The Value of RL in Fine-Tuning

Gokul Swamy

Outline for Today

1. What assumption on the preference dataset did we make in the DPO derivation and what happens when it breaks?

2. When are two-stage RLHF and DPO equivalent?

3. Why does two-stage RLHF work *much* better in practice?

Recap: Forward and Reverse KL

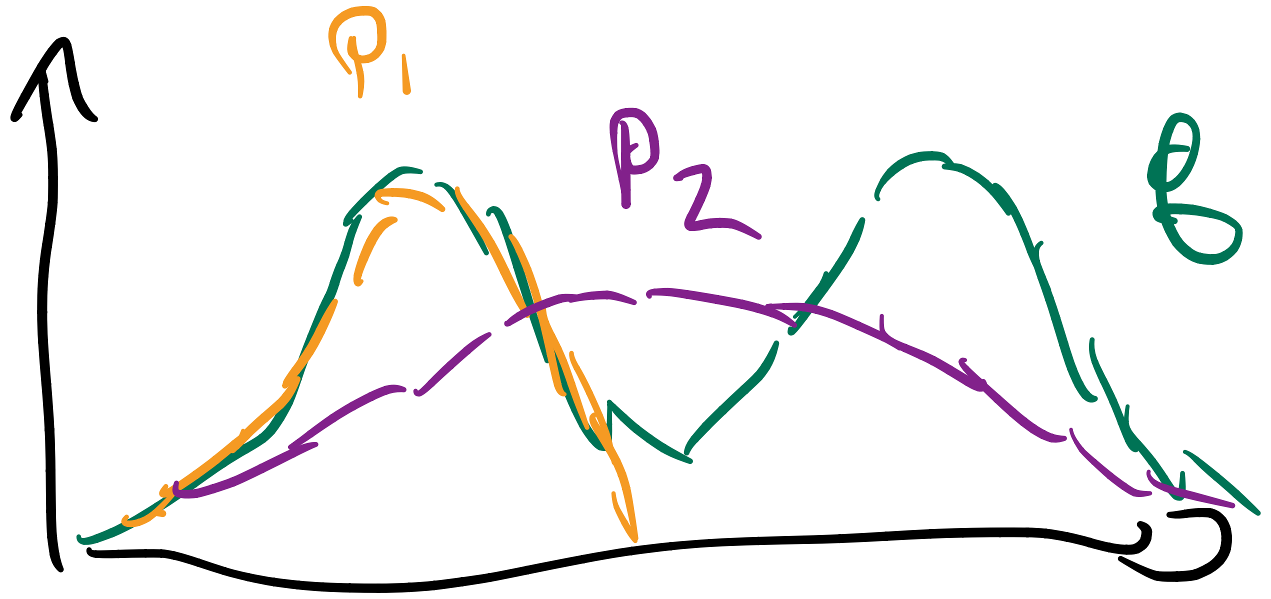
~ mode covering ~

FKL: $\min_{p} \mathbb{D}_{\mathrm{KL}}(q\,||\,p) = \min_{p} \sum_{x} q(x) \log \left(\frac{q(x)}{p(x)} \right)$ MLE

RKL: $\min_{p} \mathbb{D}_{\mathrm{KL}}(p\,||\,q) = \min_{p} \sum_{x} p(x) \log \left(\frac{p(x)}{q(x)} \right)$ Soft RL

