Online vs Offline RLHF

Konwoo Kim, REFORM 2.5

Papers

1. The Importance of Online Data: Understanding Preference Fine-tuning via Coverage
2. Understanding the performance gap between online and offline alignment algorithms

Overview

1. Online vs offline RLHF
2. Coverage conditions
3. Empirical experiments

1. Online vs Offline RLHF

Online vs Offline RLHF

$$\widehat{r} \in \operatorname*{argmax}_{r \in \mathcal{R}} \widehat{\mathbb{E}}_{x, y^+, y^- \sim \mathcal{D}} \left[\log \left(\frac{\exp(r(x, y^+))}{\exp(r(x, y^+)) + \exp(r(x, y^-))} \right) \right]$$

$$\pi_{\mathsf{rlhf}} \in \operatorname*{argmax}_{\pi} \widehat{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathbb{E}_{y \sim \pi(\cdot|x)} [\widehat{r}(x, y)] - \beta \mathsf{KL}(\pi(\cdot \mid x) || \pi_{\mathsf{ref}}(\cdot \mid x)) \right]$$

Online vs Offline RLHF

$$\widehat{r} \in \operatorname*{argmax}_{r \in \mathcal{R}} \widehat{\mathbb{E}}_{x,y^+,y^- \sim \mathcal{D}} \left[\log \left(\frac{\exp(r(x,y^+))}{\exp(r(x,y^+)) + \exp(r(x,y^-))} \right) \right]$$

$$\pi_{\mathsf{rlhf}} \in \operatorname*{argmax}_{\pi} \widehat{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathbb{E}_{y \sim \pi(\cdot|x)} [\widehat{r}(x,y)] - \beta \mathsf{KL}(\pi(\cdot \mid x) || \pi_{\mathsf{ref}}(\cdot \mid x)) \right]$$

$$\ell_{\mathsf{dpo}}(\pi) = \widehat{\mathbb{E}}_{x,y^+,y^- \sim \mathcal{D}} \left[\log \left(\frac{\exp\left(\beta \log\left(\frac{\pi(y^+|x)}{\pi_{\mathsf{ref}}(y^+|x)}\right)\right)}{\exp\left(\beta \log\left(\frac{\pi(y^+|x)}{\pi_{\mathsf{ref}}(y^+|x)}\right)\right) + \exp\left(\beta \log\left(\frac{\pi(y^-|x)}{\pi_{\mathsf{ref}}(y^-|x)}\right)\right)} \right) \right]$$

2. Coverage Conditions

Global and Local Coverage

Assumption 4.1 (Global Coverage). *For all* π, *we have*

$$\max_{x,y:\rho(x)>0} \frac{\pi(y \mid x)}{\pi_{\mathsf{ref}}(y \mid x)} \leq C_{\mathsf{glo}}.$$

Global and Local Coverage

Assumption 4.1 (Global Coverage). *For all π, we have*

$$\max_{x,y:\rho(x)>0} \frac{\pi(y \mid x)}{\pi_{\mathsf{ref}}(y \mid x)} \leq C_{\mathsf{glo}}.$$

Assumption 4.2 (Local KL-ball Coverage). *For all $\varepsilon_{\mathsf{kl}} < \infty$ and all policy π such that $\mathbb{E}_{x\sim\rho}[\mathsf{KL}(\pi(\cdot \mid x)||\pi_{\mathsf{ref}}(\cdot \mid x))] \leq \varepsilon_{\mathsf{kl}}$, we have*

$$\max_{x,y:\rho(x)>0} \frac{\pi(y \mid x)}{\pi_{\mathsf{ref}}(y \mid x)} \leq C_{\varepsilon_{\mathsf{kl}}}.$$

Global Coverage is Necessary for DPO

Assumption 4.1 (Global Coverage). *For all* π, *we have*

$$\max_{x,y:\rho(x)>0} \frac{\pi(y \mid x)}{\pi_{\mathsf{ref}}(y \mid x)} \leq C_{\mathsf{glo}}.$$

