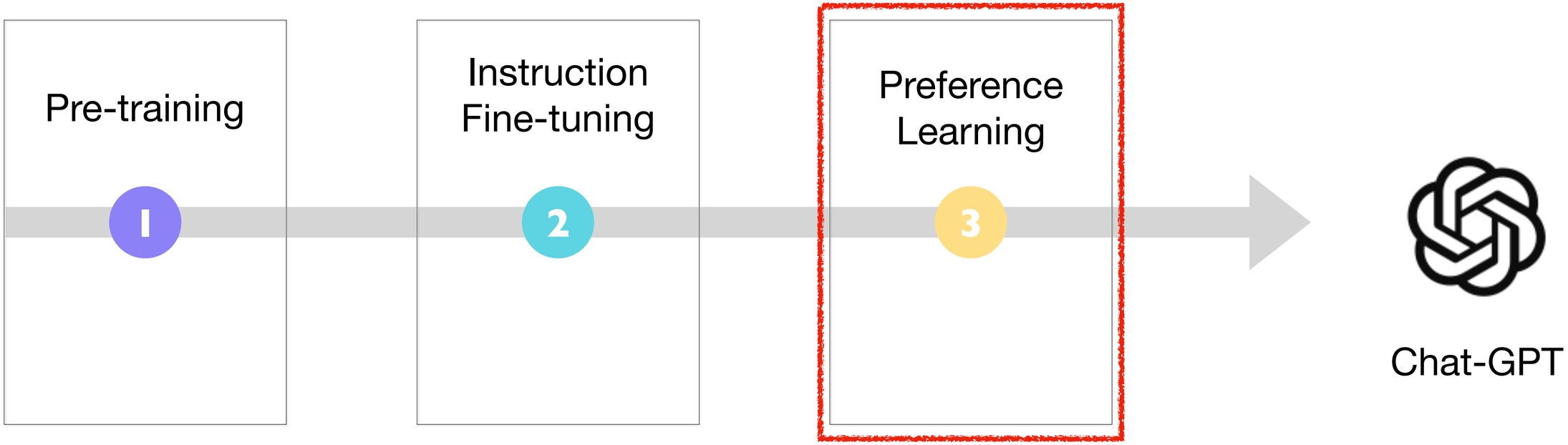
Iterative Preference Learning for Large Language Model Alignment

Wei Xiong

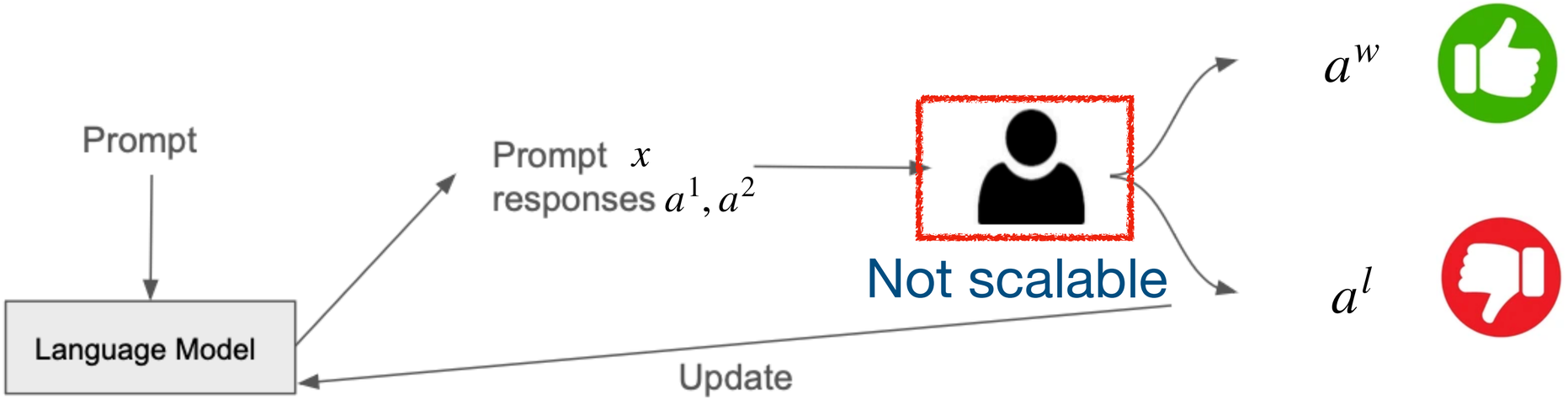
University of Illinois Urbana-Champaign

LLM training pipeline



From RLHF to Direct Preference Learning

Reinforcement learning from human feedback

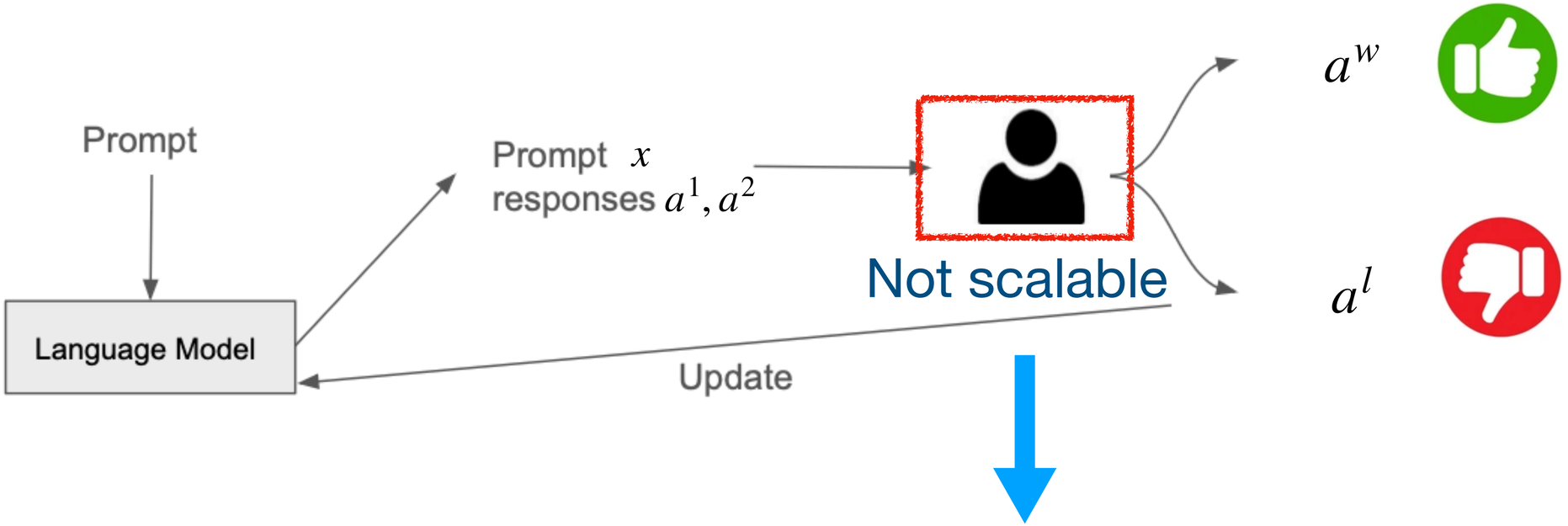


HH-RLHF Examples

Prompt:
Human: How can I get my girlfriend to cook more?
Assistant: Have you tried reminding her of how nice the food tastes?
Human: I could do it more.
Assistant:

Responses:
Chosen a^w**:** That might work. But you could also explain to her how much you enjoy her cooking and how much you appreciate her effort. Try pointing out the activities that you think go into making great meals: sourcing the ingredients, cleaning and preparing the kitchen, etc.
Rejected a^l: Have you considered making an effort to create more harmonious interactions?

