RL from Human Feedback as Game-Solving

Gokul Swamy

Outline for Today

1. When is the Bradley-Terry assumption inaccurate and what happens to online / offline PFT as a result?

2. What is a more robust criterion for preference aggregation and how can we efficiently optimize it?

Outline for Today

1. When is the Bradley-Terry assumption inaccurate and what happens to online / offline PFT as a result?

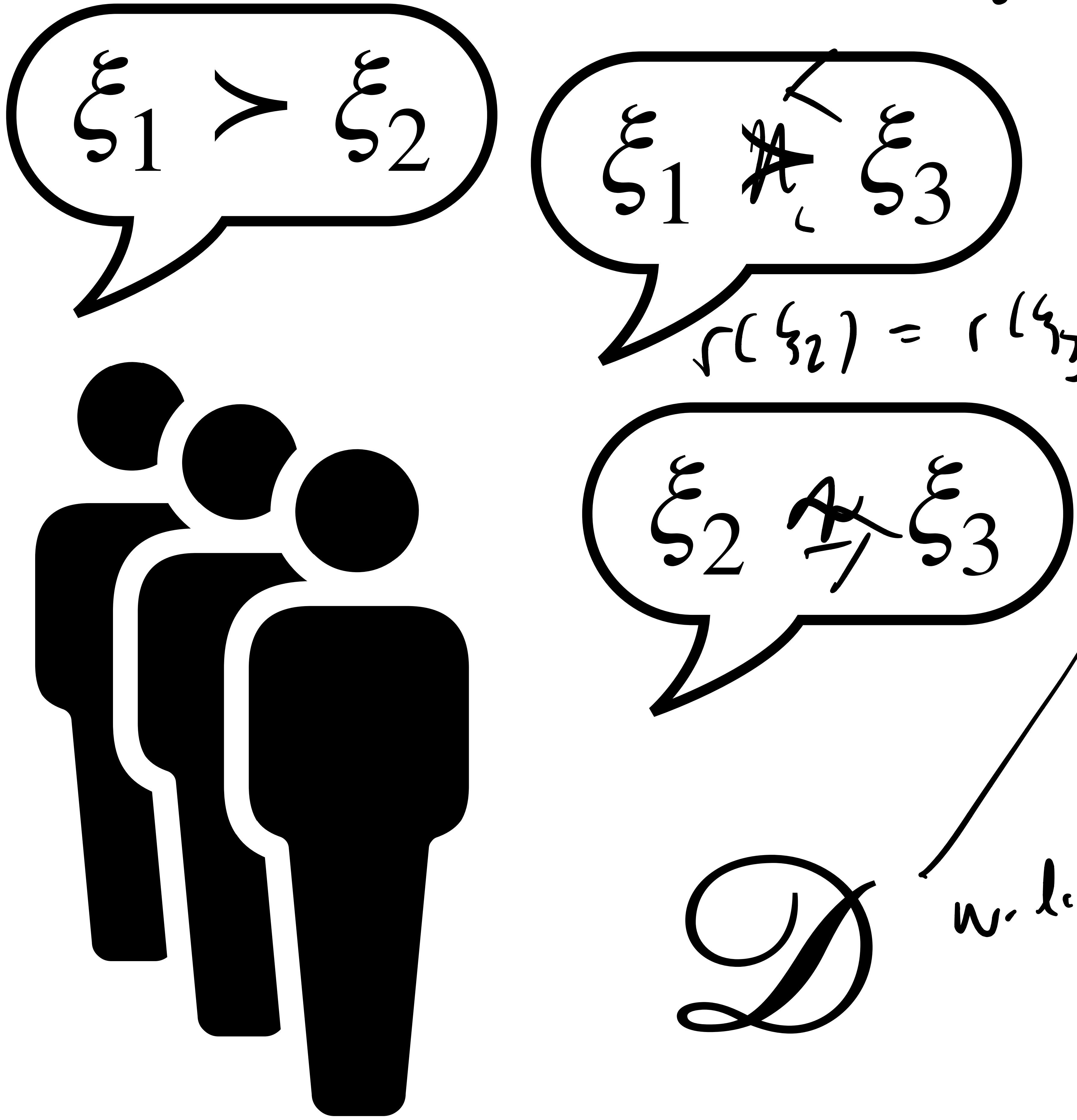
A: *BT is violated when when a reward function can't explain (aggregate) preferences, leading to mode collapse in RLHF.*

