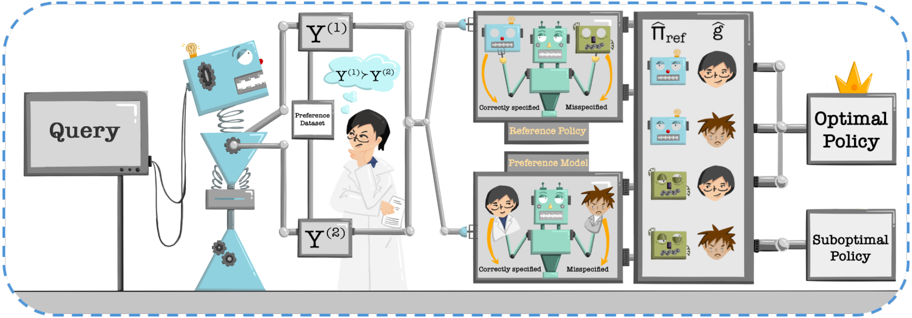
Doubly robust alignment for LLMs

Erhan Xu*, Kai Ye*, Hongyi Zhou*, Luhan Zhu, Francesco Quinzan†, Chengchun Shi†

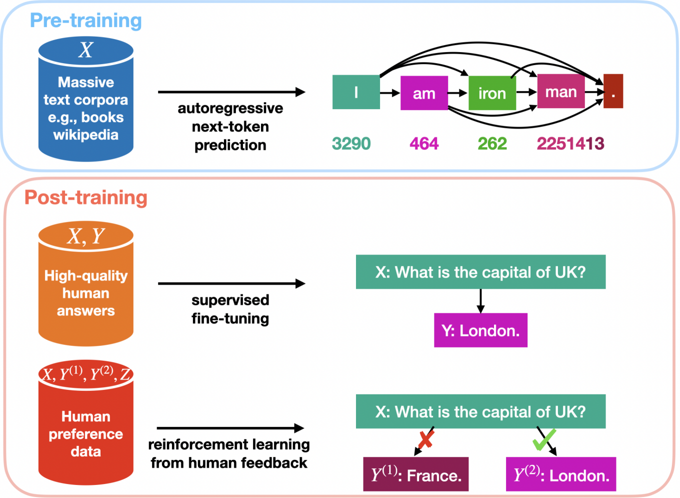
LSE@Stats-Powered AI, Tsinghua Unversity, Oxford University, UAL



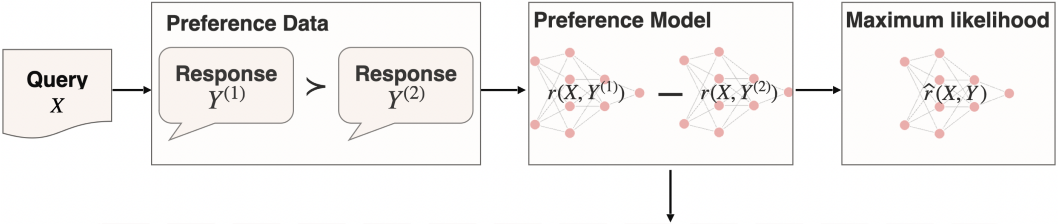
How to train an LLM

Notation

- X: a sentence or prompt.
- Y: responses.
- $Z := \mathbb{I}(Y^{(1)} \succ Y^{(2)})$ represents the resulting human feedback.