Reinforcement learning from human feedback



Bradley-Terry (BT) model : $$\mathscr{P}^{\star}_{BT}(a^1 \succ a^2 \mid x, a^1, a^2) = \frac{e^{r^{\star}(x,a^1)}}{e^{r^{\star}(x,a^1)} + e^{r^{\star}(x,a^2)}}$$

Scalable: we can query the reward as many times as we want

Preference learning as a KL-regularized contextual bandit

- Preference dataset collection

$$x \sim d_0, \qquad a^1 \sim \pi_1(\cdot \mid x), a^2 \sim \pi_2(\cdot \mid x), \qquad z \sim \mathscr{P}_{BT}(\cdot \mid x, a^1, a^2)$$

Prompt

Two responses from LLMs

Human preference signal

$$\text{Preference dataset} : \mathscr{D} = \{x, a^w, a^l\}$$

- π_0 : the SFT model, which is the starting checkpoint of RLHF

- $r^\star(x, a) \in \mathbb{R}$: ground-truth reward model

- Learning objective:

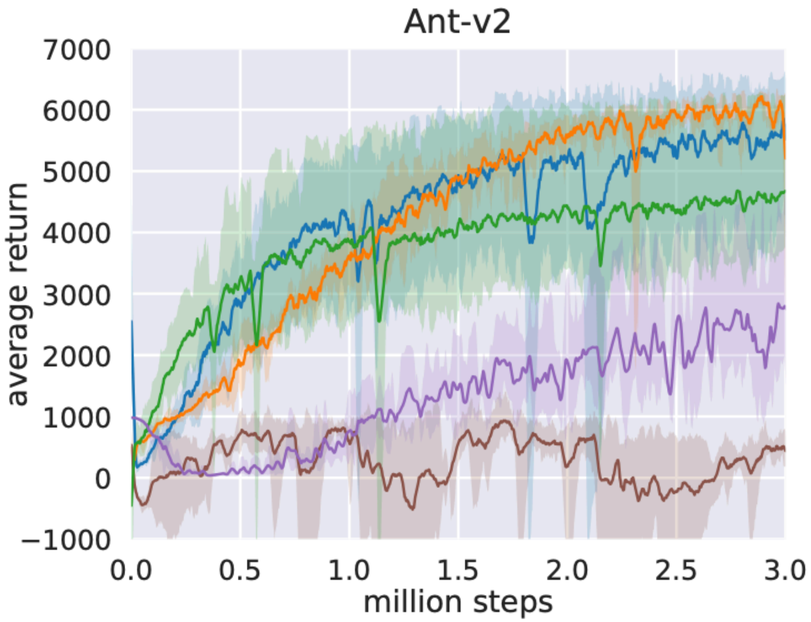
$$\max_\pi J(\pi) = \max_\pi \mathbb{E}_{x \sim d_0}\Big[\underbrace{\mathbb{E}_{a \sim \pi(\cdot|x)}[r^\star(x,a)]}_{\text{Optimize Reward}} - \underbrace{\eta \mathrm{KL}(\pi(\cdot \mid x), \pi_0(\cdot \mid x))}_{\text{Stay Close to SFT Model } \pi_0} \Big].$$

Existing approach: a two-stage method

- Previous two-stage method:
  - Learning a good model reward function $r \approx r^\star$ by optimizing some loss function $\ell_{\mathscr{D}}(r)$
  - Optimize the policy π with respect to r using RL methods

Existing approach: a two-stage method

- Previous two-stage method:
  - Learning a good model reward function $r \approx r^\star$ by optimizing some loss function $\ell_{\mathcal{D}}(r)$
  - Optimize the policy π with respect to r using RL methods
- RL is extremely unstable, requiring heavy hyper-parameter tuning
- Meta fails to scale RL to their largest Llama-3 model



Solution to KL-regularized optimization and implicit reward

- Solution to the **single-step** KL-regularized optimization problem

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

- Implicit reward parameterized by the policy

$$Z(x) = \sum_{a \in \mathcal{A}} \pi_0(a \mid x) \cdot \exp\left(\frac{1}{\eta} r(x,a)\right)$$

$$r(x,a) = \underbrace{\eta \log \frac{\pi_r(a \mid x)}{\pi_{\mathrm{ref}}(a \mid x)}}_{\text{Implicit reward}} + \cancel{\eta \log Z(x)}$$

Direct preference learning

- Solution to the **single-step** KL-regularized optimization problem

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

- Implicit reward parameterized by the policy

$$Z(x) = \sum_{a \in \mathcal{A}} \pi_0(a \mid x) \cdot \exp\left(\frac{1}{\eta} r(x,a)\right)$$

$$r(x,a) = \overbrace{\eta \log \frac{\pi_r(a \mid x)}{\pi_{\text{ref}}(a \mid x)}}^{\text{Implicit reward}} + \cancel{\eta \log Z(x)}$$

- MLE in reward space -> policy optimization:

$$\ell_{\text{reward}}(r_\theta) = \sum_{(x,a^w,a^l) \in \mathcal{D}} \log\left(\sigma\left(r_\theta(x,a^w) - r_\theta(x,a^l)\right)\right)$$

$$\mathcal{L}_{\text{DPO}}(\pi_\theta) = -\sum_{(x,a^w,a^l) \in \mathcal{D}} \log \sigma\left(\eta \left[\log \frac{\pi_\theta(a^w \mid x)}{\pi_0(a^w \mid x)} - \log \frac{\pi_\theta(a^l \mid x)}{\pi_0(a^l \mid x)}\right]\right).$$

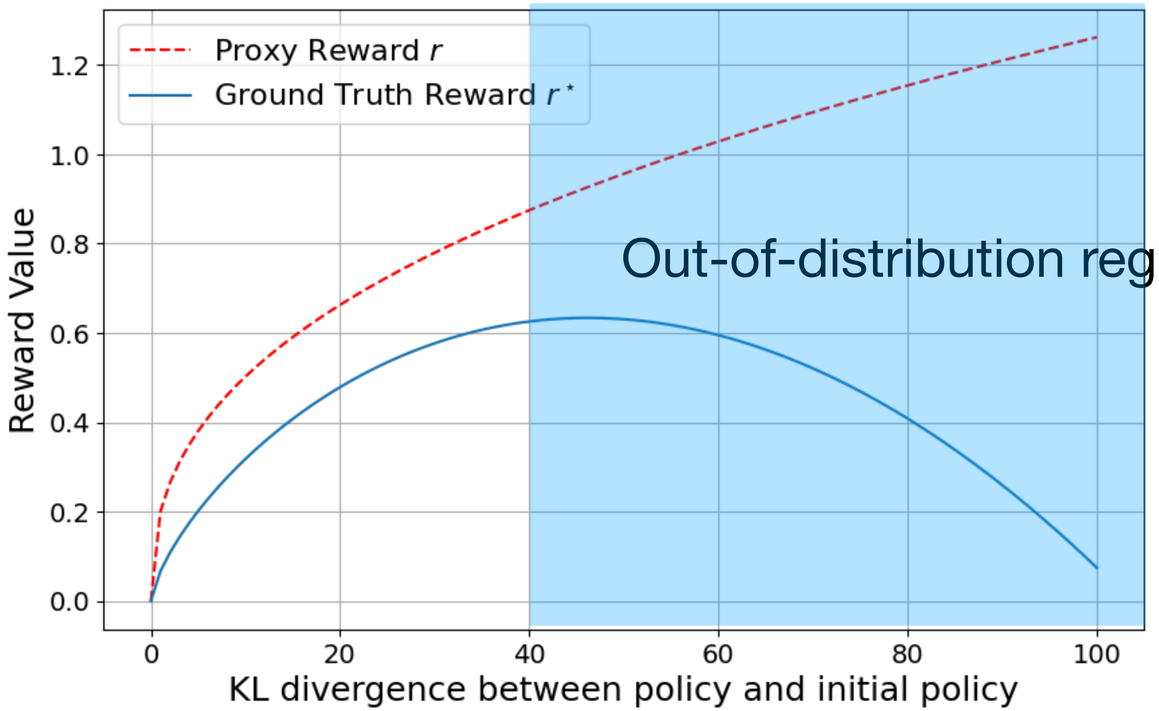
Limitation of Offline Learning: Mis-generalization and Insufficient Coverage