Assumption 4.3 (In Distribution Reward Learning). *We assume the DPO policy* π_{dpo} *satisfies that:*

$$\mathbb{E}_{x,y\sim\rho\circ\pi_{\mathsf{ref}}}\left[\left(\beta\log\left(\frac{\pi_{\mathsf{dpo}}(y \mid x)}{\pi_{\mathsf{ref}}(y \mid x)Z(x)}\right) - r^*(x,y)\right)^2\right] \leq \varepsilon_{\mathsf{dpo}}.$$

Global Coverage is Necessary for DPO

Proposition 4.1. *Denote* π_{ref} *as* any *reference policy such that* Assumption 4.1 *breaks. Let* Π_{dpo} *be the set of DPO returned policies such that* Assumption 4.3 *holds. Then there exists policy* $\pi \in \Pi_{dpo}$ *such that* $J(\pi) = -\infty$.

Proof sketch. Without loss of generality, we consider a promptless setting, and assume that the response space is $\mathcal{Y} = \{y_1, y_2, y_3\}$. Again without loss of generality, we assume π_{ref} only covers y_1 and y_2, and thus Assumption 4.1 breaks. We assume partition function $Z = 1$ for all π but we will be rigorous in the formal proof. Then consider the following policy π such that

$$\beta \log\left(\frac{\pi(y_1)}{\pi_{\mathsf{ref}}(y_1)}\right) = r^*(y_1) - \sqrt{\varepsilon_{\mathsf{dpo}}}, \quad \text{and} \quad \beta \log\left(\frac{\pi(y_2)}{\pi_{\mathsf{ref}}(y_2)}\right) = r^*(y_2) - \sqrt{\varepsilon_{\mathsf{dpo}}},$$

One can check π satisfies Assumption 4.3. Now consider the optimal policy $\pi^*(y_i) = \pi_{\mathsf{ref}}(y_i)\exp\left(\frac{1}{\beta}r^*(y_i)\right)$, for $i \in \{1, 2\}$, and $\pi^*(y_3) = 0$. Since $\pi^*(y_1) + \pi^*(y_2) = 1$, combining everything we get $\pi(y_3) > 0$, which implies $\mathsf{KL}(\pi||\pi_{\mathsf{ref}})$ is unbounded, thus we complete the proof. $\square$

Online RLHF

Lemma 4.1. *Suppose that Assumption 4.4 holds. Then for any RLHF policy π_{rlhf}, we have that*

$$\mathsf{KL}(\pi_{\mathsf{rlhf}} || \pi_{\mathsf{ref}}) := \mathbb{E}_{x\sim\rho}\left[\mathbb{E}_{y\sim\pi_{\mathsf{rlhf}}(\cdot|x)}\left[\log\left(\frac{\pi_{\mathsf{rlhf}}(y \mid x)}{\pi_{\mathsf{ref}}(y \mid x)}\right)\right]\right] \leq \frac{2R'}{\beta}.$$

Then we can show that the RLHF algorithm can guarantee performance under partial coverage:

Theorem 4.2. *Suppose that Assumption 4.4 holds. Then for any reference policy π_{ref} for which Assumption 4.2 holds with $\varepsilon_{\mathsf{kl}} = \frac{2R'}{\beta}$, and any RLHF policy π_{rlhf} with $\widehat{r}$ such that (c.r. Assumption 4.3)*

$$\mathbb{E}_{x,y\sim\rho\circ\pi_{\mathsf{ref}}}\left[\left(r^*(x,y) - \widehat{r}(x,y)\right)^2\right] \leq \varepsilon_{\mathsf{reward}},$$

we have

$$J(\pi^*) - J(\pi_{\mathsf{rlhf}}) \leq O(C_{\varepsilon_{\mathsf{kl}}}\sqrt{\varepsilon_{\mathsf{reward}}}).$$

Hybrid Preference Optimization

Algorithm 1 Hybrid Preference Optimization (HyPO)

require Pretrained LLM π_{θ_0}, reference policy π_{ref}, offline data $\mathcal{D}$, learning rate α, KL coefficient λ.

for $t = 1, \ldots, T$ **do**

Sample a minibatch of offline data $D_{\mathsf{off}} := \{x, y^+, y^-\} \sim \mathcal{D}$.