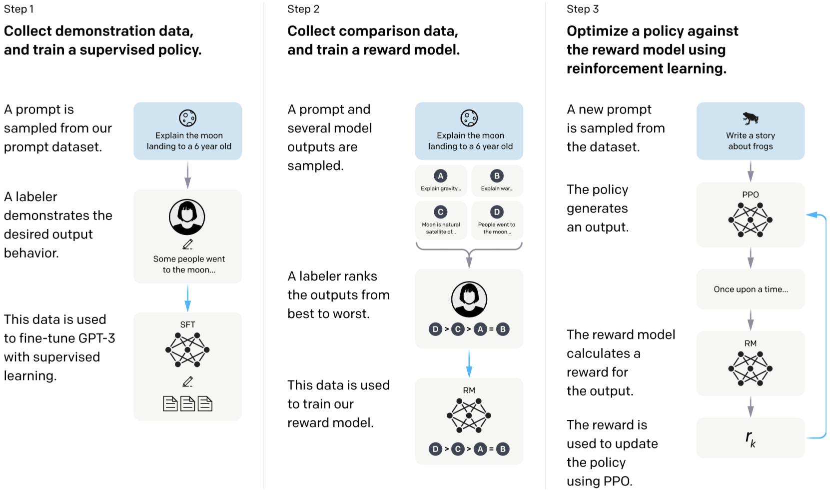
Reward learning in RLHF



Bradley-Terry (BT) model (Bradley & Terry, 1952) is most widely adopted to model human preferences:

$$p(Y^{(1)} \succ Y^{(2)} \mid X) = \sigma(r(X, Y^{(1)}) - r(X, Y^{(2)}))$$

Baseline algorithm I: PPO-based approach



– from InstructGPT (Ouyang et al., 2022)

Baseline algorithm II: DPO-based approach

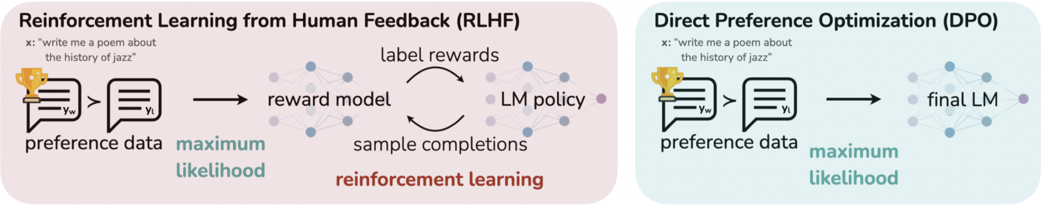


Figure 1: **DPO optimizes for human preferences while avoiding reinforcement learning** **(Rafailov et al., 2023)**

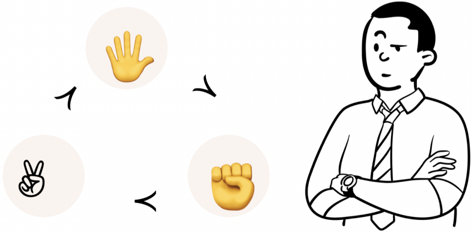
Reward function can be derived in closed-form using the optimal policy

$$r(y, x) = \beta \log\left(\frac{\pi^*(y \mid x)}{\pi_{ref}(y \mid x)}\right) + C(x)$$

BT model can be misspecified

Both **PPO**- and **DPO**-based algorithms rely on **BT model** assumption for human preference modelling, which is likely violated due to **transitivity** ...

What's the best way to learn a new language?		
Practice speaking daily and immerse yourself in the culture through media and conversation.	*Use apps like Duolingo and review flashcards.*	*Join a local language group and travel to countries where the language is spoken.*

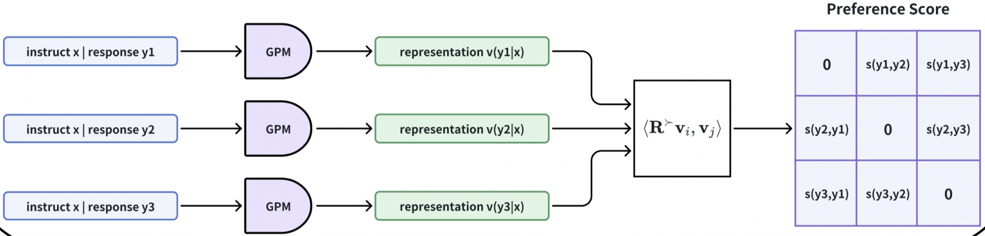


Even when BT model is correct

- **PPO**-based algorithms are highly sensitive to the **reward model**. Misspecifying the reward can
  1. lead to reward hacking (Skalse et al., 2022; Laidlaw et al., 2024)
  2. misguide policy learning (Kaufmann et al., 2023; Zheng et al., 2023; Chen et al., 2024)
- **DPO**-based algorithms are highly sensitive to the **reference policy** (Liu et al., 2024; Gorbatovski et al., 2024; Xu et al., 2024)