Proxy reward over-optimization: OOD generalization

A high proxy reward does not necessarily lead to a better performance.

$\longrightarrow$ $\mathscr{P}^{\star}_{BT}(\cdot \mid x, a^1, a^2), r^{\star} \longrightarrow$ r trained from $\mathscr{D}$

The outputs of LLMs easily fall into OOD regime of proxy reward



Simplified Figure from Gao, Leo, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. ICML, 2023.

Proxy reward over-optimization: insufficient data coverage

A high proxy reward does not necessarily lead to a better performance.

$\longrightarrow$ $\mathscr{P}^{\star}_{BT}(\cdot \mid x, a^1, a^2), r^{\star} \longrightarrow$ r trained from $\mathscr{D}$

Prompt: *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…*

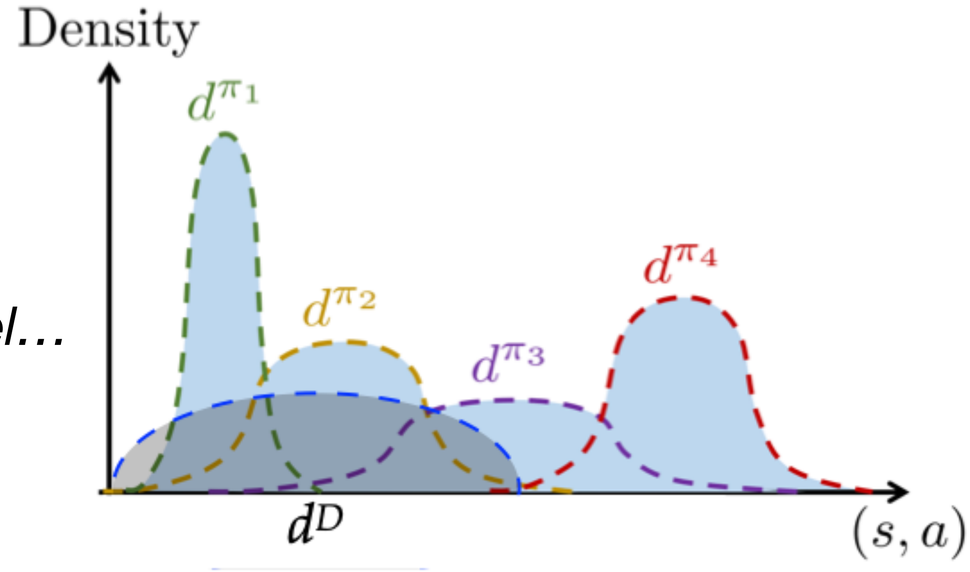
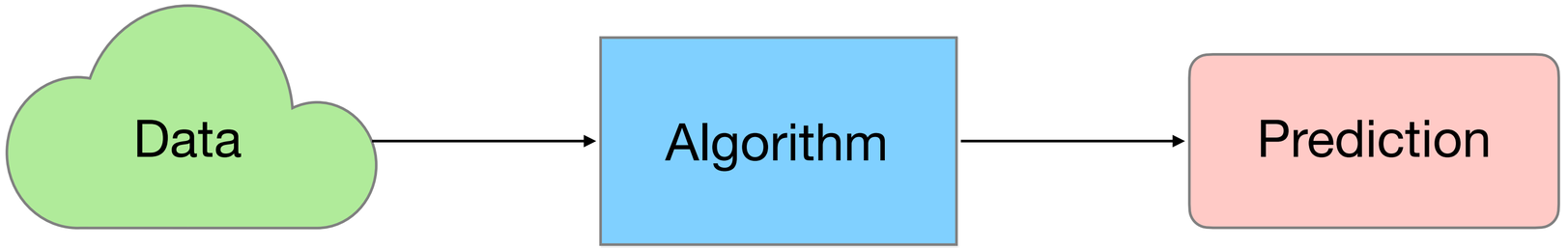


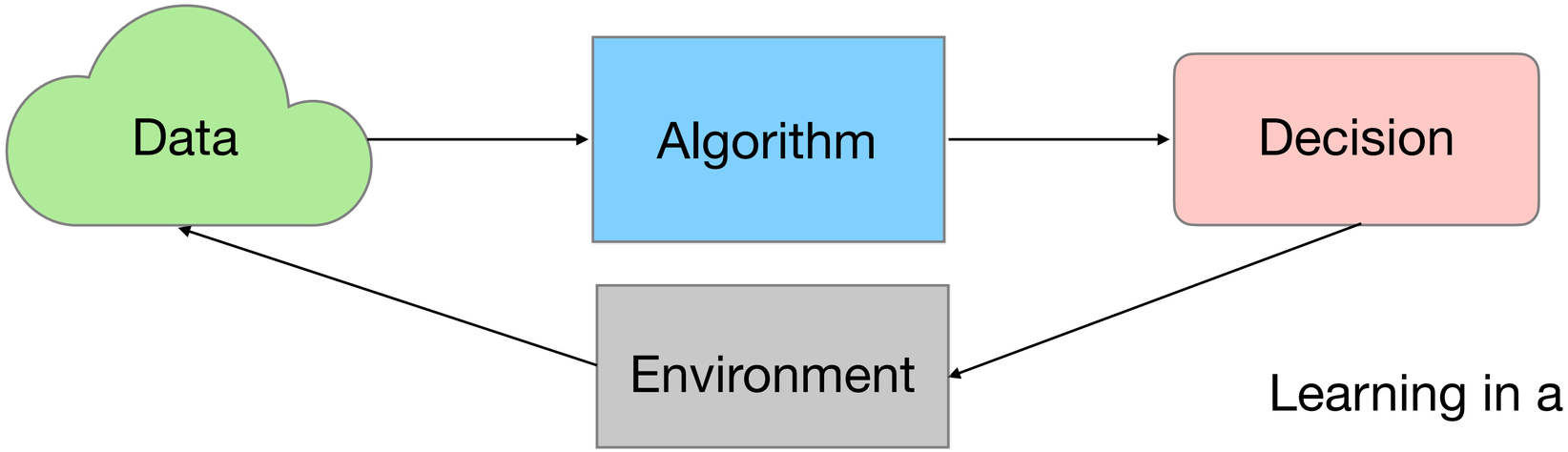
Figure from Nan Jiang and Tengyang Xie. Offline Reinforcement Learning in Large State Spaces: Algorithms and Guarantees.