Compute DPO loss $\ell_{\mathsf{dpo}} := \sum_{x,y^+,y^- \in D_{\mathsf{off}}} \log\left(\sigma\left(\beta \log\left(\frac{\pi_{\theta_{t-1}}(y^+|x)}{\pi_{\mathsf{ref}}(y^+|x)}\right) - \beta \log\left(\frac{\pi_{\theta_{t-1}}(y^-|x)}{\pi_{\mathsf{ref}}(y^-|x)}\right)\right)\right)$.

Sample (unlabeled) online data $D_{\mathsf{on}} := \{x, y\}$ where $x \sim \mathcal{D}, y \sim \pi_{\theta_{t-1}}(x)$.

Compute $\ell_{\mathsf{kl}} := \sum_{x,y \in D_{\mathsf{on}}} \log(\pi_{\theta_{t-1}}(y|x)) \cdot \mathrm{sg}\left(\log\left(\frac{(\pi_{\theta_{t-1}}(y|x))}{(\pi_{\mathsf{ref}}(y|x))}\right)\right)$.

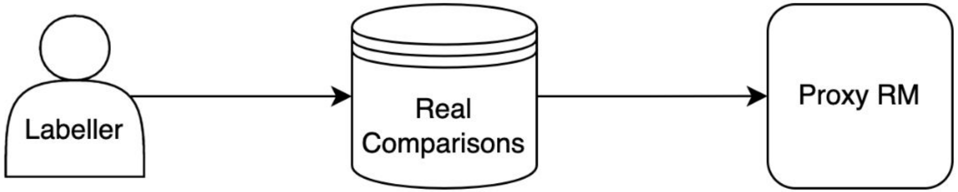
Update $\theta_t = \theta_{t-1} + \alpha \cdot \nabla_{\theta_{t-1}}(\ell_{\mathsf{dpo}} - \lambda \ell_{\mathsf{kl}})$.

return π_T.