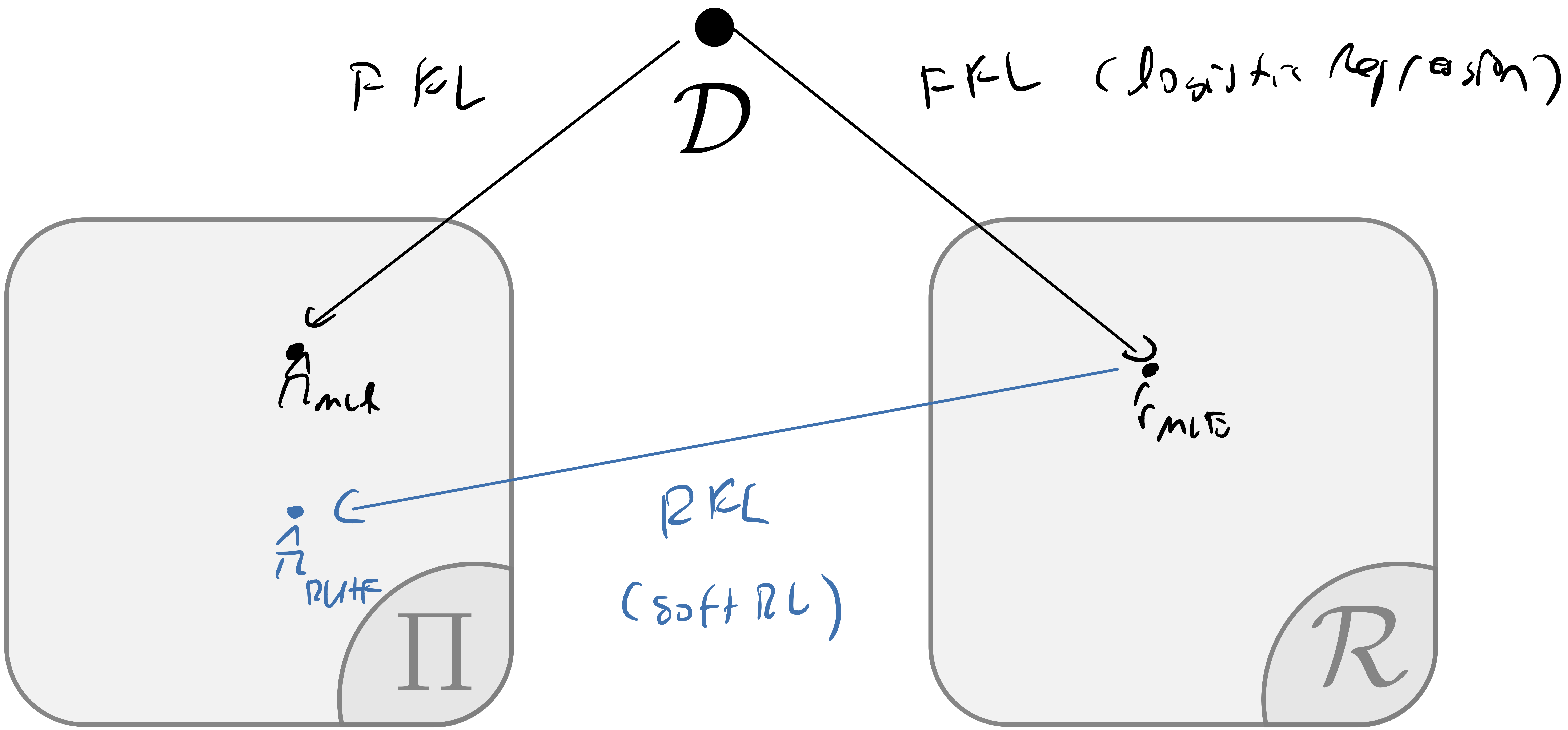
~ mode-seeking ~

~ mode-seeking

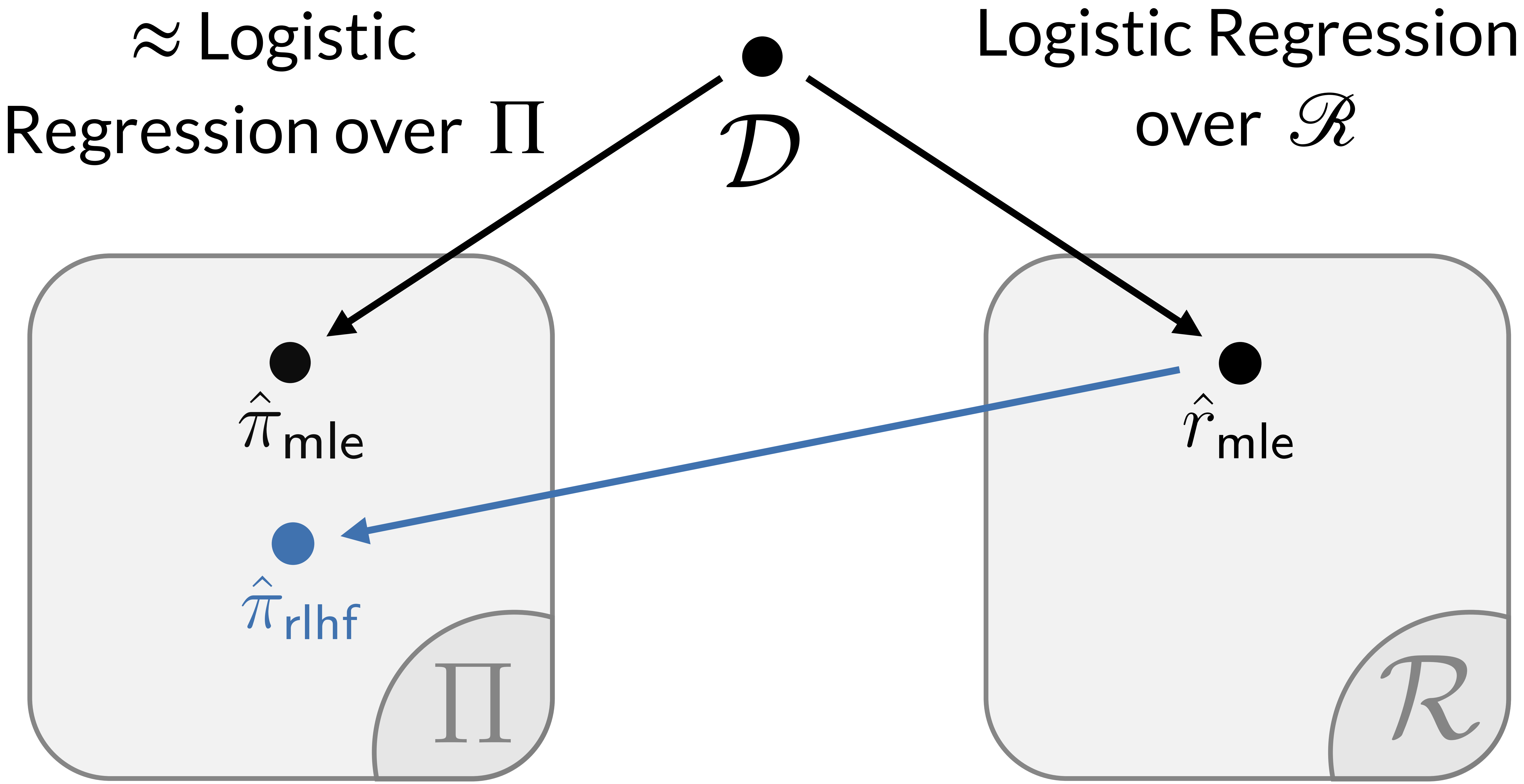


FKL is “mode-covering” while RKL is “mode-seeking”

Recap: Information Geometry of RLHF



Recap: Information Geometry of RLHF



$$\hat{\pi}_{\text{rlhf}} = \arg\max_{\pi \in \Pi} \mathbb{E}_{\xi \sim \pi}[\hat{r}_{\text{mle}}(\xi)] + \mathbb{D}_{KL}(\pi \,||\, \pi_{\text{ref}})$$

Outline for Today

$\hookrightarrow \mathbb{E}_{\xi \sim \mu}$

1. What assumption on the preference dataset did we make in the DPO derivation and what happens when it breaks?

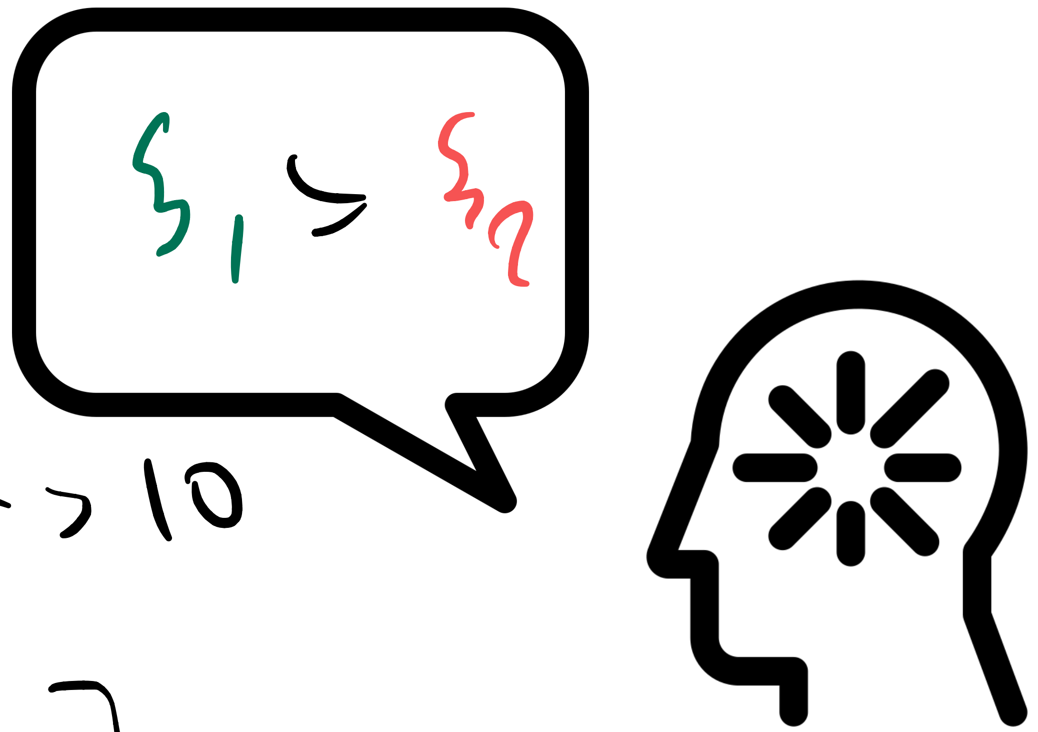
***A**: Full coverage of $\mathscr{D}$. Without it, we can't control the RKL.*

2. When are two-stage RLHF and DPO equivalent?

3. Why does two-stage RLHF work *much* better in practice?

Why does DPO break with Partial Coverage?

	ξ_1	ξ_2	ξ_3
π_{ref}	0.5	0.5	0
$\hat{\pi}_{\text{rlhf}}$	1	0	0
$\hat{\pi}_{\text{dpo}}$	~~1~~ .1	0 0	0 .9



$r > 10$

$[1, 0, ?]$

$$\underset{\pi\in\Pi}{\arg\max}\ \mathbb{E}_{\xi\sim\pi}[\overbrace{\hat{r}_{\text{mle}}(\xi)}^{r}] + \mathbb{D}_{KL}(\pi\,||\,\overbrace{\pi_{\text{ref}}}^{0})$$

$$\underset{\pi\in\Pi}{\arg\max}\ \mathbb{E}_{\mathscr{D}}\left[\log\sigma\left(\sum_{h}^{H}\log\frac{\pi(a_h^+\,|\,s_h^+)}{\cancel{\pi_{\text{ref}}(a_h^+\,|\,s_h^+)}} - \log\frac{\pi(a_h^-\,|\,s_h^-)}{\cancel{\pi_{\text{ref}}(a_h^-\,|\,s_h^-)}}\right)\right]$$

$\hookrightarrow \xi_1 \succ \xi_2$

DPO Doesn't Regularize to π_{ref} and can produce OOD responses.