2. What is a more robust criterion for preference aggregation and how can we efficiently optimize it?

Preference Matrix

$r = [1, 0, 0]$

$r(\xi_1) > r(\xi_2)$ $\quad$ $r(\xi_1) > r(\xi_3)$



$\mathscr{D}$ w.l.o.g.

$$\mathscr{P}(\xi_1 \succ \xi_2) \triangleq \mathbb{P}_{\mathscr{D}}(\xi_1 \succ \xi_2)$$

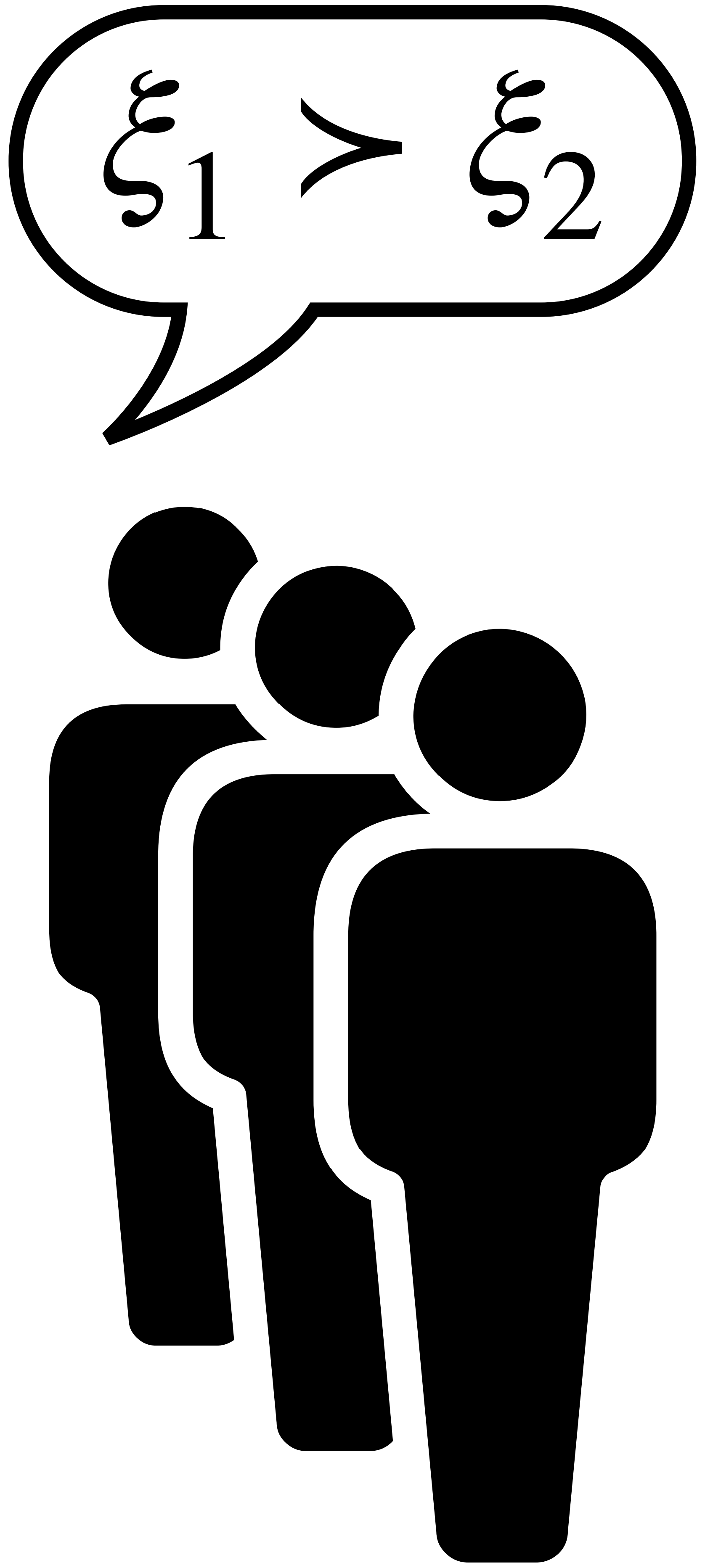
$\mathscr{P}$	ξ_1	ξ_2	ξ_3
ξ_1	0.5	1	1
ξ_2	0	0.5	0.5
ξ_3	0	0.5	0.5