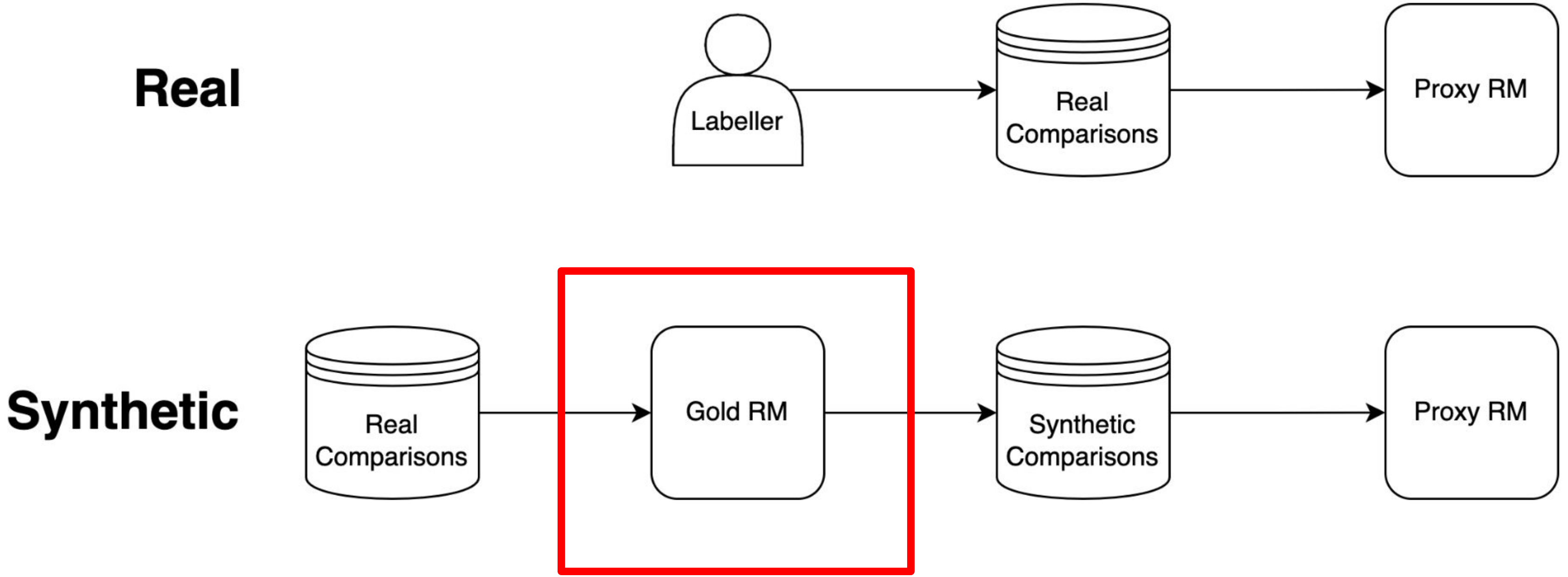
3. Empirical Experiments

Controlled Setup

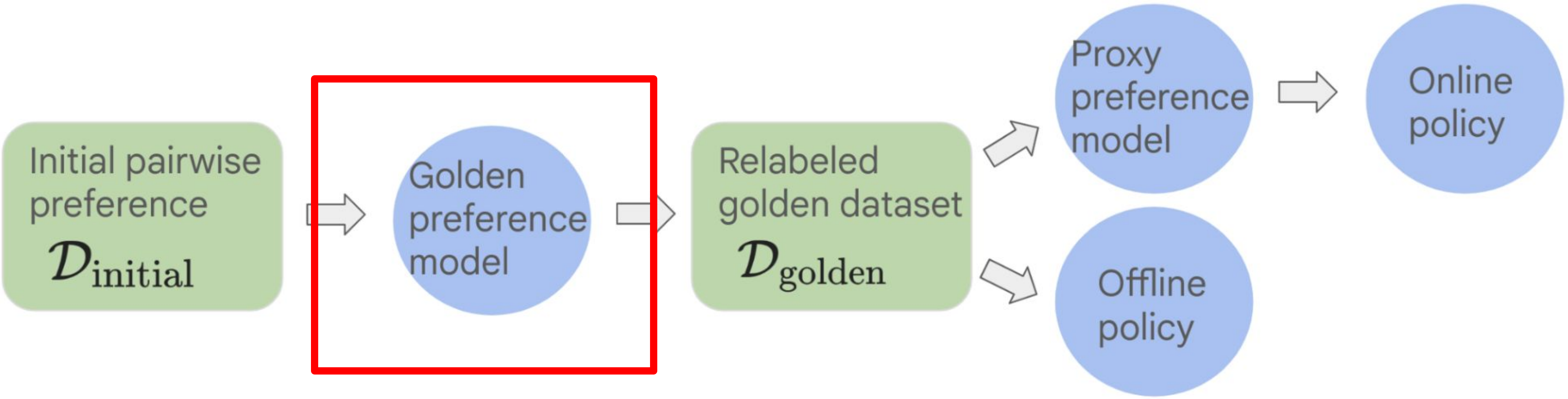
Real



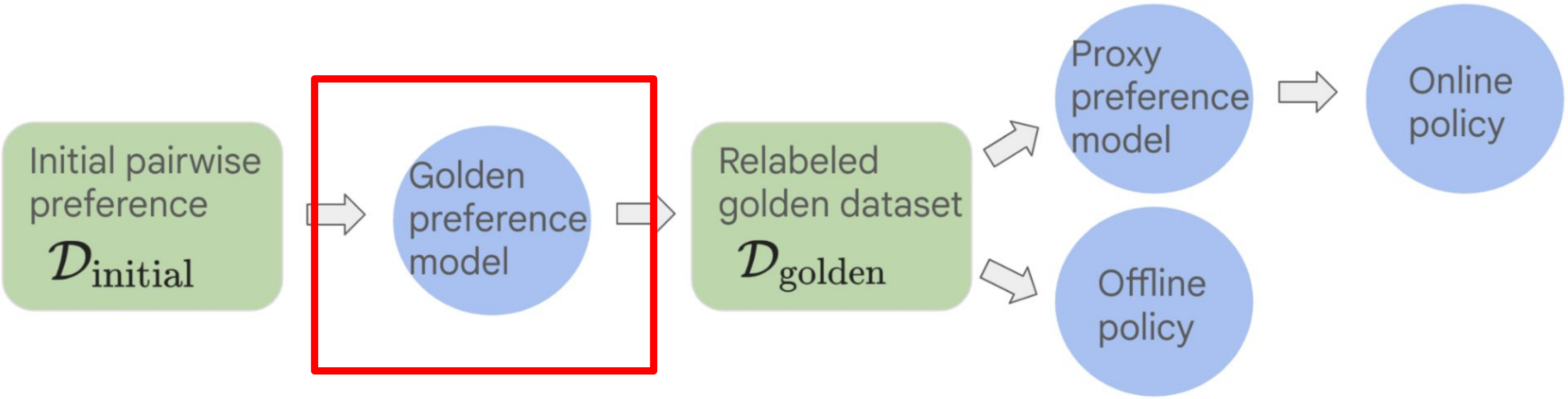