Outline for Today

1. What assumption on the preference dataset did we make in the DPO derivation and what happens when it breaks?

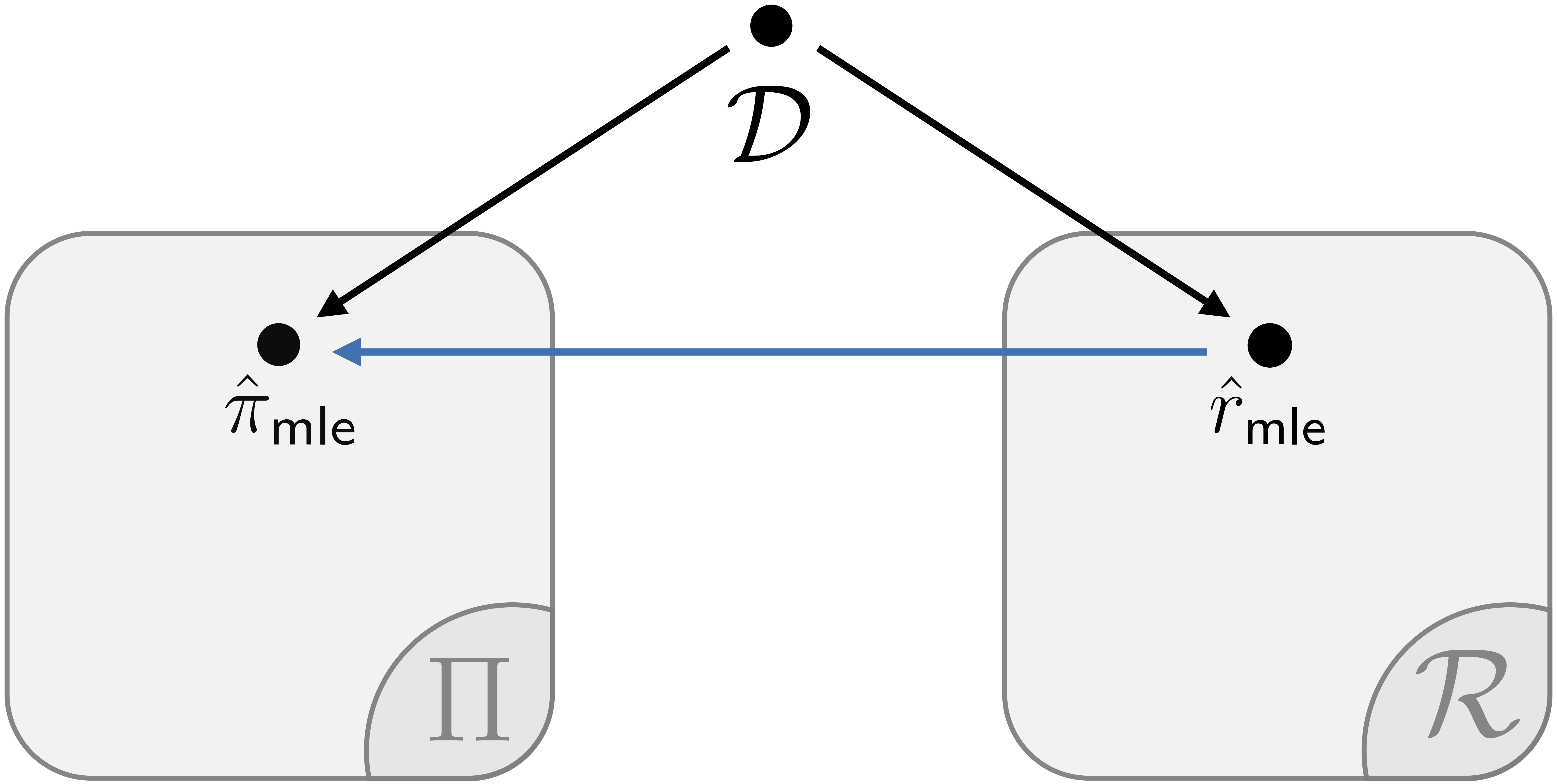
***A**: Full coverage of $\mathscr{D}$. Without it, we can't control the RKL.*

2. When are two-stage RLHF and DPO equivalent?

***A**: When Π and $\mathscr{R}$ are isomorphic and all projections are exact.*

3. Why does two-stage RLHF work *much* better in practice?

When are two-stage RLHF and DPO equivalent?



A: *If (1) $\Pi \Leftrightarrow \mathscr{R}$ and (2) all projections exact.*

If ⏱️: RLHF = DPO when $\Pi \Leftrightarrow \mathscr{R}$

($\hat{\pi}_{\text{mle}}$

(minimizing same fn over the same class)

$$\mathbb{E}_{\mathscr{D}}\left[\log\sigma\left(\sum_h^H \log\hat{\pi}_{\text{mle}}(a_h^+ \mid s_h^+) - \log\hat{\pi}_{\text{mle}}(a_h^- \mid s_h^-)\right)\right] = \mathbb{E}_{\mathscr{D}}\left[\log\sigma\left(\hat{r}_{\text{mle}}(\xi^+) - \hat{r}_{\text{mle}}(\xi^-)\right)\right]$$

(uniqueness of minimizers)

$$\Rightarrow \forall \xi \in \Xi, \sum_h^H \log\hat{\pi}_{\text{mle}}(a_h \mid s_h) = \hat{r}_{\text{mle}}(\xi)$$

(soft RL = RLHF)

$$\hat{\pi}_{\text{rlhf}} = \arg\min_{\pi\in\Pi} \mathbb{D}_{KL}\left(\mathbb{P}_\pi \,||\, \mathbb{P}^\star_{\hat{r}}\right)$$

$$= \arg\min_{\pi\in\Pi} \mathbb{D}_{KL}\left(\mathbb{P}_\pi \,||\, \mathbb{P}_{r_{\hat{\pi}}}\right)$$

$$= \hat{\pi}_{\text{mle}}$$

the same function

$\hat{\pi}_{\text{mle}} \in \Pi$

NPG covariant

MLE is invariant to reparameterization.

Policies vs. Reward Models

Policies: $\pi : \mathscr{S} \rightarrow \Delta(\mathscr{A}) \in \Pi$

(Prefixes) *(Tokens)*

Rewards: $r : \Xi \rightarrow \mathbb{R} \in \mathscr{R}$

(Completions)

Both of these are fine-tuned from the same SFT checkpoint!

Outline for Today

1. What assumption on the preference dataset did we make in the DPO derivation and what happens when it breaks?

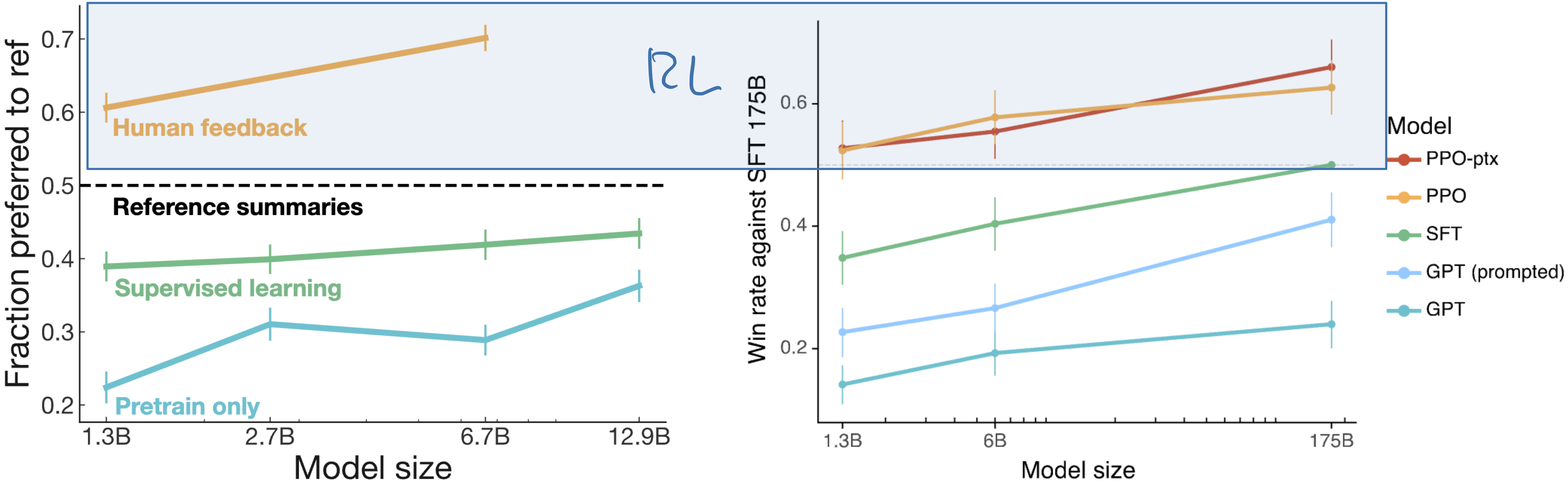
A: *Full coverage of $\mathscr{D}$. Without it, we can't control the RKL.*