Baseline algorithm III: preference-based approach

General preference modelling (GPM, Zhang et al., 2024)



Nash learning from human feedback (NLHF, Munos et al., 2023)

$$\max_{\pi} \min_{\nu} \mathbb{E}_{y^{(1)} \sim \pi, y^{(2)} \sim \nu} p(y^{(1)} \succ y^{(2)})$$

Identity preference optimization (IPO, Azar et al., 2023)

$$\max_{\pi} \mathbb{E}_{y^{(1)} \sim \pi, y^{(2)} \sim \pi_{ref}} p(y^{(1)} \succ y^{(2)})$$

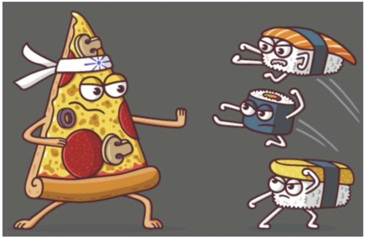
Accurate preference model is vital

Many preference-based approaches do **not** require the BT model assumption. However, they still suffer from potential misspecification of **preference model**

Should I start a pizzeria or sushi restaurant?

Preference: pizza vs sushi

- **In Italy, 80% vs 20%**
- **In Japan, 10% vs 90%**



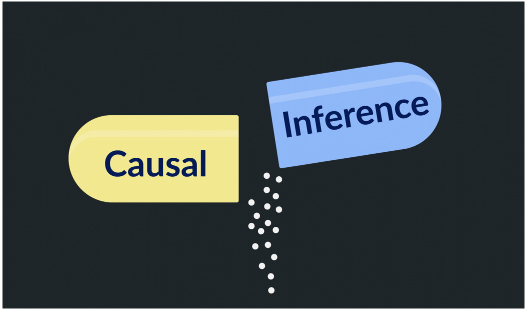
In summary, all three baseline algorithms suffer from certain model misspecification

		Robust to misspecified:	preference model	reward model	reference policy
RLHF	Reward-based	PPO-based	✗	✗	✓
		DPO-based	✗	✓	✗
	Preference-based	IPO	✓	-	✗
		GPM	✗	-	✓
		DRPO	✓	✓	✓

Table: Robustness of different algorithms to model misspecification. Our algorithm is denoted by DRPO, short for doubly robust preference optimization.

Doubly robust (DR) methods

Doubly robust methods originate from the **missing data** and **causal inference** literature (see e.g., Robins et al., 1994; Scharfstein et al., 1999)



Doubly robust methods (Cont'd)

Consider the estimation of **average treatment effect** (ATE) in causal inference. These methods estimate two models:

- A **propensity score** model for treatment assignment mechanism
- Similar to **reference policy** in LLMs

- An **outcome regression** model for patient's outcome given treatment
- Similar to **reward model** in LLMs

- Consistency of the ATE estimator only requires **one** model to be correct
- When **both** are correct, the ATE estimator becomes **semiparametrically efficient**

When DR methods meet LLMs

- **Preference evaluation:** for any target policy $\boldsymbol{\pi}$, evaluate its **total preference**

$$p(\boldsymbol{\pi}) = \mathbb{E}_{y^{(1)}\sim\pi, y^{(2)}\sim\pi_{ref}} p(y^{(1)} \succ y^{(2)})$$

  We estimate two models from the data:

  1. a preference model
  2. a reference policy[1]

  and develop a **doubly robust** and **semiparametrically efficient** estimator $\widehat{p}(\boldsymbol{\pi})$