Controlled Setup



Controlled Setup



Controlled Setup



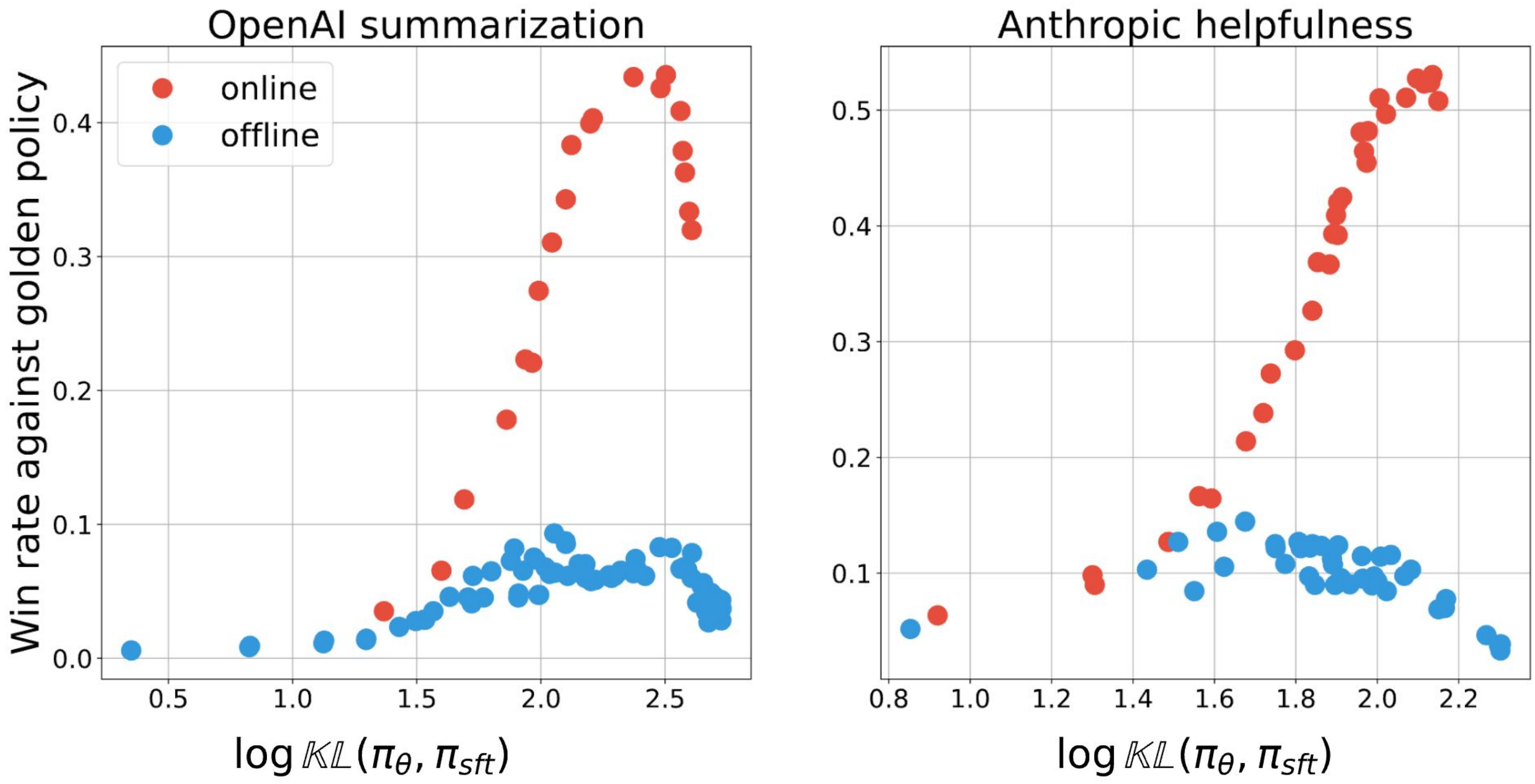
- Eval: win-rate against golden online baseline
- Judged by golden preference model