2. When are two-stage RLHF and DPO equivalent?

A: *When Π and $\mathscr{R}$ are isomorphic and all projections are exact.*

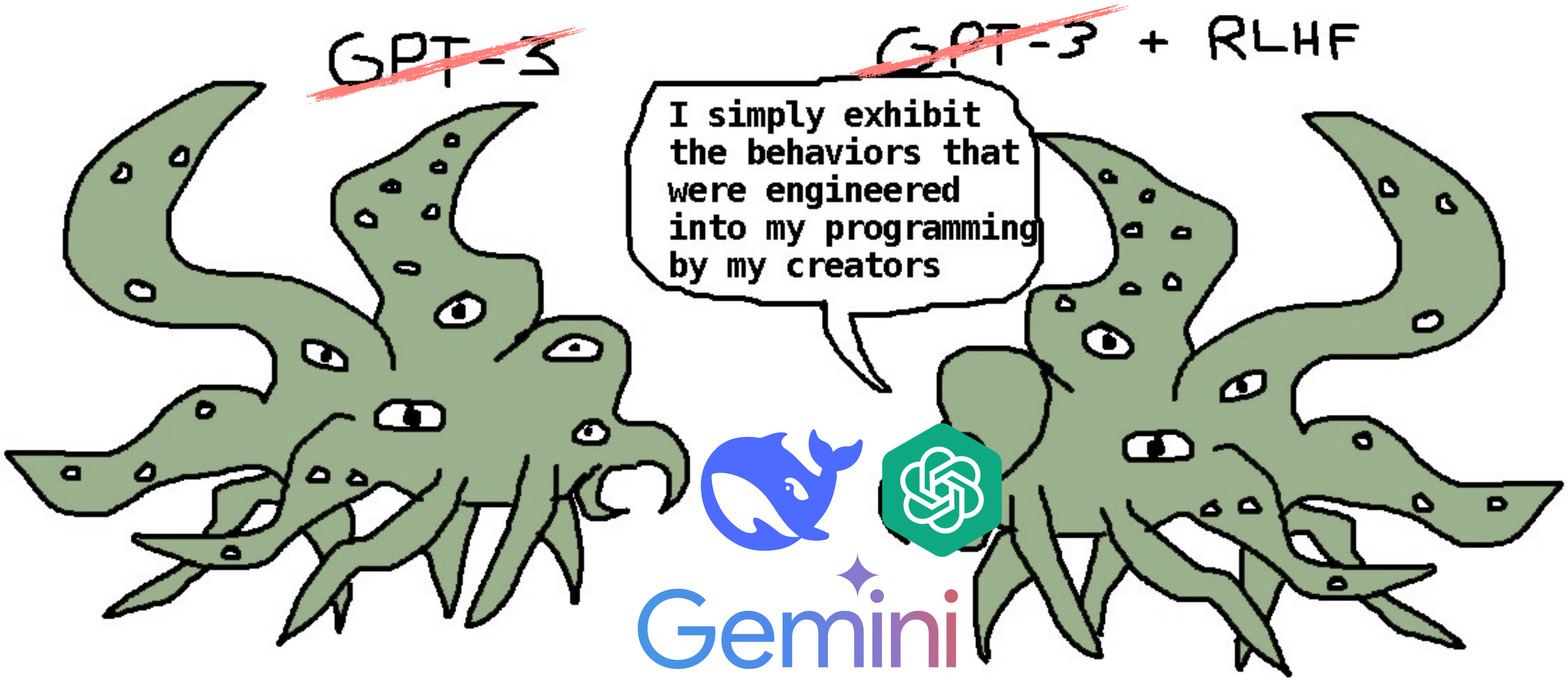
3. Why does two-stage RLHF work *much* better in practice?

Two Stage RLHF > DPO



[Stiennon et al., Ouyang et al.]

Two Stage RLHF > DPO



Base

FT

Grounding the Gap in Summarization

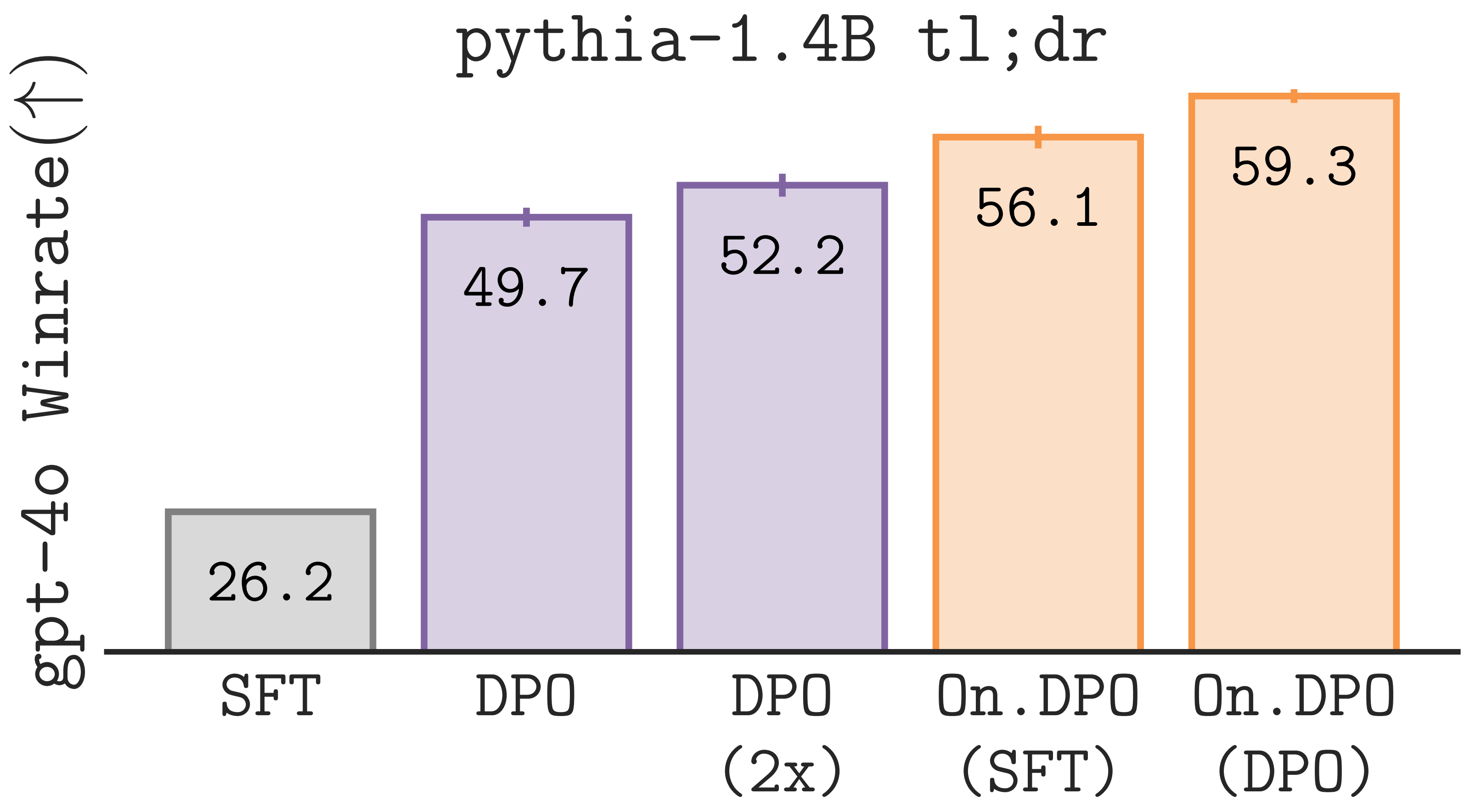
1. We will focus on the task of summarization of Reddit posts, using models from the Pythia family pre-trained on the Pile.

