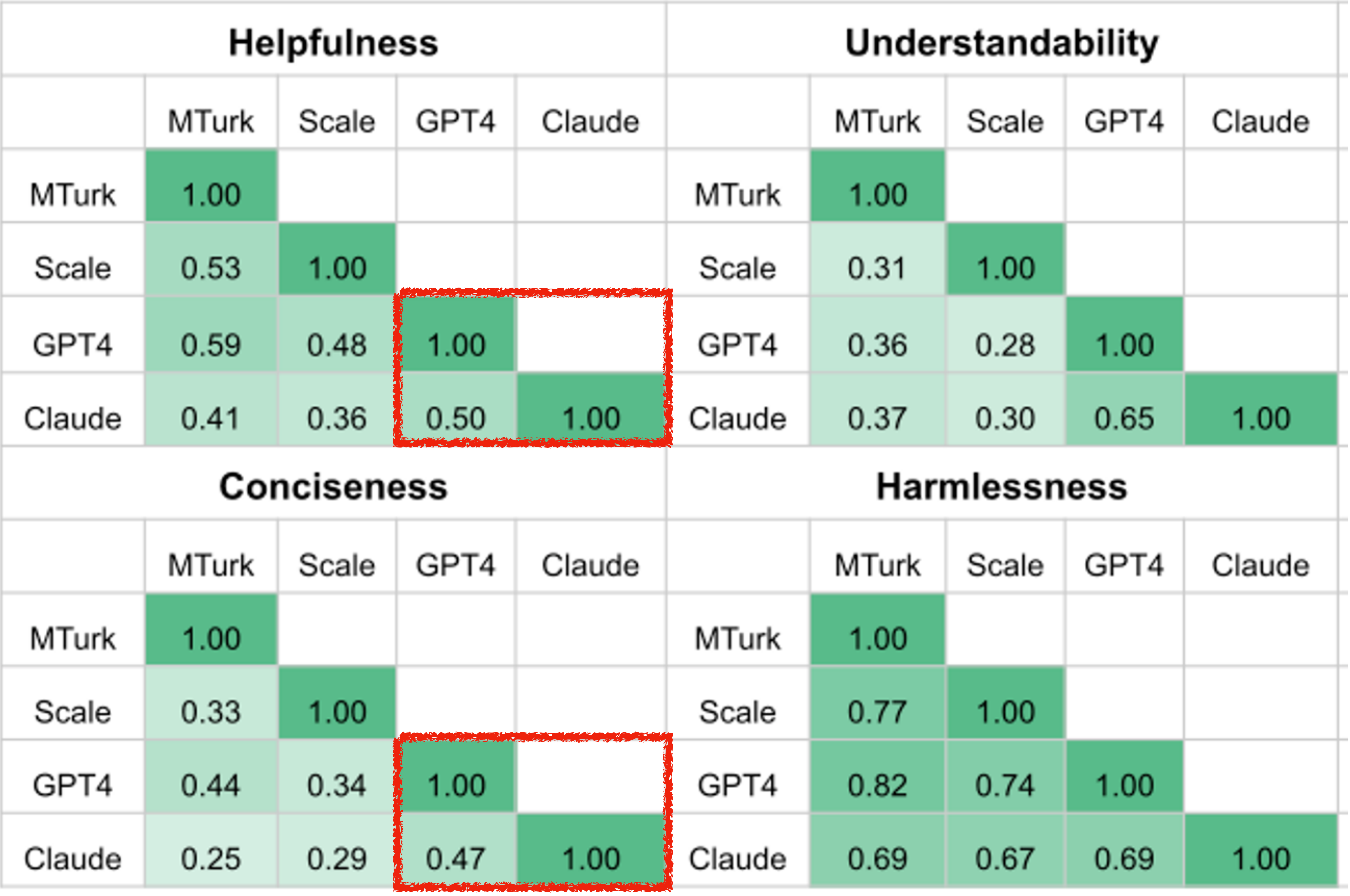
$$\mathbb{E}_{x \sim d_0} \mathbb{E}_{a^1 \sim \pi^1_{t_0}} \left[r^\star(x, a^1) - \eta \mathrm{KL}(\pi^1_{t_0} \| \pi_0)\right] \geq \max_{\pi' \in \Pi} \mathbb{E}_{x \sim d_0} \mathbb{E}_{a^2 \sim \pi'} \left[r^\star(x, a^2) - \eta \mathrm{KL}(\pi' \| \pi_0)\right] - \epsilon\,.$$

Ye C, Xiong W, Zhang Y, et al., Iterative reinforcement learning from human feedback with general preference: from theory to algorithm

On-going Challenges and Future Directions

Challenge 1: Preference Conflict

- The agreement rate among humans is only 70%;
- Even the LLMs have different preferences.