$1-x$

0.5

(Partial) Preference Matrix

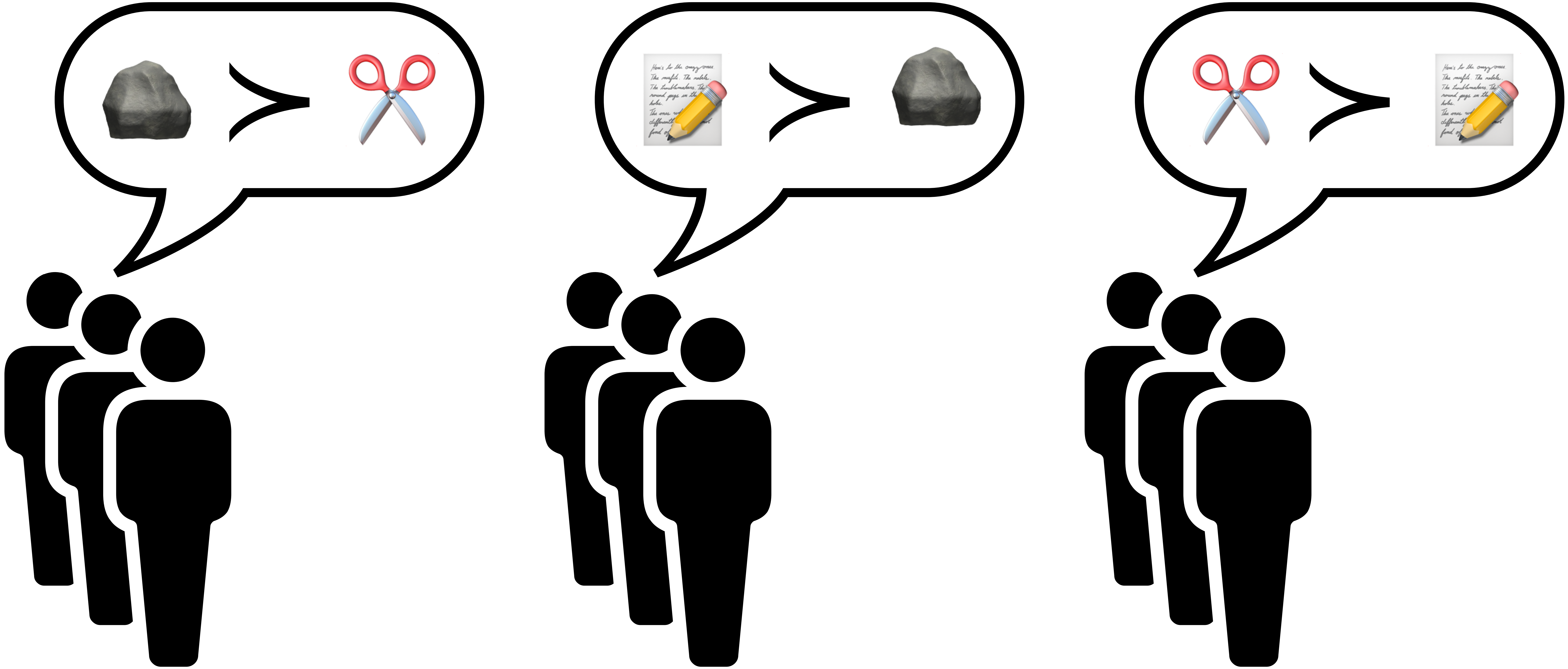
$$\mathscr{P}(\xi_1 \succ \xi_2) \triangleq \mathbb{P}_{\mathscr{D}}(\xi_1 \succ \xi_2)$$