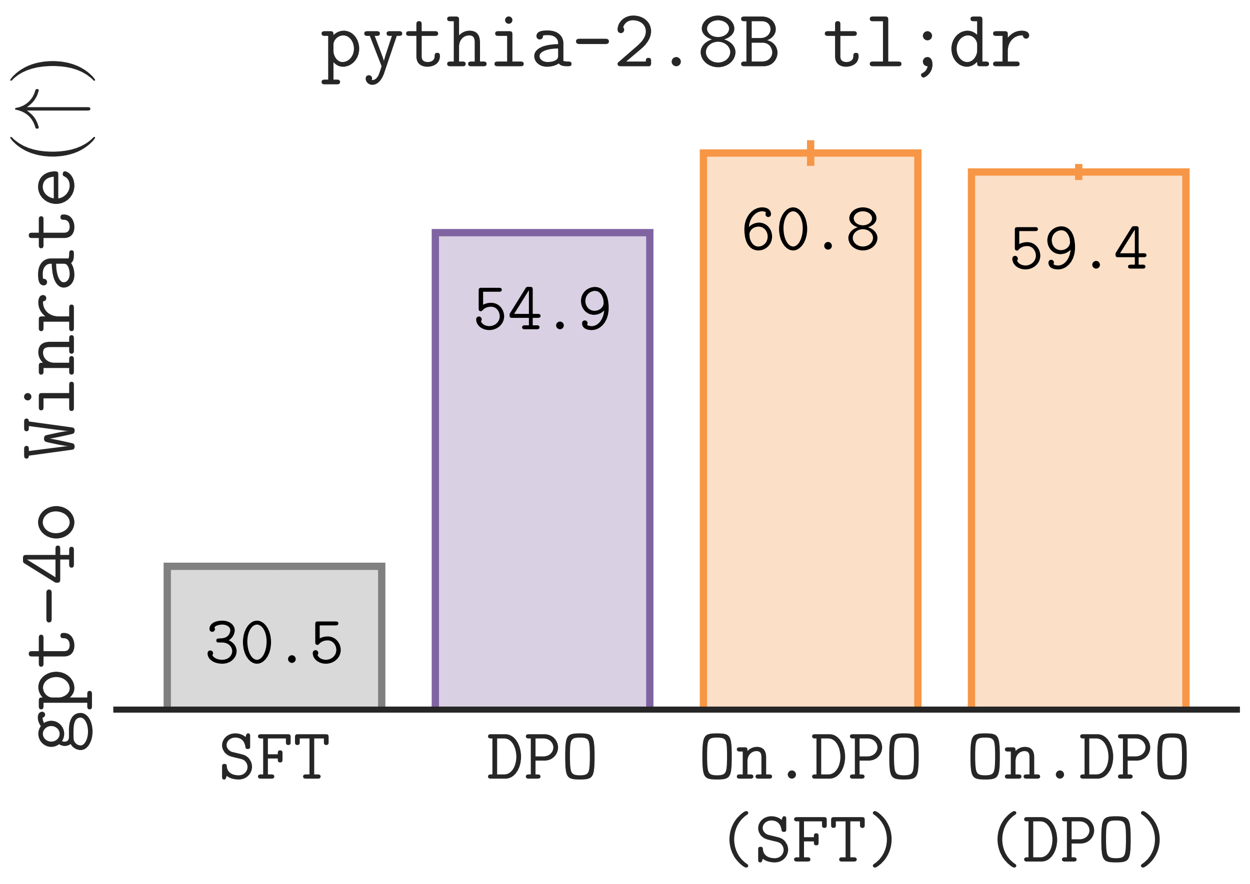
2. We will use the ***same*** dataset to train both policies and reward models.

3. We will start from the ***same*** SFT checkpoint to train both.

4. We will use the ***same*** optimizer (DPO) for both online and offline PFT with the ***same*** hyperparameters.

Gap Appears in "🍎s to 🍎s" Comparison





6 Hypotheses for the Online-Offline Gap

$\mathbb{H}_1$

$\mathbb{H}_2$

$\mathbb{H}_3$

$\mathbb{H}_4$

$\mathbb{H}_5$

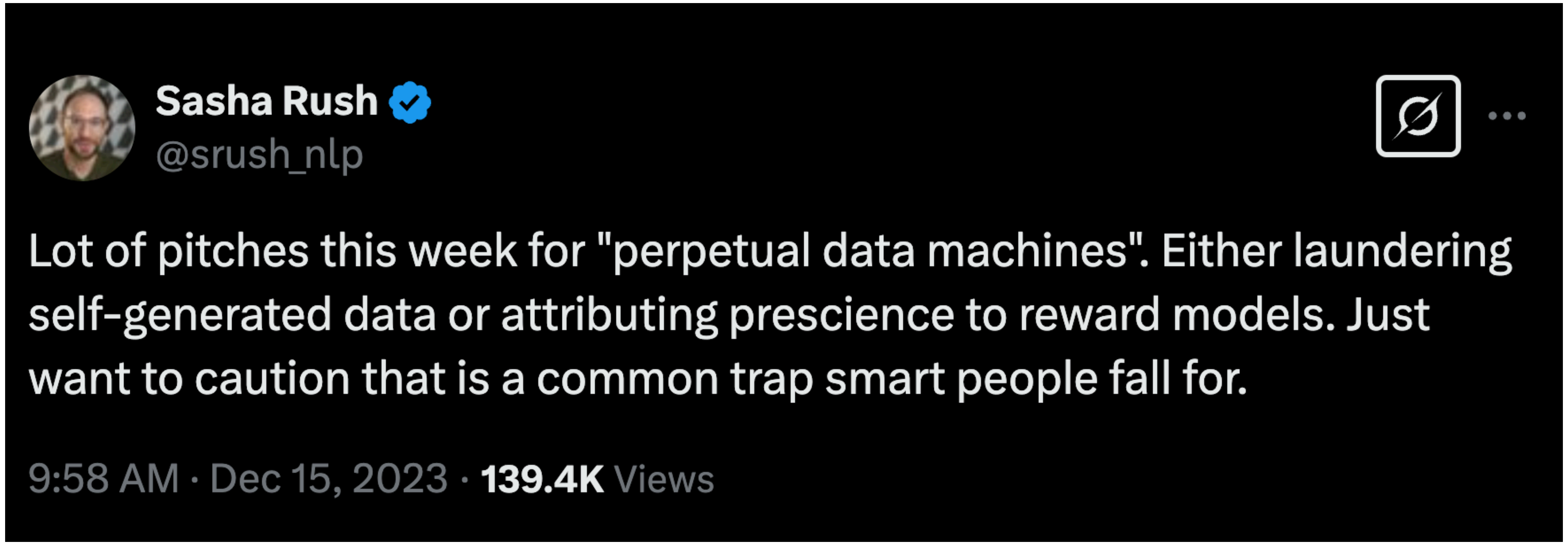
$\mathbb{H}_6$

$\mathbb{H}_1$: Intrinsic Value of On-Policy Feedback

... but the on-policy labels are just *imputed*

... from an RM trained on the same data as the policy

... and we can't create any new info via sampling.


Sasha Rush
@srush_nlp

Lot of pitches this week for "perpetual data machines". Either laundering self-generated data or attributing prescience to reward models. Just want to caution that is a common trap smart people fall for.

9:58 AM · Dec 15, 2023 · **139.4K** Views

$\mathbb{H}_2$: Failure of Offline Regularization to π_{ref}

$\mathbb{E}_{\xi \sim \pi}$

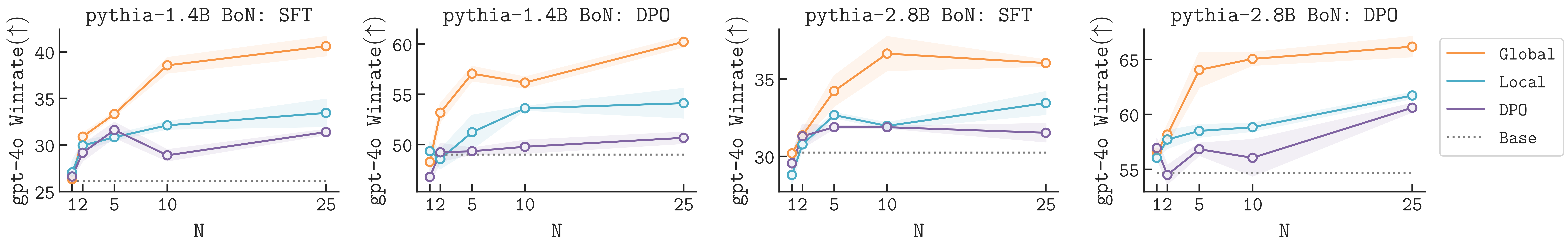
$$\pi^\star = \arg\min_{\pi \in \Pi} \mathbb{D}_{KL}\left(\mathscr{D}\,||\,\pi\right) \quad + \quad \mathbb{D}_{KL}\left(\pi\,||\,\pi_{\text{ref}}\right)$$

Reverse KL has an on-policy expectation.