- **Preference optimization:**

$$\widehat{\pi} = \arg\max_{\pi} \widehat{p}(\boldsymbol{\pi}) - \beta \mathrm{KL}(\pi, \widehat{\pi}_{ref})$$

[1] In practice, usually we directly use a pre-trained or SFT model

More detailed details: DRPE

- denote $g(X, Y^{(1)}, Y^{(2)}) := \mathbb{P}(Y^{(1)} \succ Y^{(2)} \mid X)$:
  - PPO-based: $\mathbb{E}_{X\sim\mathcal{D},\, y\sim\pi(\cdot|X)}\left[\widehat{r}(y, X)\right] - \beta\,\mathrm{KL}\left[\pi(y \mid X) \,\|\, \pi_{\mathrm{ref}}(y \mid X)\right]$
  - DPO-based: $\widehat{r}(y, x) = \beta \log\left(\frac{\widehat{\pi}(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right) - C(x)$
- DR Policy Evaluation:

$$\widehat{p}_{\mathrm{DR}}(\pi) = \frac{1}{2}\mathbb{E}_{(X,Y^{(1)},Y^{(2)},Z)\sim\mathcal{D}}\Bigg\{\sum_{a=1}^{2}\mathbb{E}_{y\sim\pi(\cdot|X)}[\widehat{g}(X, y, Y^{(a)})]$$
$$+\sum_{a=1}^{2}(-1)^{a-1}\frac{\pi(Y^{(a)}|X)}{\widehat{\pi}_{\mathrm{ref}}(Y^{(a)}|X)}[Z - \widehat{g}(X, Y^{(1)}, Y^{(2)})]\Bigg\}$$

More detailed details: DRPO

- DRPO Loss function $\mathcal{L}_{\mathsf{DRPO}}$:

$$
\begin{aligned}
&-\frac{1}{2}\mathbb{E}_{X,Y^{(1)},Y^{(2)}\sim\widetilde{\mathcal{D}}}\Bigg[\underbrace{\mathbb{E}_{Y^*\sim\mathcal{D}_X^*}\left[\widehat{g}(Y^*,Y^{(2)},X)\log\pi_\theta(Y^*|X)\right]}_{\text{term I}}\\
&+\mathrm{sg}\Bigg(\underbrace{\mathrm{clip}\left(\frac{\pi_\theta(Y^{(1)}|X)}{\pi_{\mathrm{ref}}(Y^{(1)}|X)},1-\epsilon_1,1+\epsilon_2\right)\left(Z-\widehat{g}(Y^{(1)},Y^{(2)},X)\right)}_{\text{term II}}\Bigg)\log\pi_\theta(Y^{(1)}\mid X)\Bigg]\\
&+\beta\mathbb{E}_{Y^*\sim\mathcal{D}_X^*,X\sim\widetilde{\mathcal{D}}}\left[\frac{\widehat{\pi}_{\mathrm{ref}}(Y^*\mid X)}{\pi_\theta(Y^*\mid X)}-1-\log\frac{\widehat{\pi}_{\mathrm{ref}}(Y^*\mid X)}{\pi_\theta(Y^*\mid X)}\right]
\end{aligned}
$$

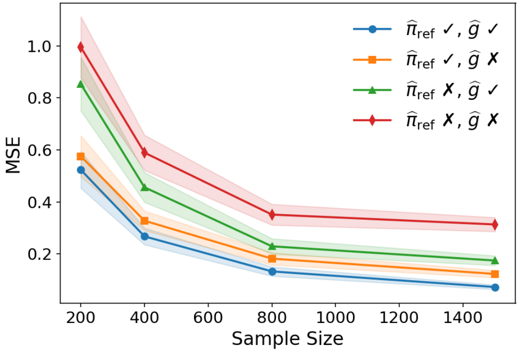
- sg: stop gradient (detach); $\mathrm{clip}(\bullet, a, b)$: clip to range $[a, b]$
- hyperparameters: β, ϵ_1, ϵ_2, temperature of policies; size of $\mathcal{D}_X$ (minor).