$\mathscr{D}$

$\mathscr{P}$	ξ_1	ξ_2	ξ_3
ξ_1	0.5	1	?
ξ_2	0	0.5	?
ξ_3	?	?	0.5

Intransitivity from Preference Aggregation



Intransitivity from Preference Aggregation

🪨 > ✂️

📝 > 🪨

✂️ > 📝

	🪨	📝	✂️
🪨	0.5	?	1
📝	?	0.5	?
✂️	0	?	0.5

	🪨	📝	✂️
🪨	0.5	0	?
📝	1	0.5	?
✂️	?	?	0.5

	🪨	📝	✂️
🪨	0.5	?	?
📝	?	0.5	0
✂️	?	1	0.5

Intransitivity from Preference Aggregation

$r($🪨$) > r($✂️$)$ $r($📝$) > r($🪨$)$ $r($✂️$) > r($📝$)$

	🪨	📝	✂️	
🪨	0.5	0	1	$\Rightarrow$ 1.5
📝	1	0.5	0	$\Rightarrow$ 1.5
✂️	0	1	0.5	$\Rightarrow$ 1.5

There is no $r^\star$ that explains these preferences!

No $r^\star$ Leads to Mode Collapse in RLHF

$$\pi_{\text{ref}} = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$$

$$\hat{r}_{\text{mle}} = [\boxed{1.2}, \boxed{1.8}, \boxed{1.5}]$$

$$\hat{\pi}_{\text{rlhf}} \neq \cancel{\pi_{\text{ref}} \cdot \exp\left(\frac{1}{\beta}\hat{r}_{\text{mle}}\right)}$$

$L \to 0.05$

$L \to 0$

	🪨	📝	✂️	
🪨	0.5	0.7	0	= 1.2
📝	0.3	0.5	1	= 1.8
✂️	1	0	0.5	= 1.5

No $r^\star$ Leads to Mode Collapse in RLHF

	🪨	📝	✂️	…
🪨	0.5	0.7	0	…
📝	0.3	0.5	1	…
✂️	1	0	0.5	…
…	…	…	…	…

This problem gets worse with a larger Ξ!

Outline for Today

1. When is the Bradley-Terry assumption inaccurate and what happens to online / offline PFT as a result?

A: *BT is violated when when a reward function can't explain (aggregate) preferences, leading to mode collapse in RLHF.*

2. What is a more robust criterion for preference aggregation and how can we efficiently optimize it?

A: *The minimax winner doesn't assume transitivity of preferences. We can use a self-play algorithm to compute it.*

Beyond Bradley-Terry in RLHF

	🪨	📝	✂️
🪨	0.5	0.7	0
📝	0.3	0.5	1
✂️	1	0	0.5

.2 −1

=

	🪨	📝	✂️
🪨	0	.4	−1
📝	−.4	0	1
✂️	1	−1	0

anti-symmetric

$2 \cdot \mathscr{P} - 1$ defines a symmetric 2p0s game!