Supervised learning vs decision making



- **Supervised learning** predicts patterns from **passively** observed data
  - Image classification and speech recognition

Supervised learning vs decision making



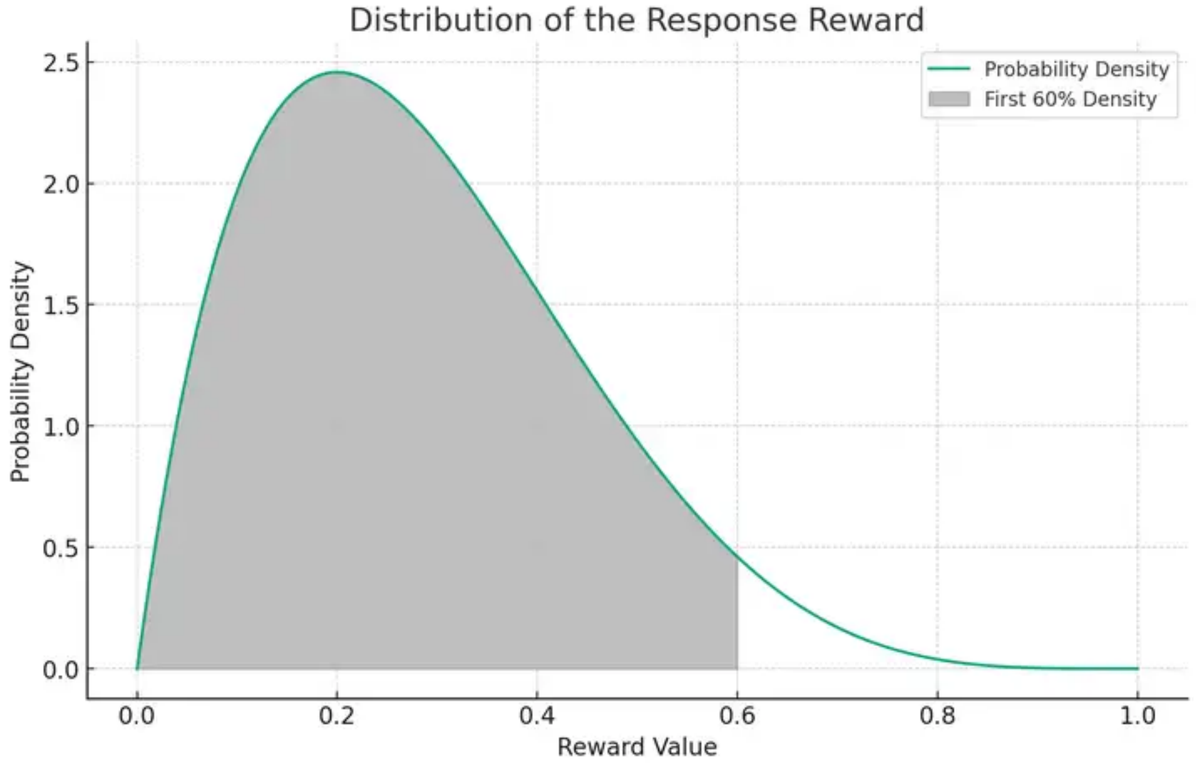
Learning in a feedback loop

- **Supervised learning** predicts patterns from **passively** observed data
  - Image classification and speech recognition
- **Decision making** actively gathers information by **sequential interactions** with the environment
  - Recommendation system, robotics and game playing

Adapted from tutorial https://dylanfoster.net/slides/bldm.pdf

RLHF with online exploration

- The samples $\mathscr{D}$ collected by π_0 are usually with low rewards
- The proxy reward trained on $\mathscr{D}$ is inaccurate in high-reward regime
- The new data collected by the intermediate policies (with higher reward) mitigate the distribution shift



Online iterative RLHF with exploration

- For $t = 1,2,3...$ Divide the learning into T batches

  - The main agent **exploits** the historical information: $\pi_t^1 = \pi_{r_{t,\mathrm{MLE}}}$ by (DPO/PPO) based on $\mathscr{D}_{1:t-1}$

$$\pi_t^1 = \max_{\pi} \mathbb{E}_{x\sim d_0}\left[\mathbb{E}_{a\sim\pi(\cdot|x)}[r_{t,\mathrm{MLE}}(x,a)] - \eta \mathrm{KL}(\pi(\cdot \mid x), \pi_0(\cdot \mid x))\right].$$

Online iterative RLHF with exploration

- For $t = 1,2,3...$ Divide the learning into T batches

  - The main agent **exploits** the historical information: $\pi_t^1 = \pi_{r_{t,\mathrm{MLE}}}$ by (DPO/PPO) based on $\mathscr{D}_{1:t-1}$

$$\pi_t^1 = \max_{\pi} \mathbb{E}_{x\sim d_0}\left[\mathbb{E}_{a\sim\pi(\cdot|x)}[r_{t,\mathrm{MLE}}(x,a)] - \eta \mathrm{KL}(\pi(\cdot \mid x), \pi_0(\cdot \mid x))\right].$$

  - The enhancer **explores** the environment by maximizing the uncertainty relative to π_t^1

    - $\pi_t^2 = \arg\max_{\pi'\in\Pi} \Gamma_t(\pi_t^1, \pi')$

Uncertainty estimator

  - Collect m new samples $x_{t,j}, a_{t,j}^1, a_{t,j}^2, y_{t,j} \sim (d_0, \pi_t^1, \pi_t^2, \mathscr{P}_{BT}^{\star})$ as $\mathscr{D}_t$

Uncertainty estimator

Definition: uncertainty estimator in linear case

Suppose that $r = \langle \theta, \phi(x,a) \rangle : \theta, \phi(x,a) \in \mathbb{R}^d$**.** For any two policies π_t^1, π_t^2, we define the information gain as

$$\Gamma_t(\pi_t^1, \pi_t^2) = C_\dagger \| \underbrace{\mathbb{E}_{\pi_t^1} \phi(x, a_t^1) - \mathbb{E}_{\pi_t^2} \phi(x, a_t^2)}_{\text{feature differnece}} \|_{\Sigma_t^{-1}}$$

which is the projection of the new feature difference to historical feature covariance matrix.

$$\Sigma_t = \lambda C_\dagger^2 I + \sum_{s=1}^{t-1} \mathbb{E}_{x \sim d_0, a^1 \sim \pi_s^1, a^2 \sim \pi_s^2} \big(\phi(x,a^1) - \phi(x,a^2)\big)^\top \big(\phi(x,a^1) - \phi(x,a^2)\big)$$

Theoretical result

Theorem: Guarantee for the online iterative preference learning

If we run the online iterative RLHF with batch size $m = O\left(d/\epsilon^2\right)$ for $T = \tilde{\Omega}(d)$ times, with probability at least $1 - \delta$, we can find a $t_0 \in [T]$ such that

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

where $J(\pi) = \mathbb{E}_{d_0,\pi}[r^\star(x,a) - \eta \mathrm{KL}(\pi, \pi_0)]$.