Helpfulness					Understandability				
	MTurk	Scale	GPT4	Claude		MTurk	Scale	GPT4	Claude
MTurk	1.00				MTurk	1.00			
Scale	0.53	1.00			Scale	0.31	1.00		
GPT4	0.59	0.48	1.00		GPT4	0.36	0.28	1.00	
Claude	0.41	0.36	0.50	1.00	Claude	0.37	0.30	0.65	1.00

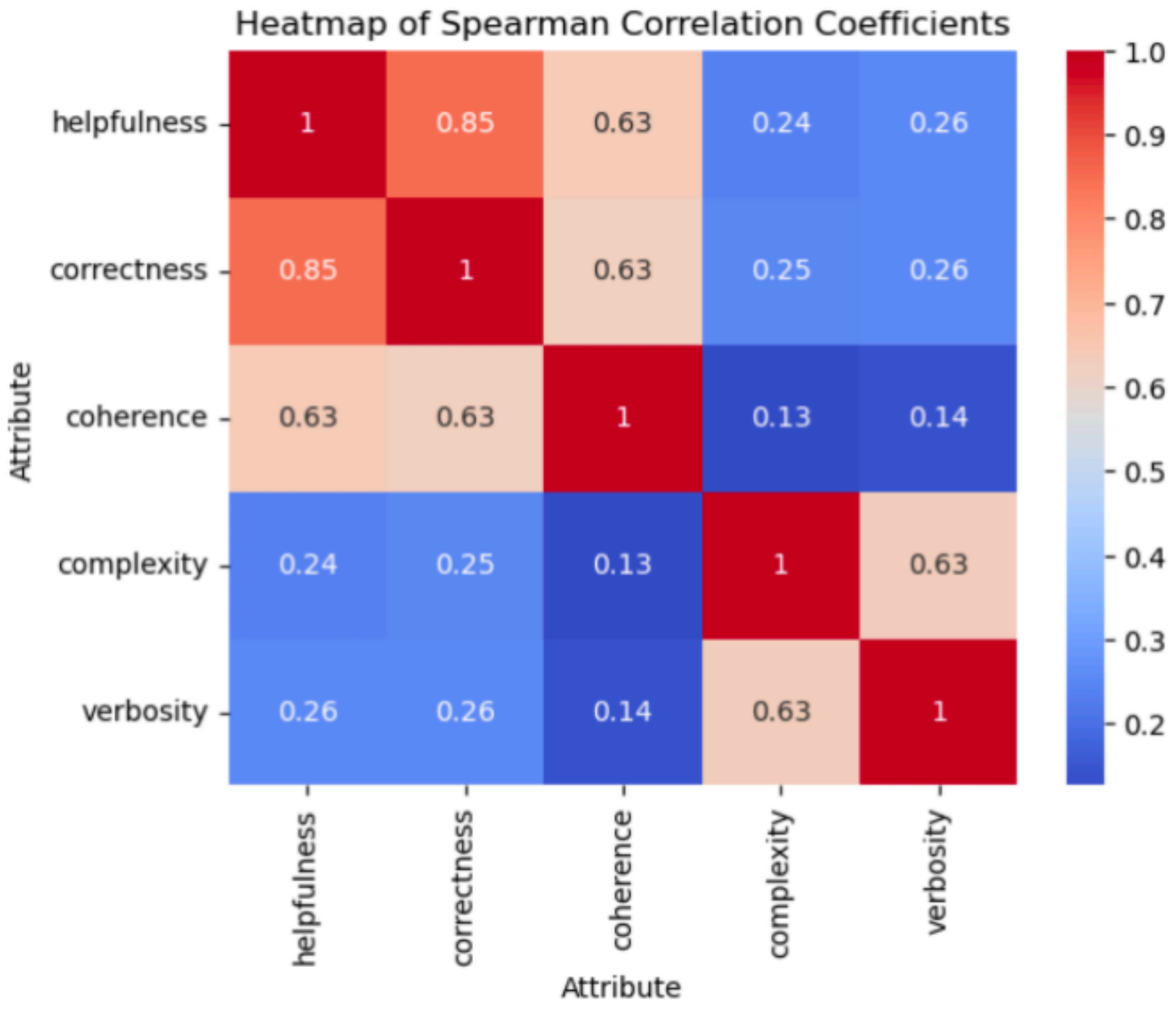
Conciseness					Harmlessness				
	MTurk	Scale	GPT4	Claude		MTurk	Scale	GPT4	Claude
MTurk	1.00				MTurk	1.00			
Scale	0.33	1.00			Scale	0.77	1.00		
GPT4	0.44	0.34	1.00		GPT4	0.82	0.74	1.00	
Claude	0.25	0.29	0.47	1.00	Claude	0.69	0.67	0.69	1.00