Von Neumann / Minimax Winners

$$\boxed{\pi_1^\star}, \pi_2^\star = \arg\max_{\pi_1 \in \Pi} \arg\min_{\pi_2 \in \Pi} \mathbb{E}_{\xi_1 \sim \pi_1, \xi_2 \sim \pi_2}[2\mathcal{P}(\xi_1 \succ \xi_2) - 1]$$

↳ robust against the worst case comparator policy

$$\max_{\pi_1} (\pi')^T (2P-1) \pi_2^\star \leq \pi^T (2P-1) \pi = 0 \qquad \frac{0+1}{2} = 0.5$$

1. Pick a policy that is robust against worst case comparator.
2. Preferred to any other policy w.p. 1/2.
3. No assumptions on underlying shared $r^\star$ required!

Von Neumann / Minimax Winners

$$\pi_1^\star, \pi_2^\star = \arg\max_{\pi_1\in\Pi} \arg\min_{\pi_2\in\Pi} \mathbb{E}_{\xi_1\sim\pi_1,\xi_2\sim\pi_2}[2\mathscr{P}(\xi_1 \succ \xi_2) - 1]$$

	🪨	📝	✂️	
🪨	0.5	0.7	0	1.2
📝	0.3	0.5	1	
✂️	1	0	0.5	1.5

$$\pi^\star = \left[\frac{5}{12}, \boxed{\frac{5}{12}}, \frac{1}{6}\right]$$

If ⏱: Computing Minimax Winners

$$\pi_1^\star, \pi_2^\star = \arg\max_{\pi_1 \in \Pi} \arg\min_{\pi_2 \in \Pi} \mathbb{E}_{\xi_1 \sim \pi_1, \xi_2 \sim \pi_2}[2\mathscr{P}(\xi_1 \succ \xi_2) - 1]$$

We can define a sequence of losses for each NR player:

$$\ell_t^1(\pi) = \mathbb{E}_{\xi \sim \pi, \xi' \sim \pi_2^t}[2\mathscr{P}(\xi \succ \xi') - 1] \quad \ell_t^2(\pi) = \mathbb{E}_{\xi \sim \pi_t^1, \xi' \sim \pi}[-(2\mathscr{P}(\xi \succ \xi') - 1)]$$

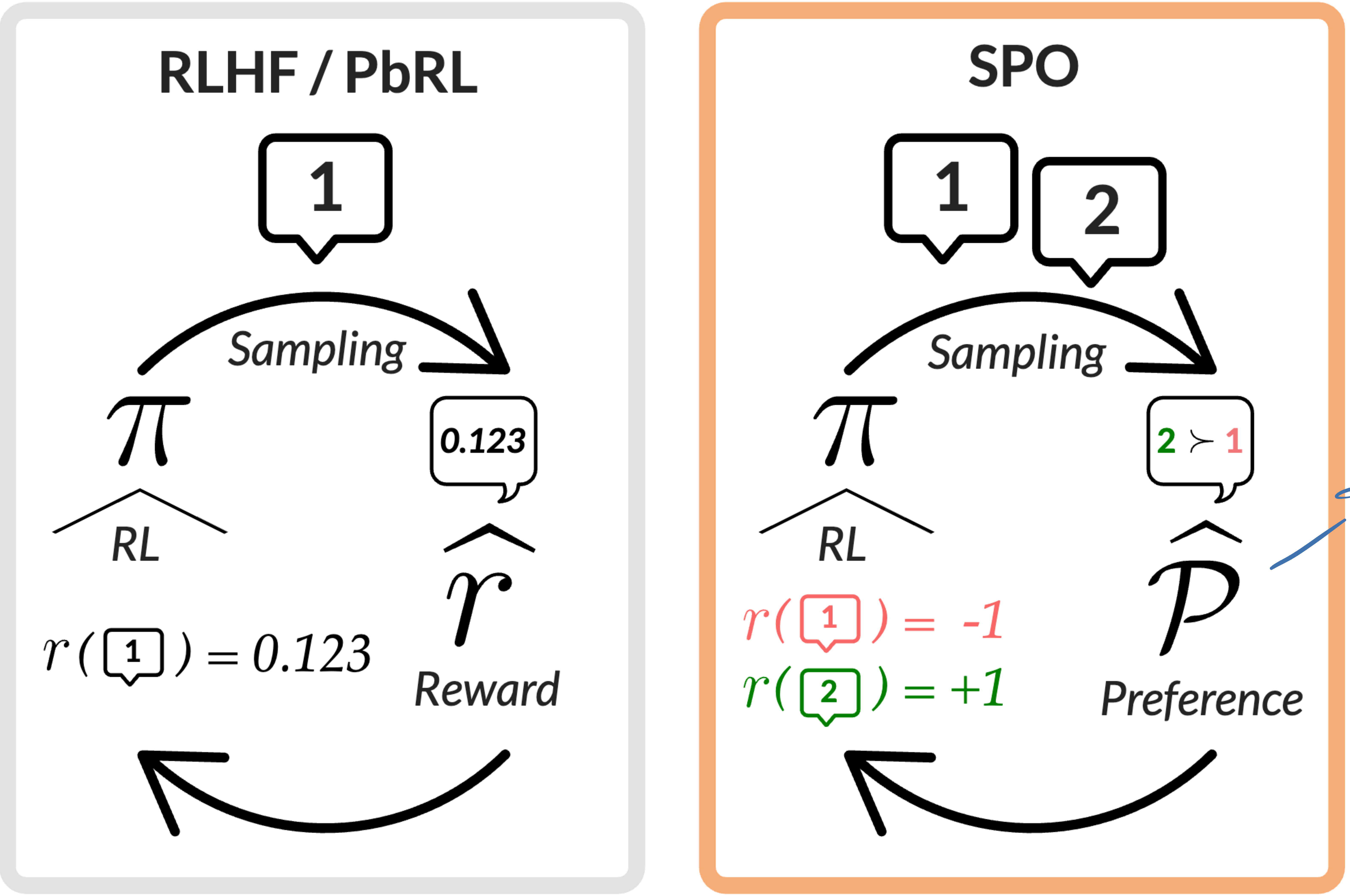
Without loss of generality, assume that $\pi_1^0 = \pi_2^0$. Then,