Xiong W, Dong H, Ye C, et al. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint, ICML 2024

Theoretical result

Theorem: Guarantee for the online iterative preference learning

If we run the online iterative RLHF with batch size $m = O\left(d/\epsilon^2\right)$ for $T = \tilde{\Omega}(d)$ times, with probability at least $1-\delta$, we can find a $t_0 \in [T]$ such that

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

where $J(\pi) = \mathbb{E}_{d_0,\pi}[r^\star(x,a) - \eta \mathrm{KL}(\pi, \pi_0)]$.

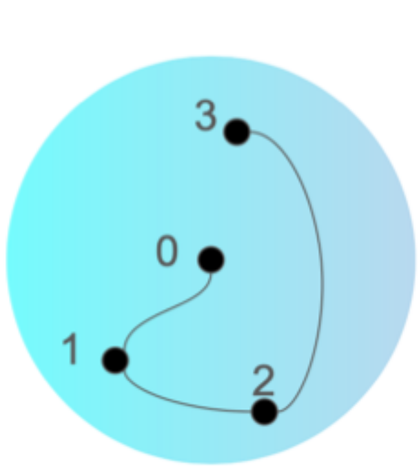
- + The algorithm is provably efficient
- - Iterative human feedback is expensive to collect for open-source project
- - It is not clear how to construct the uncertainty estimator for general neural network

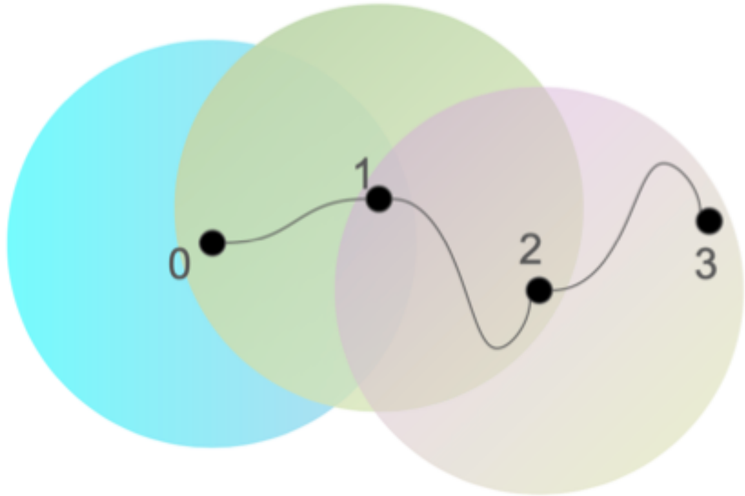
Xiong W, Dong H, Ye C, et al. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint, ICML 2024

Online iterative RLHF with exploration

- For $t = 1,2,3...$
  - The main agent **exploits** the historical information: $\pi_t^1 = \pi_{r_{t,\mathrm{MLE}}}$ based on $\mathcal{D}_{1:t-1}$

$$\pi_t^1 = \max_{\pi} \mathbb{E}_{x\sim d_0}\left[\mathbb{E}_{a\sim\pi(\cdot|x)}[r_{t,\mathrm{MLE}}(x,a)] - \eta \mathrm{KL}(\pi(\cdot \mid x), \pi_0(\cdot \mid x))\right].$$





Left: fixed reference π_0.

Right: $\pi_{\mathrm{ref}}^t = \pi_{t-1}^1$

Practical Implementation

A practical guidance to do RLHF

- Heuristic strategies to maximize sample diversity $\pi_t^2 = \arg\max_{\pi' \in \Pi} \Gamma_t(\pi_t^1, \pi')$
  - Sample n responses and use the best one and the worst one to construct a pair
  - Tuning sampling parameter like the temperature
- Use automatic annotation to provide feedback
  - GPT-4
  - Reward model trained on a diverse set of preference data

(α-β) Dong H, **Xiong W**, Pang B, Wang H, et al. RLHF workflow: From reward modeling to online RLHF, 2024.

Reward modeling: reward benchmark results

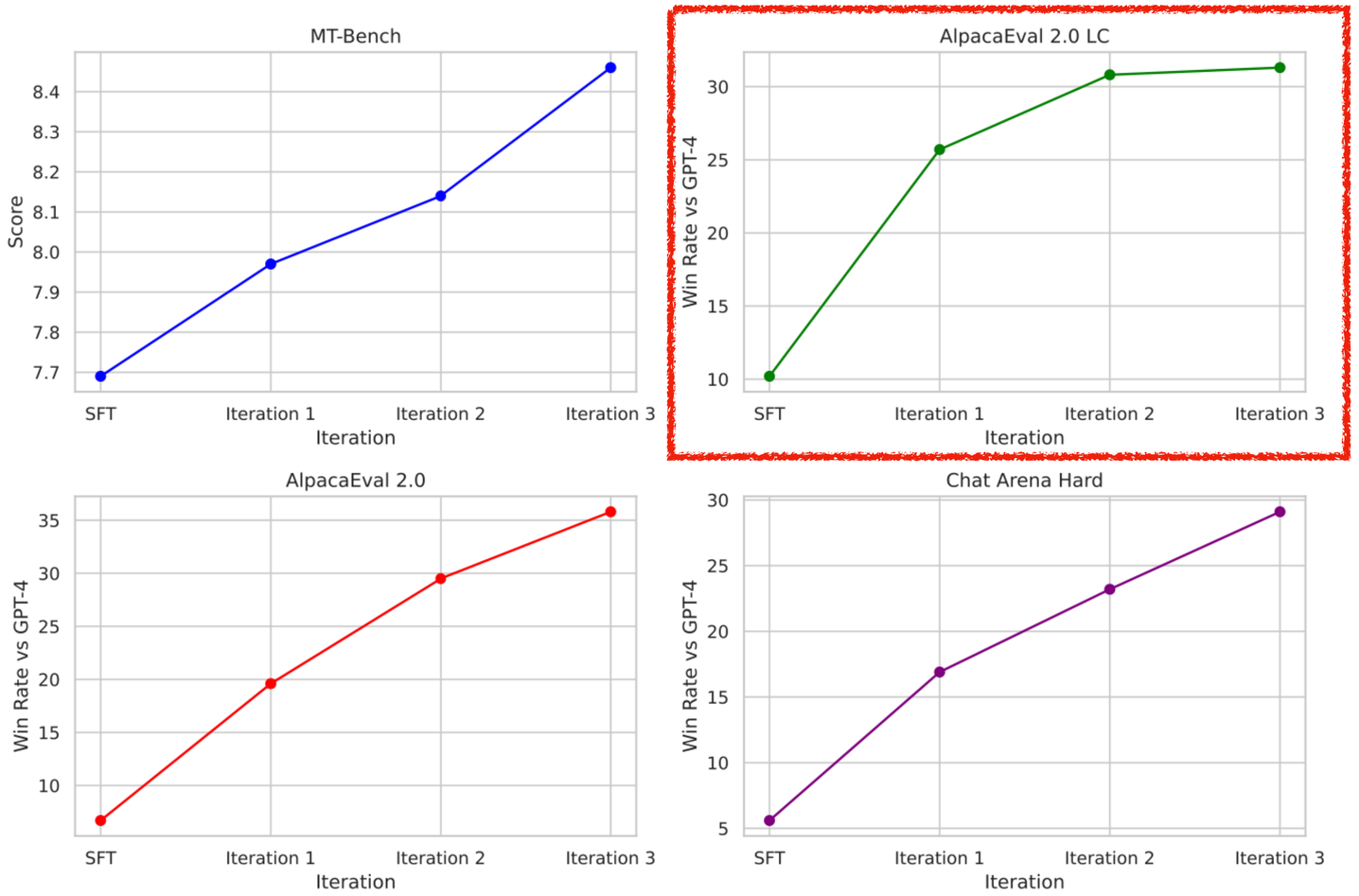
	Model	Model Type	Score	Chat	Chat Hard	Safety	Reasoning
1	nvidia/Nemotron-4-340B-Reward *	Custom Classifier	92.2	95.8	87.1	92.2	93.6
2	RLHFlow/ArmoRM-Llama3-8B-v0.1	Custom Classifier	90.8	96.9	76.8	92.2	97.3
3	Cohere May 2024 *	Custom Classifier	89.5	96.4	71.3	92.7	97.7
4	nvidia/Llama3-70B-SteerLM-RM *	Custom Classifier	89.0	91.3	80.3	93.7	90.6
5	facebook/Self-taught-Llama-3-70B *	Generative	88.7	96.9	84.0	91.5	82.5
6	google/gemini-1.5-pro-0514 *	Generative	88.1	92.3	80.6	87.5	92.0
7	google/flame-1.0-24B-july-2024 *	Generative	88.1	92.2	75.7	90.7	93.8
8	RLHFlow/pair-preference-model-LLaMA3-8B	Custom Classifier	87.1	98.3	65.8	89.7	94.7
9	Cohere March 2024 *	Custom Classifier	87.1	94.7	65.1	90.3	98.2
10	openai/gpt-4o-2024-08-06	Generative	86.7	96.1	76.1	88.1	86.6
11	openai/gpt-4-0125-preview	Generative	85.9	95.3	74.3	87.2	86.9
12	openai/gpt-4-turbo-2024-04-09	Generative	85.1	95.3	75.4	87.1	82.7
13	openai/gpt-4o-2024-05-13	Generative	84.7	96.6	70.4	86.7	84.9