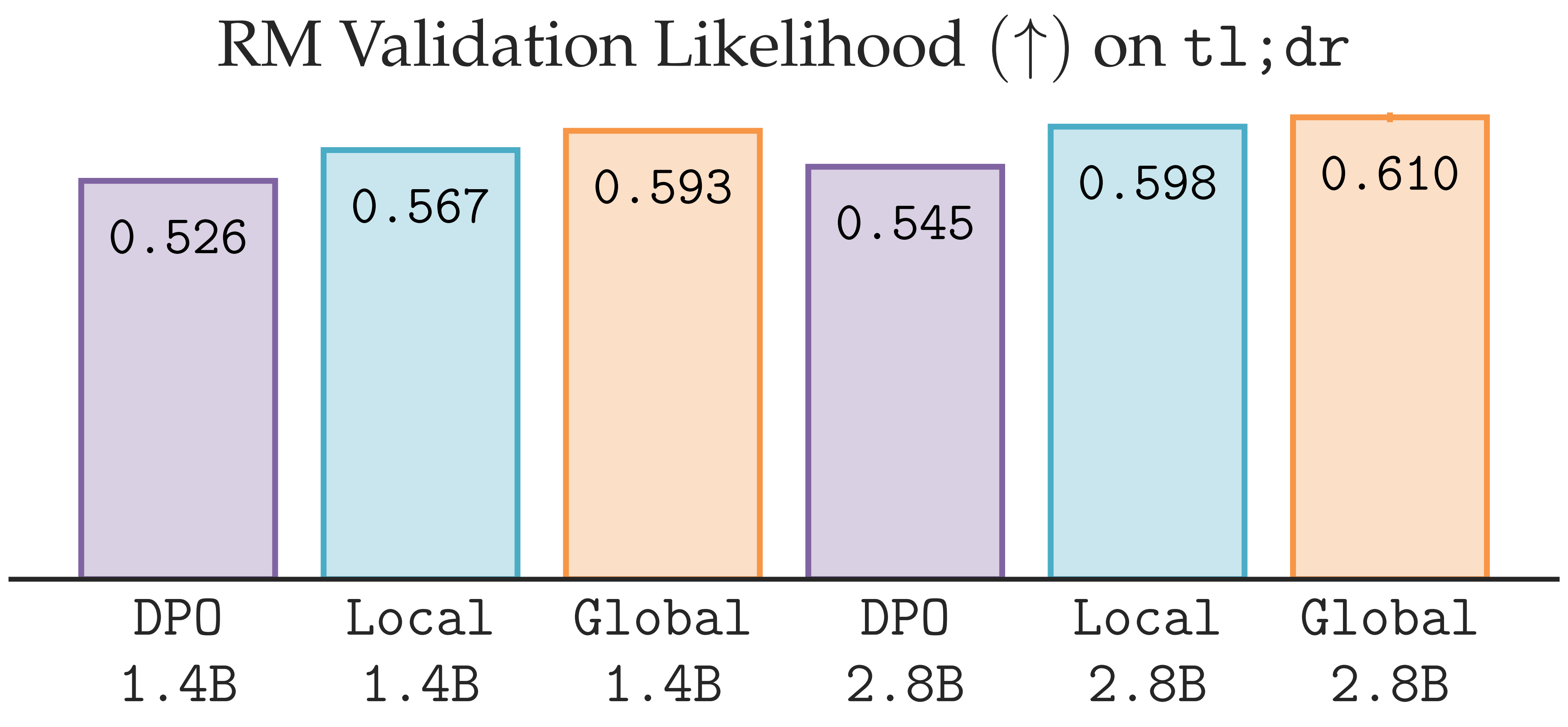
... but we used the same regularizer for all experiments.

[Song et al.]