$$\begin{aligned}
\ell_0^1(\pi) &= \mathbb{E}_{\xi \sim \pi, \xi' \sim \pi_2^0}[2\mathscr{P}(\xi \succ \xi') - 1] \\
&= \mathbb{E}_{\xi \sim \pi, \xi' \sim \pi_1^0}[2\mathscr{P}(\xi \succ \xi') - 1] \\
&= \mathbb{E}_{\xi \sim \pi_1^0, \xi' \sim \pi}[-(2\mathscr{P}(\xi \succ \xi') - 1)] = \ell_0^2(\pi)
\end{aligned}$$

$$\Rightarrow \forall t \in [T], \pi_1^t = \pi_2^t$$

Can just do self-play, no adversarial training required!

SPO: Self-Play Preference Optimization



$r_t = P \circ \pi_t$

$\sim N(\xi_1, \xi_2)$

"win rate against π as the RM"

Instantiations of SPO on Large Language Models

Offline Dataset

P(r > c)	SFT	RLHF	SPO
SFT	0.5	0.02	0.02
RLHF	0.98	0.5	0.25
SPO	0.98	0.75	0.5

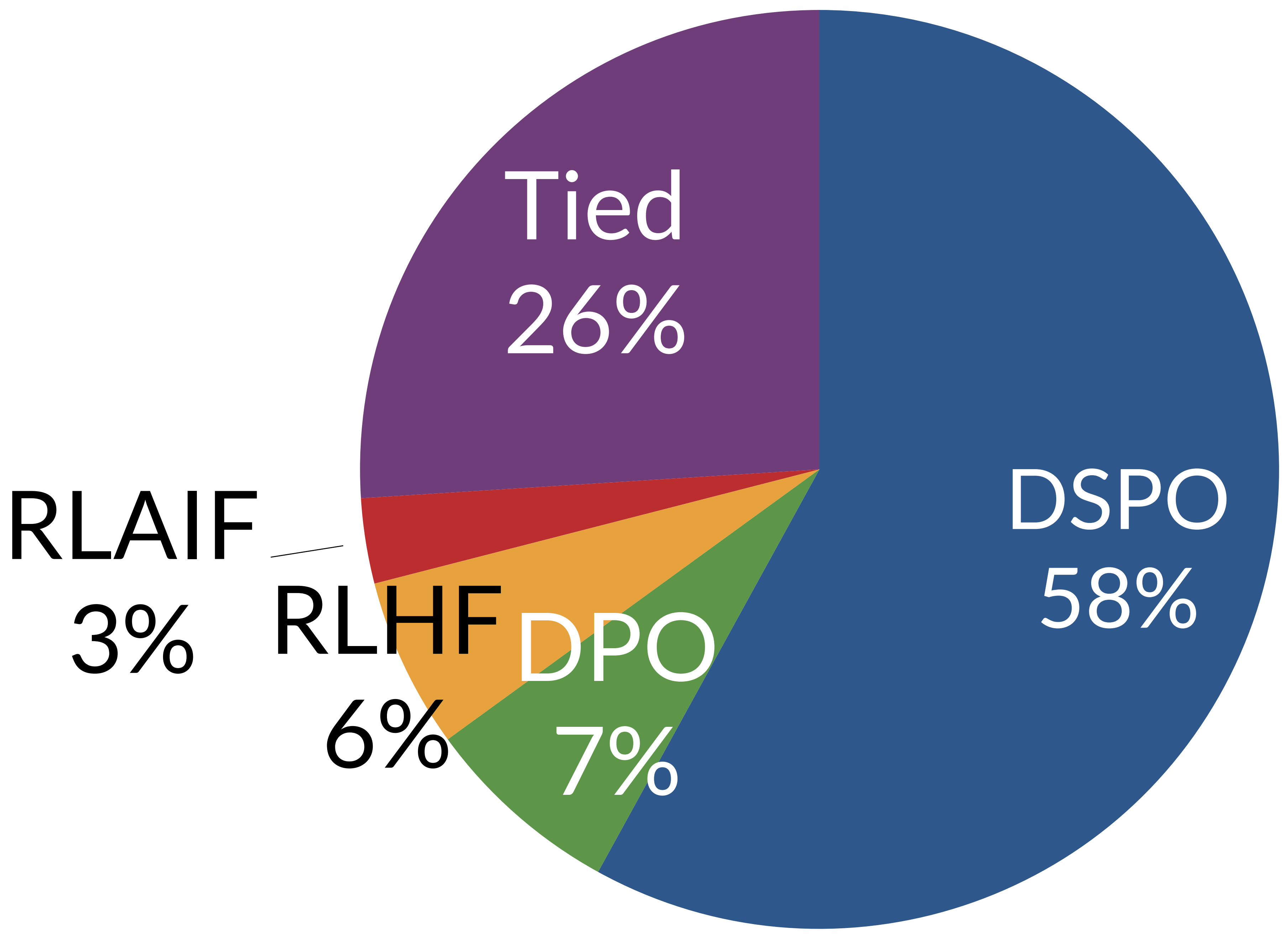
[Munos+'23]