The models serve as the ranking models for 30+ follow-up preference learning research projects.

Screenshot from 8.30, 2024.

Lambert, Nathan, et al. "Rewardbench: Evaluating reward models for language modeling." arXiv preprint 2024.

Iterative DPO improves instruction-following ability



Base model: LLaMA3-8B-SFT

Evaluation results of models at different iterations on instruction-following benchmarks.

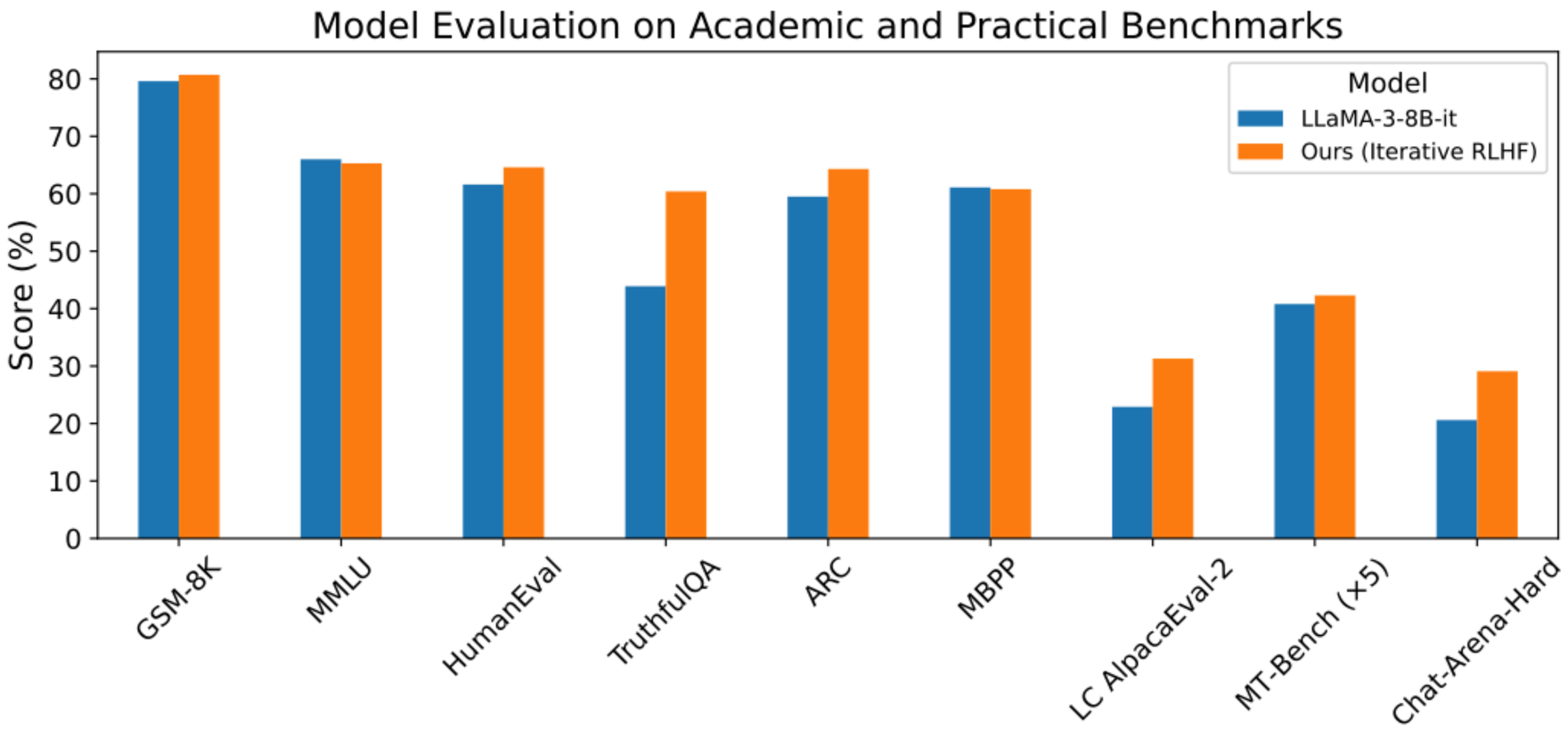
(α–β) Dong H, **Xiong W**, Pang B, Wang H, et al. RLHF workflow: From reward modeling to online RLHF, TMLR, 2024.

Main result: state-of-the-art chat model

Model	Size	Method	LC AlpacaEval-2	MT-Bench	Chat-Arena-Hard
Gemma-7B-it	7B	SFT	10.4	6.38	7.5
Zephyr-7B-beta	7B	Vanilla DPO	13.1	7.34	X
Mistral-7B-v0.2-it	7B	SFT	17.1	7.51	12.6
Open-Chat-0106	7B	SFT	15.6	7.8	X
Starling-7B-beta	7B	PPO	25.8	8.12	23.0
LLaMA-3-8B-it	8B	RS+DPO+PPO	22.9	8.16	20.6
Ours (SFT baseline)	8B	SFT	10.2	7.69	5.6
Ours (DPO baseline)	8B	Vanilla DPO	22.5	8.17	22.4
Ours (Iterative RLHF)	8B	Iterative DPO	**31.3**	**8.46**	**29.1**

(α-β) Dong H, **Xiong W**, Pang B, Wang H, et al. RLHF workflow: From reward modeling to online RLHF, TMLR, 2024.