More details: Theory

- **Preference evaluation**
  - *Double robustness* of $\widehat{p}(\pi)$: MSE of $\widehat{p}(\pi)$ decays to zero when *either* reference policy *or* preference model (not necessarily both) is correct
  - *Semiparametric efficiency*: When both models are "approximately" correct, $\widehat{p}(\pi)$ achieves the *efficiency bound* (the smallest-possible MSE one can hope for $p(\pi)$)
- **Preference optimization**
  - *Double robustness* of $\widehat{\pi}$: Regret of $\widehat{\pi}$ decays to zero when *either* reference policy *or* preference model (not necessarily both) is correct
  - *Performance gaps*:
    - PPO: $O(n^{-1/2} + \|\widehat{r} - r\|)$
    - DRPO: $O(n^{-1/2} + \|\widehat{r} - r\| \|\widehat{\pi}_{ref} - \pi_{ref}\|)$
    - DPO: $O(n^{-1/2} + \|\widehat{\pi}_{ref} - \pi_{ref}\|)$

Application to IMDb dataset

- **Task**: produce positive movie reviews
- **Objective**: evaluate total preference of a DPO-trained policy over a SFT-based reference policy
- **Ground truth**: 0.681



Applications to TL;DR and HH datasets

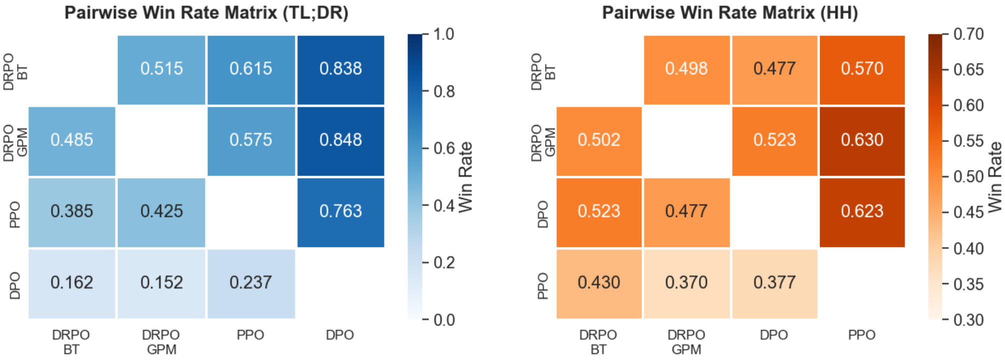


Figure: **Pairwise win rate** matrices between different methods across two datasets. **Left:** TL;DR dataset. **Right:** HH dataset. Each entry indicates how often the row method outperforms the column method; higher values denote better performance.

More Baselines and Benchmarks

Win Rate on TL;DR

Against...	Win Rate (%)
DR. DPO	72.5
rDPO	65.0
cDPO	63.5
CPO	90.0
ORPO	57.5
IPO	98.5
RSO	69.5

AlpacaEval for HH

Model	LC Win Rate (%)	Win Rate (%)
DPO	83.90	84.09
DR. DPO	92.16	90.93
rDPO	86.89	85.71
cDPO	85.05	84.28
CPO	73.59	71.28
ORPO	75.92	53.91
IPO	78.29	78.88
RSO	80.62	79.50
DRPO	86.38	84.84

Takeaways

- **Methodology**
  1. Propose a robust and efficient estimator for preference evaluation (DRPE)
  2. Leveraging this estimator, develop a doubly robust preference optimization (DRPO) algorithm for RLHF
- **Theory**
  1. Doubly robustness
  2. Statistical efficiency
- **Application to LLMs**
  1. Superior and more robust performance than PPO- and DPO-based approaches
  2. Orthogonal to other robust RLHF algorithms that address noisy preferences

Thank You!

☺ Code can be found on GitHub

`https://github.com/DRPO4LLM/DRPO4LLM`