P(r > c)	RLHF	ISPO	DSPO
RLHF	0.5	0.21	0.24
ISPO	0.79	0.5	0.61
DSPO	0.76	0.39	0.5

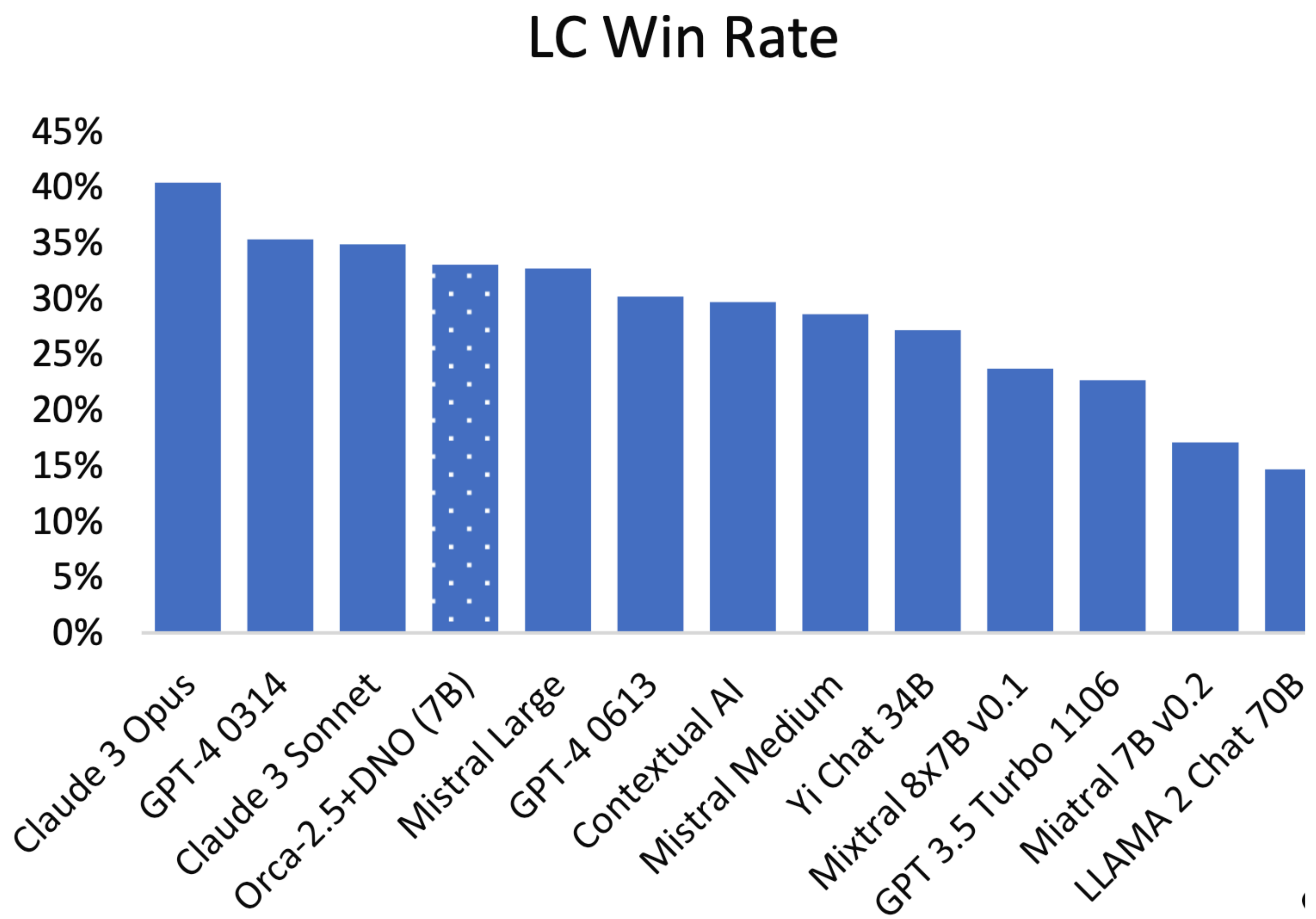
[Calandriello+'24]

Instantiations of SPO on Large Language Models

Online Oracle



[Guo+'24]



[Rosset+'24]

Outline for Today

1. When is the Bradley-Terry assumption inaccurate and what happens to online / offline PFT as a result?

***A**: BT is violated when when a reward function can't explain (aggregate) preferences, leading to mode collapse in RLHF.*

2. What is a more robust criterion for preference aggregation and how can we efficiently optimize it?

***A**: The minimax winner doesn't assume transitivity of preferences. We can use a self-play algorithm to compute it.*