$\mathbb{H}_5$: RMs Generalize Better OOD



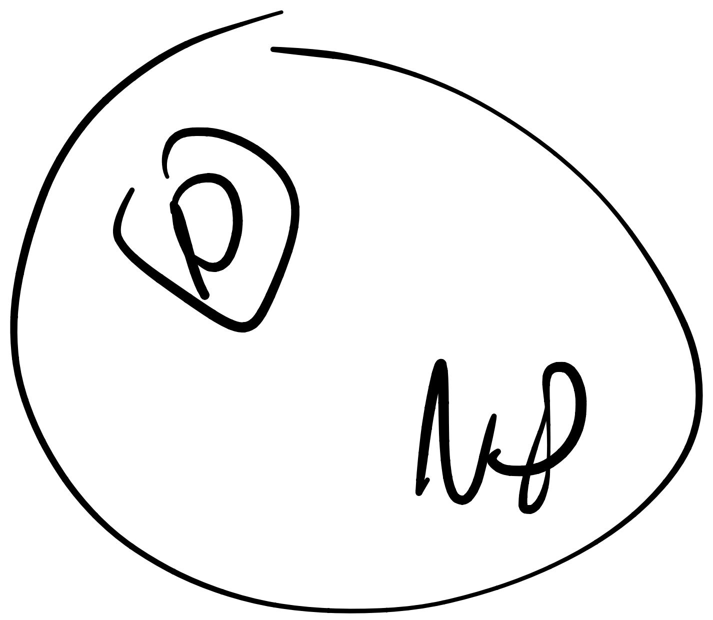
... but they also generalize better ID



Generation-Verification Gaps

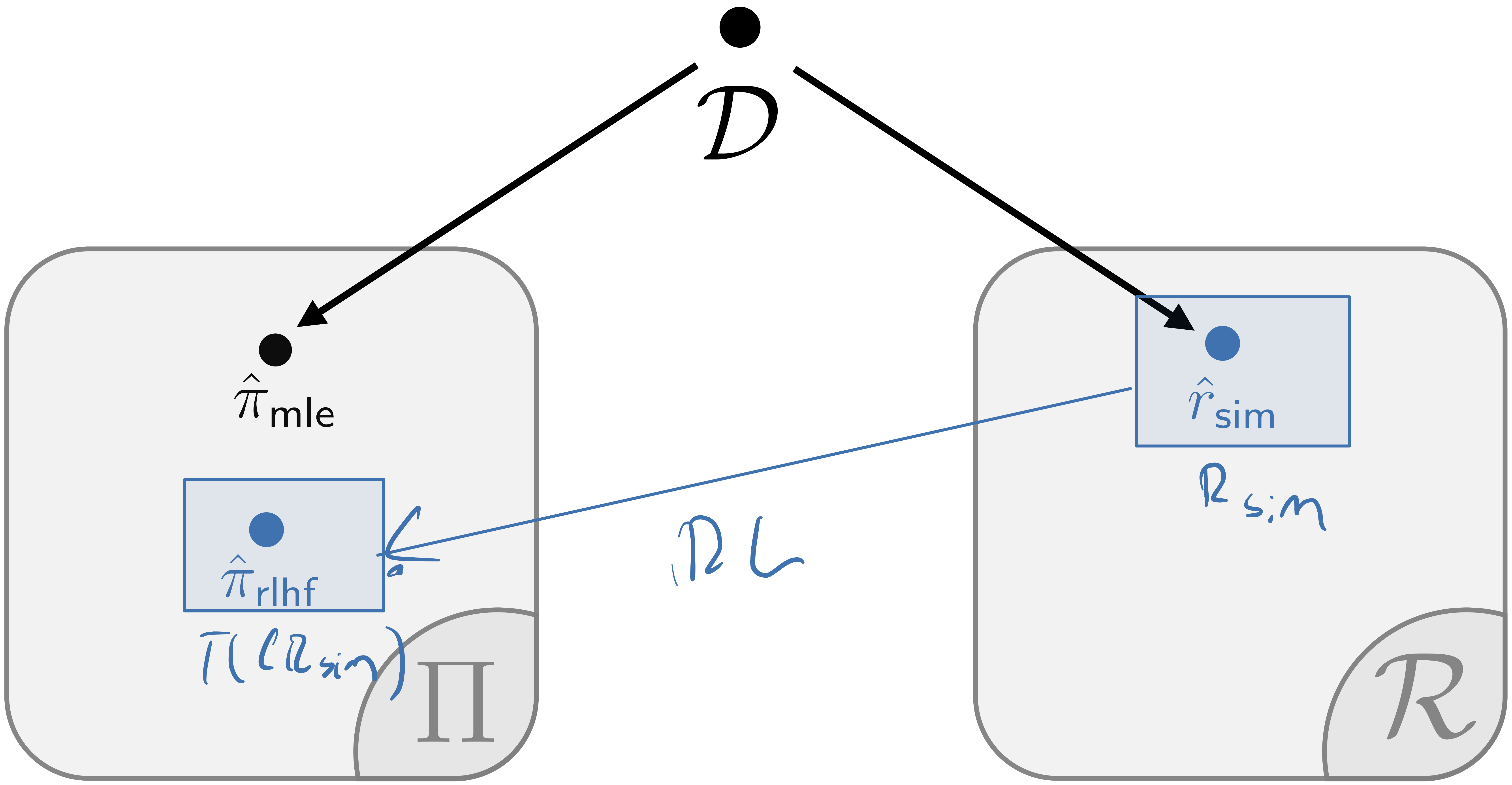
$P \mathrel{!}= NP$

4	6	2	3	1	5	7	8	9
8	3	9	7	4	2	1	5	6
5	7	1	6	8	9	4	3	2
3	1	8	4	2	7	6	9	5
6	4	5	1	9	8	2	7	3
9	2	7	5	6	3	8	1	4
2	9	3	8	7	4	5	6	1
7	5	6	2	3	1	9	4	8
1	8	4	9	5	6	3	2	7



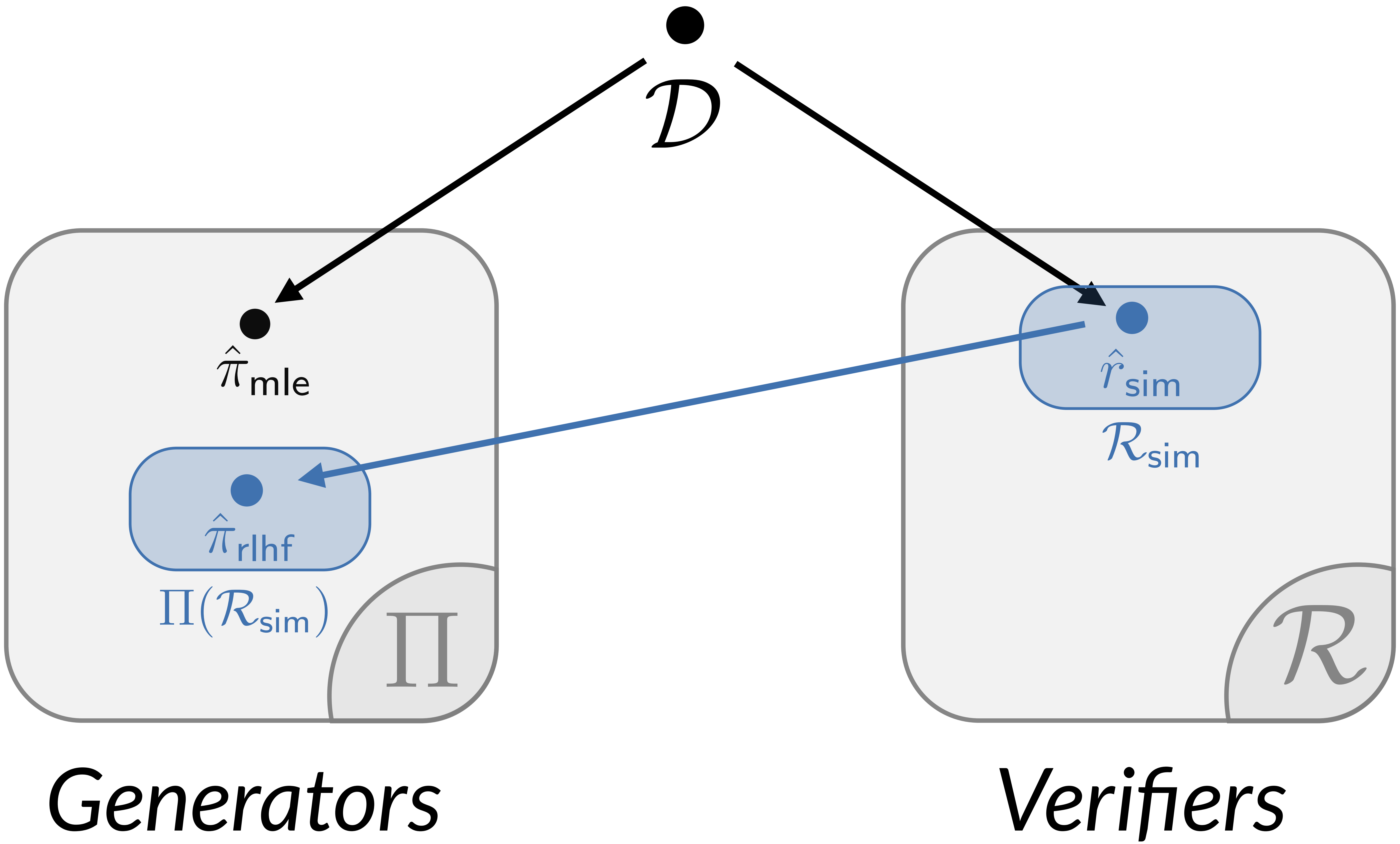
GV Gap = easier to check than to solve!

$\mathbb{H}_6$: *Proper* Learning w/ a *Generation-Verification Gap*



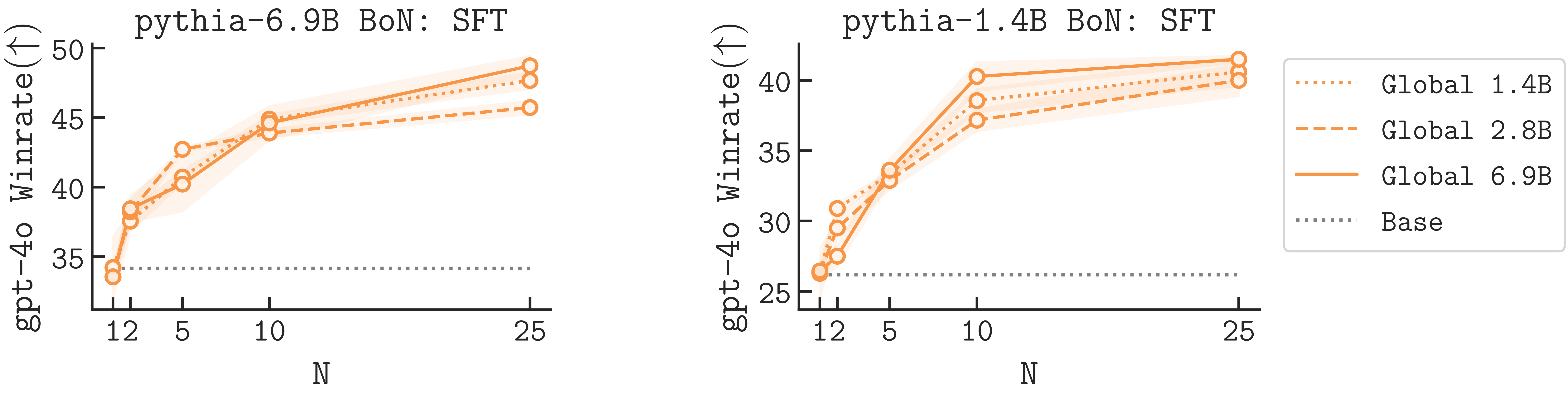
Only need to search over $\Pi\left(\mathscr{R}_{\mathit{sim}}\right) \subset \Pi$*!*

$\mathbb{H}_6$: *Proper* Learning w/ a *Generation-Verification Gap*



All roads lead to likelihood, but RL takes a shortcut!