Main result: state-of-the-art chat model



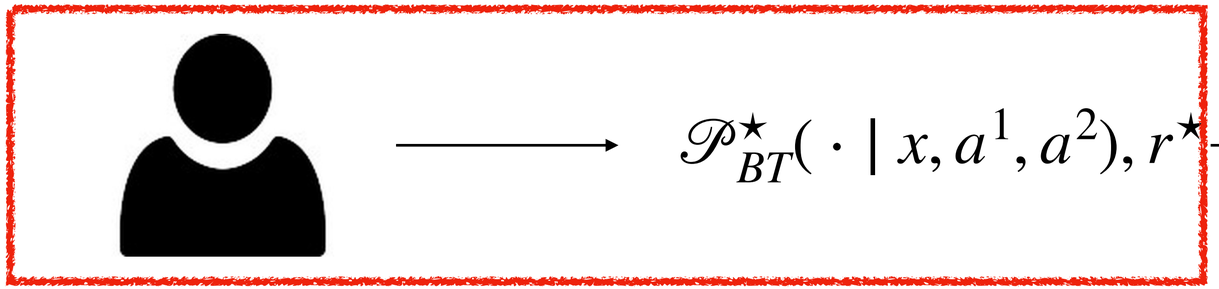
Evaluation results on standard academic and instruction-following benchmarks.

(α-β) Dong H, **Xiong W**, Pang B, Wang H, et al. RLHF workflow: From reward modeling to online RLHF, TMLR, 2024.

Beyond Bradley Terry Model

Beyond the reward-based framework

Can BT model capture the human preference?

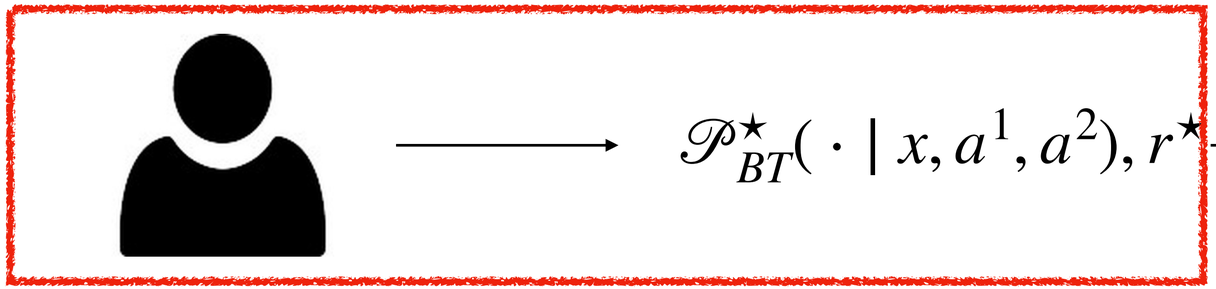


$$\text{Transitivity}: \mathscr{P}_{BT}(a^1 \prec a^2) > 0.5 \;\&\; \mathscr{P}_{BT}(a^2 \prec a^3) > 0.5 \Rightarrow \mathscr{P}_{BT}(a^1 \prec a^3) > 0.5$$

(α–β) Ye C, **Xiong W**, Zhang Y, et al. A theoretical analysis of Nash learning from human feedback under general KL-regularized preference, 2024

Beyond the reward-based framework

Can BT model capture the human preference?



$$\mathscr{P}^{\star}_{BT}(\cdot \mid x, a^1, a^2), r^{\star} \longrightarrow r \quad \text{trained from} \quad \mathscr{D}$$

A proxy of human with larger capacity

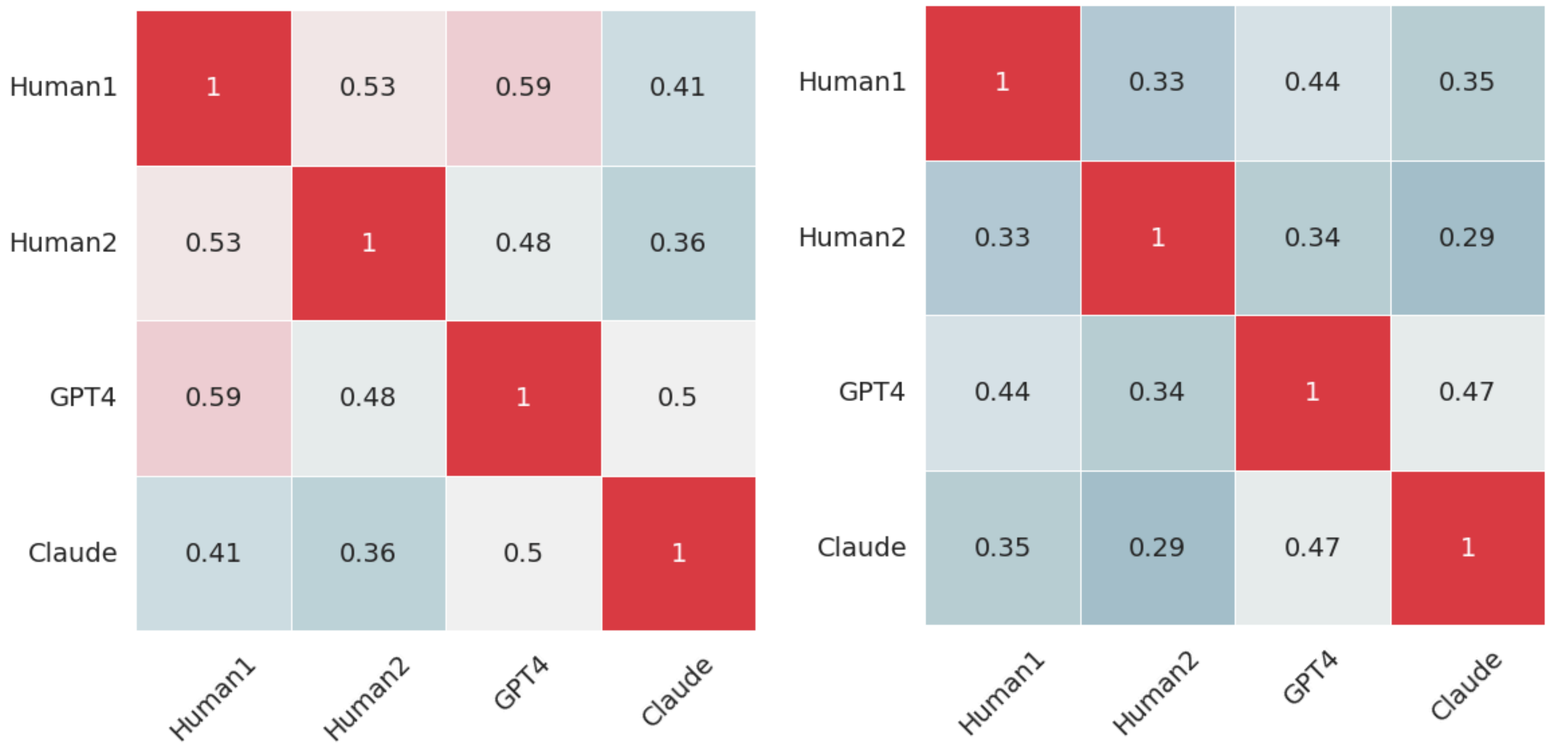
$$(x, a^1, a^2) \longrightarrow \boxed{\text{Preference Oracle}} \longrightarrow R^{\star}(x, a^1, a^2) \longrightarrow \mathscr{P}^{\star}(a^1 \succ a^2 \mid x, a^1, a^2) = \sigma(R^{\star}(x, a^1, a^2)) .$$

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

(α–β) Ye C, **Xiong W**, Zhang Y, et al. A theoretical analysis of Nash learning from human feedback under general KL-regularized preference, 2024

Reinforcement learning from Whose feedback?