Controlled Setup

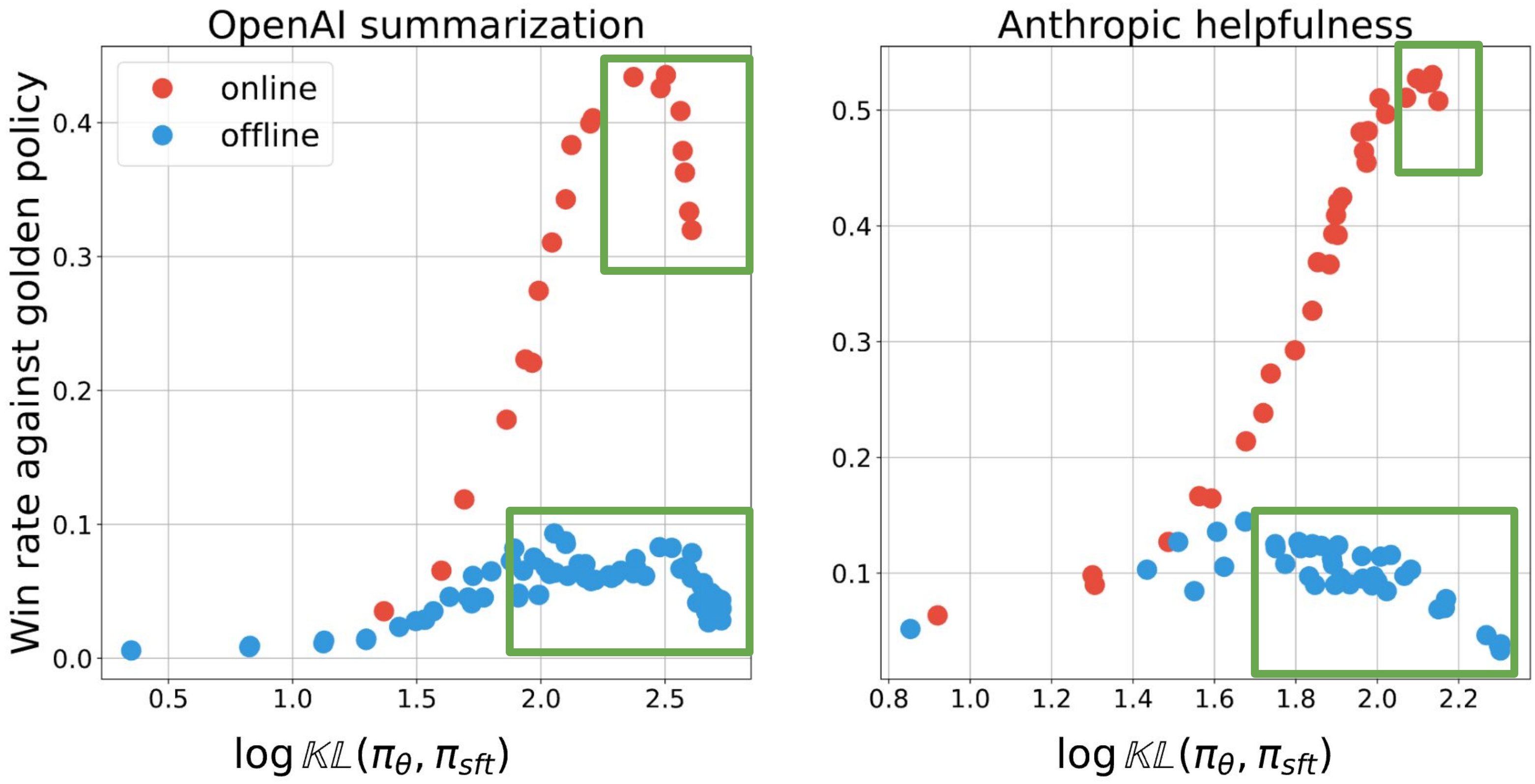
- Online vs offline versions of IPO

$$\min_{\theta} \mathbb{E}_{x\sim p,(y_w,y_l)\sim\mu}\left[\left(\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