Peering Through Preferences: Unraveling Feedback Acquisition for Aligning Large Language Models

HELM Instruct: A Multidimensional Instruction Following Evaluation Framework with Absolute Ratings

Challenge 2: Insufficiency of Scalar Reward

Human possesses *intricate* and even ***contradictory*** targets



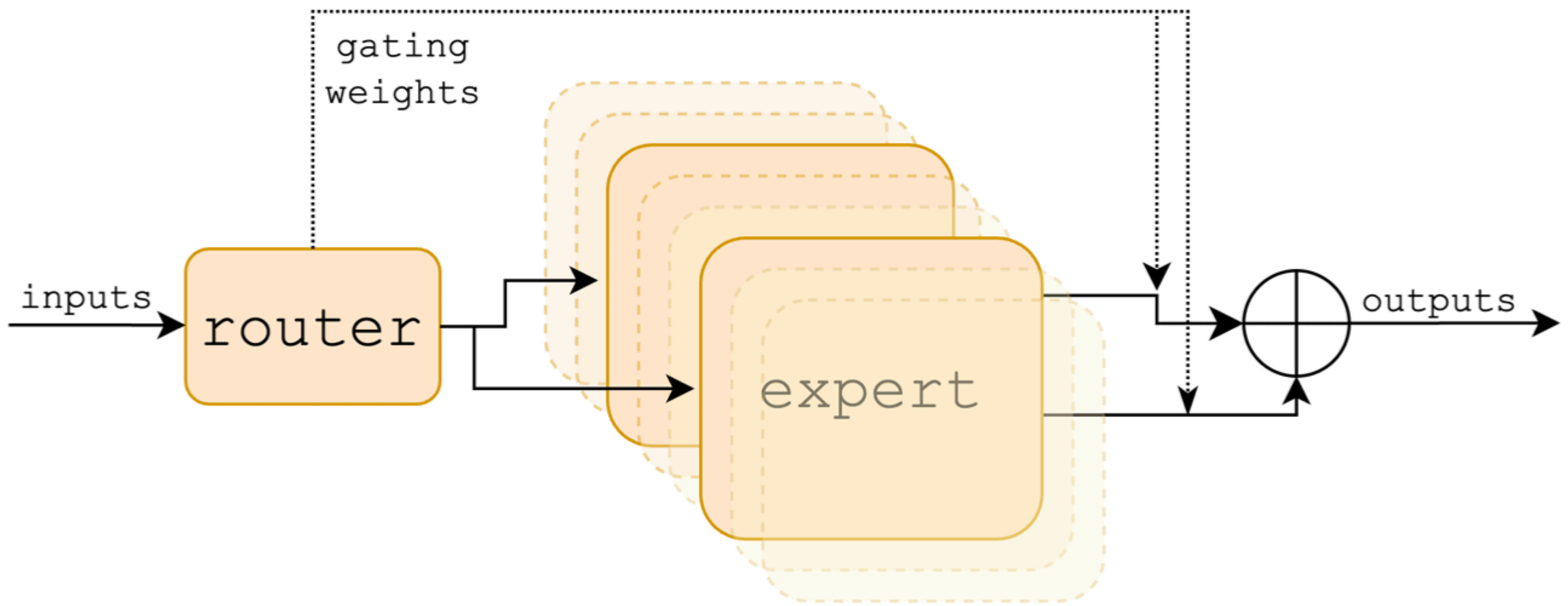
HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM

Multi-objective Reward

1. Reward Modeling: multi-objective rewards $\vec{r} = (r_1, r_2, \cdots, r_k)$
    1. Good performance
    2. Multi-head + Mixture of Expert

$$f(x, a) = \sum_{i=1}^{k} g(x)_i \cdot r_i(x, a)$$

Mixture of Experts Layer

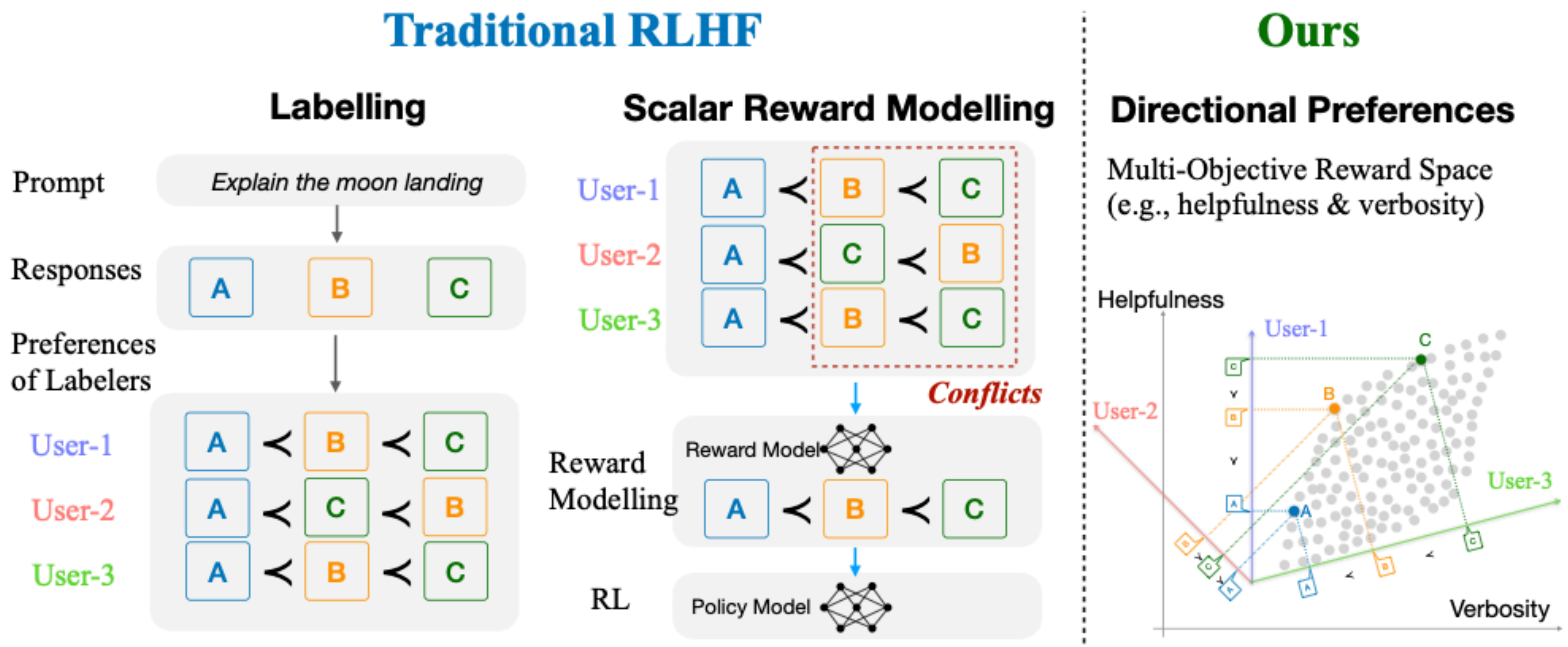


Nathan Lambert et al., RewardBench: Evaluating Reward Models for Language Modeling

User-preference-aware Alignment

1. User-preference-aware objective

$$J(\pi) = \mathbb{E}_{\nu\sim d_\nu}\left[\mathbb{E}_{x\sim d_0, a\sim\pi(\cdot\ \nu,x)} f(\nu,x,a)\right].\qquad f(\nu,x,a) = \sum_{i=1}^{k} g(\nu,x)_i \cdot r_i(x,a)$$



Wang H, Lin Y, Xiong W, et al. Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards

End Note

Central problem: how to model the preference signal

1. Offline learning: pessimism;
2. Online iterative learning: collecting new online data;
3. Use more general preference modeling:
    1. General preference
    2. Multi-objective reward
    3. User-dependent preference
4. Structured problem: math, coding, and agent…

Thanks for listening!