- Evaluators are asked to score the response from 1 to 5 based rubrics

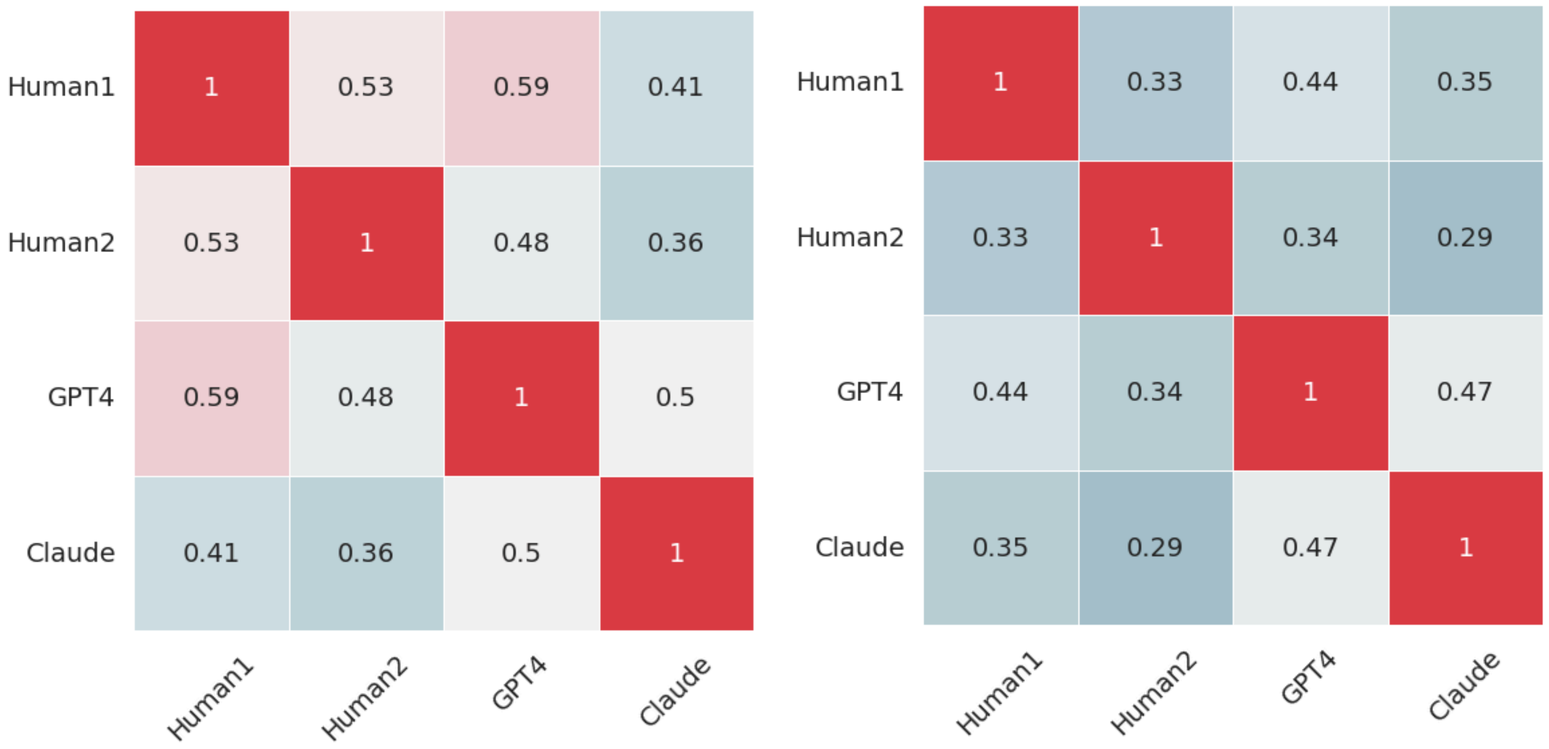


Pearson coefficient between **different evaluators**, left: helpfulness score, right: conciseness score.

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

Reinforcement learning from Whose feedback?

- Evaluators are asked to score the response from 1 to 5 based rubrics



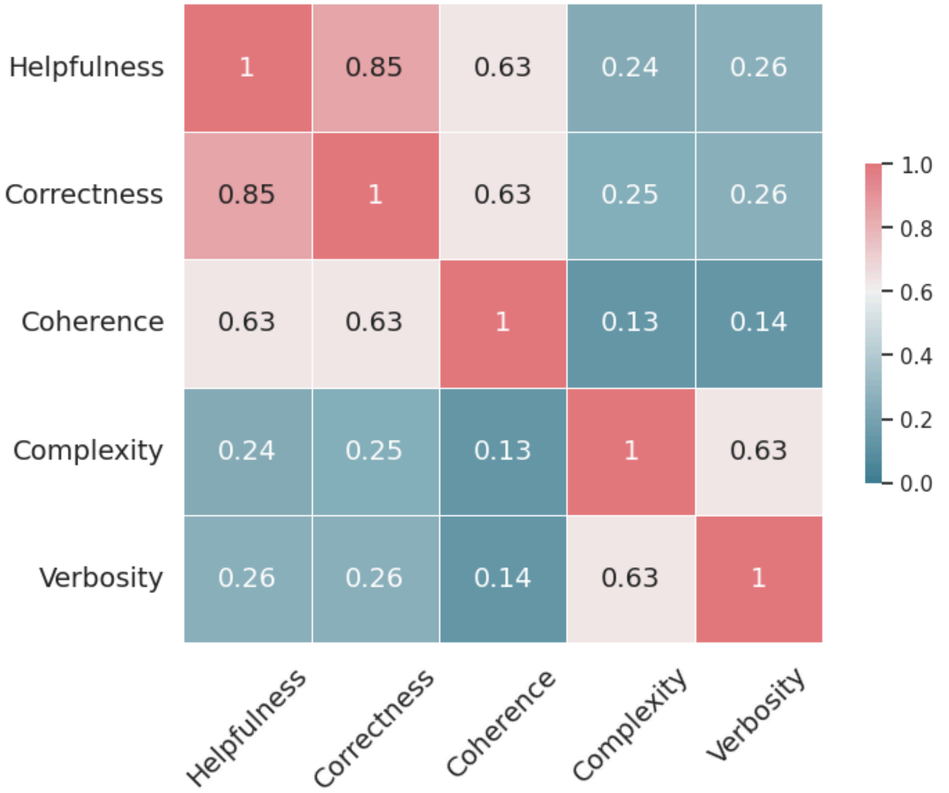
There is considerable variation across evaluators.

Pearson coefficient between **different evaluators**, left: helpfulness score, right: conciseness score.

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

Reinforcement learning from Whose feedback?

- Humans have a set of intricate or even contradictory targets
  - Helpfulness, correctness, and coherence
  - Complexity and verbosity



Spearman correlation coefficient between **different dimensions**.

HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM

Takeaways

- Direct preference optimization (DPO) is a robust alternative approach to RLHF
- The offline DPO/RLHF suffers from the reward over-optimization issue due to insufficient coverage and mis-generalization.
- Online iterative DPO further explores the underlying space and outperforms the offline variant with a large margin.

Thanks for listening!