Evidence for Generation-Verification Gap

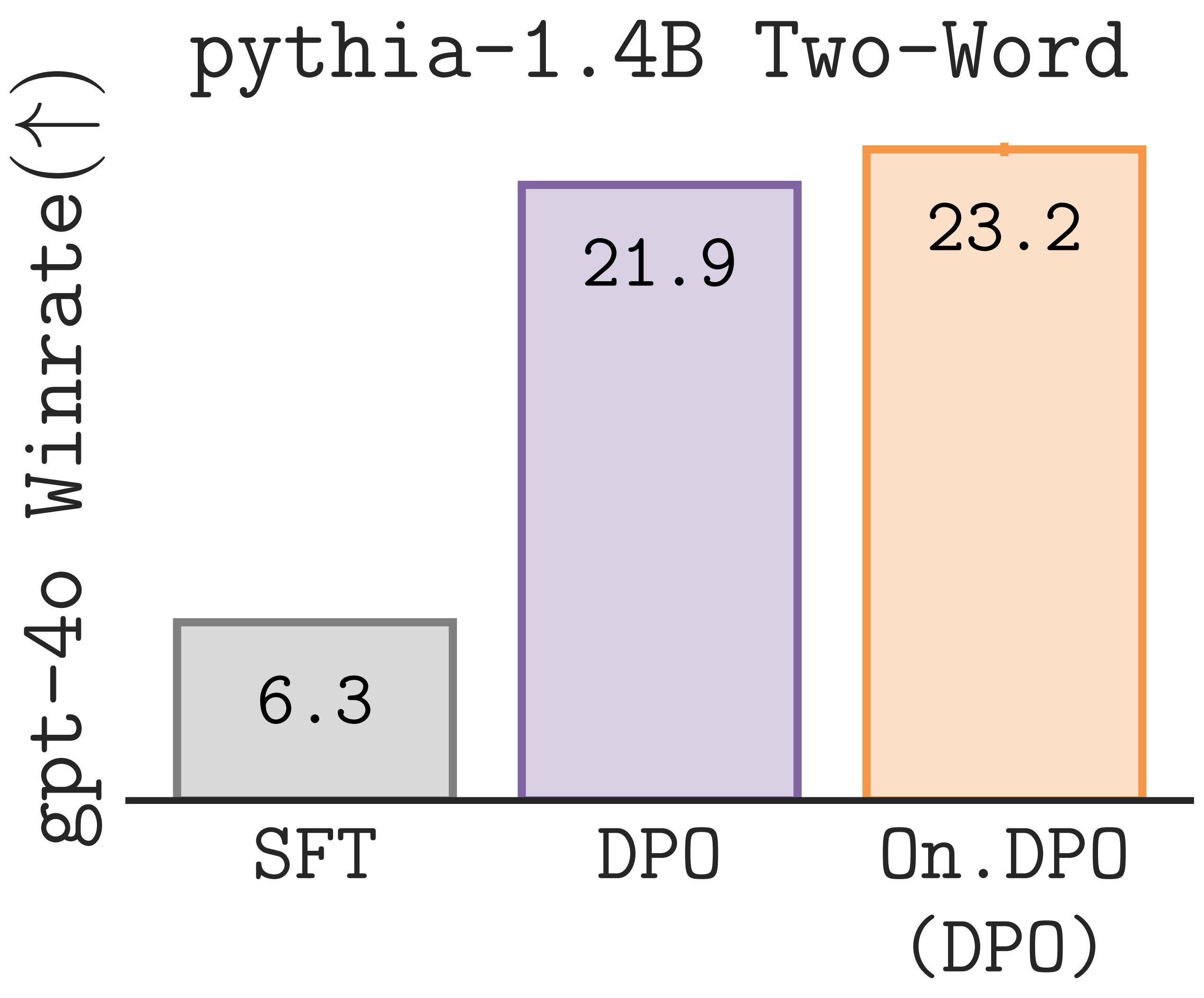


Using a much smaller RM than policy doesn't hurt.

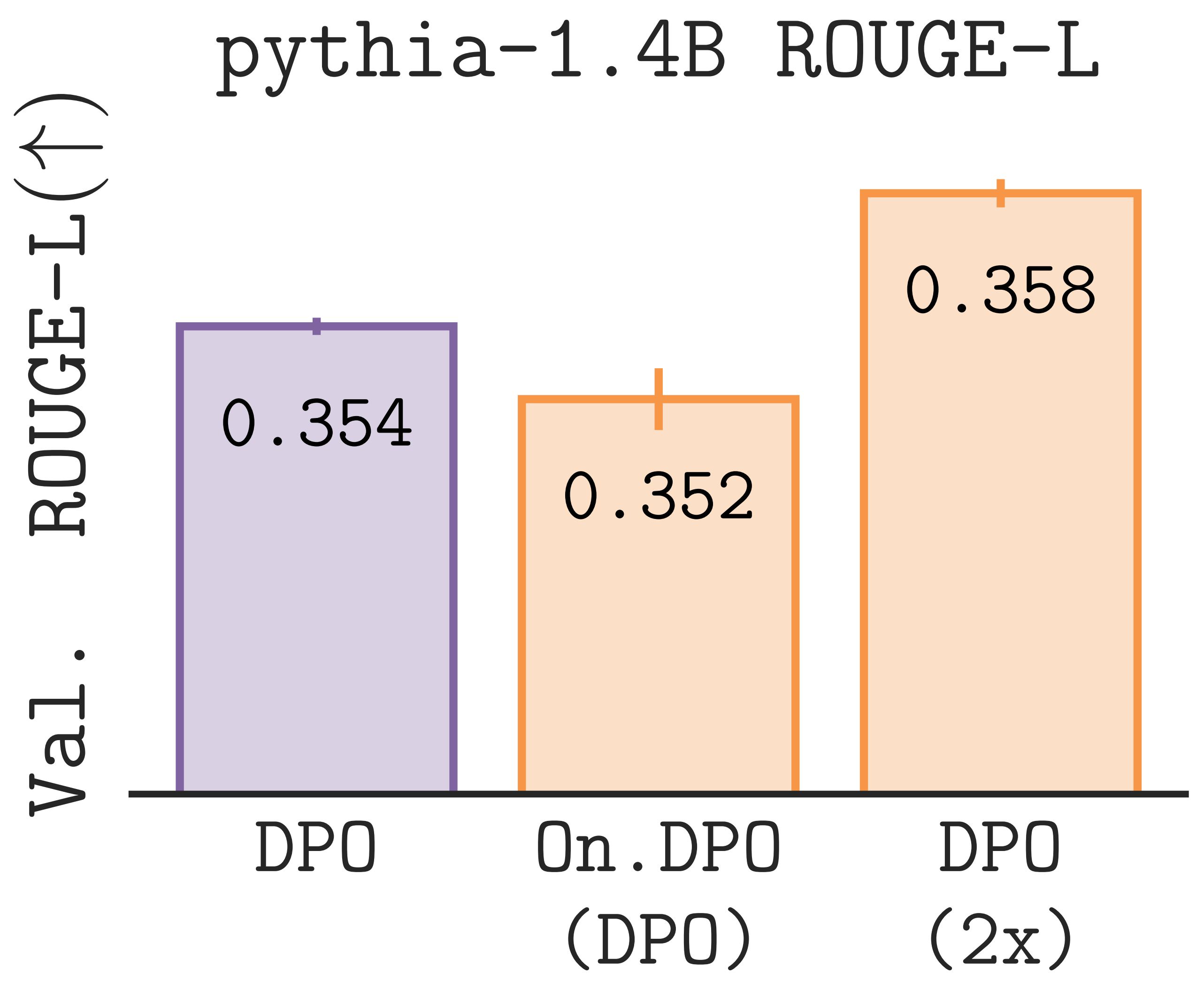
Using a much larger RM than policy doesn't help.

Closing the Generation-Verification Gap

Simplify Policy:



Complicate Reward:



Online PFT ≈ Offline PFT with no generation-verification gap!

Outline for Today

1. What assumption on the preference dataset did we make in the DPO derivation and what happens when it breaks?

***A**: Full coverage of $\mathscr{D}$. Without it, we can't control the RKL.*

2. When are two-stage RLHF and DPO equivalent?

***A**: When Π and $\mathscr{R}$ are isomorphic and all projections are exact.*

3. Why does two-stage RLHF work *much* better in practice?

***A**: RLHF only has to search over policies (generators) that are optimal for simple rewards (verifiers) rather than over all of Π.*