{sft}}(y_w|x)} - \log\frac{\pi_\theta(y_l|x)}{\pi_{\text{sft}}(y_l|x)} - \frac{\beta}{2}\right)^2\right]$$

Understanding the performance gap



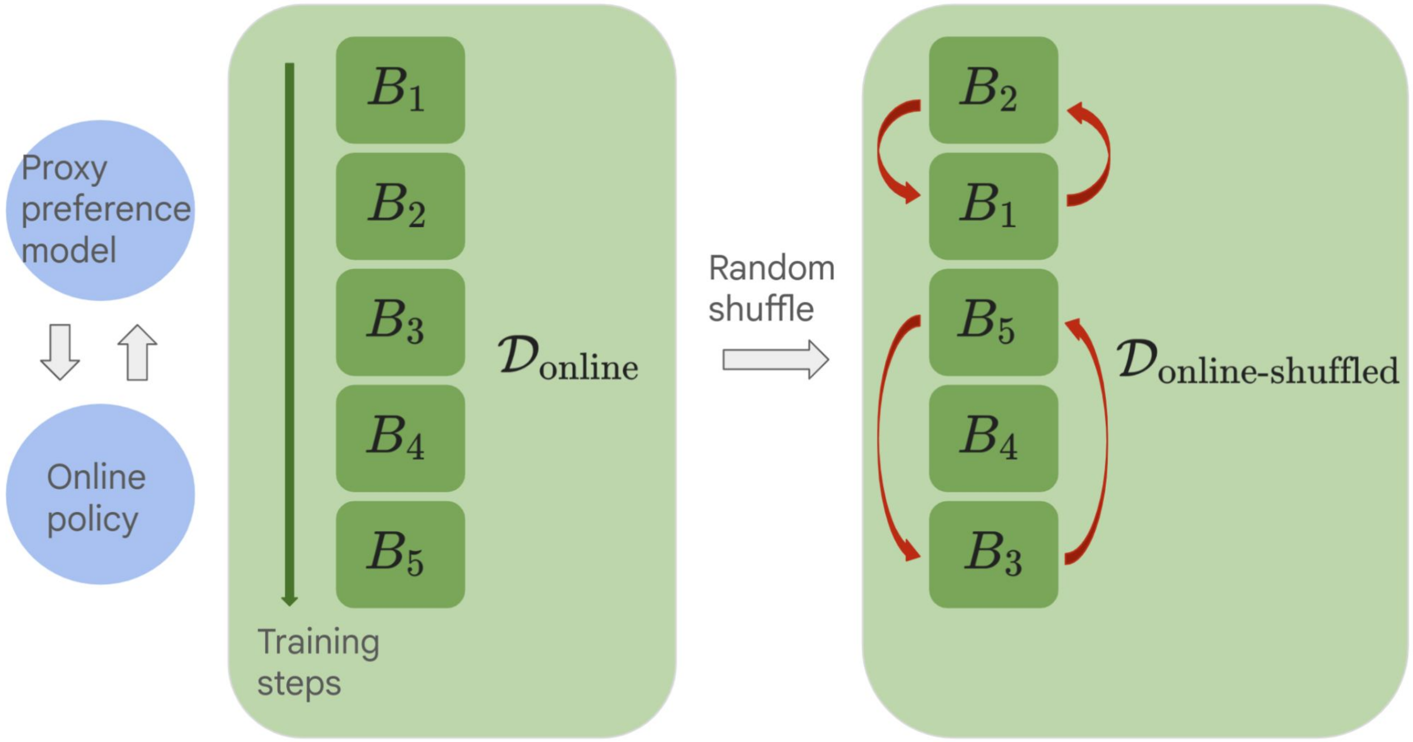
Goodhart's law



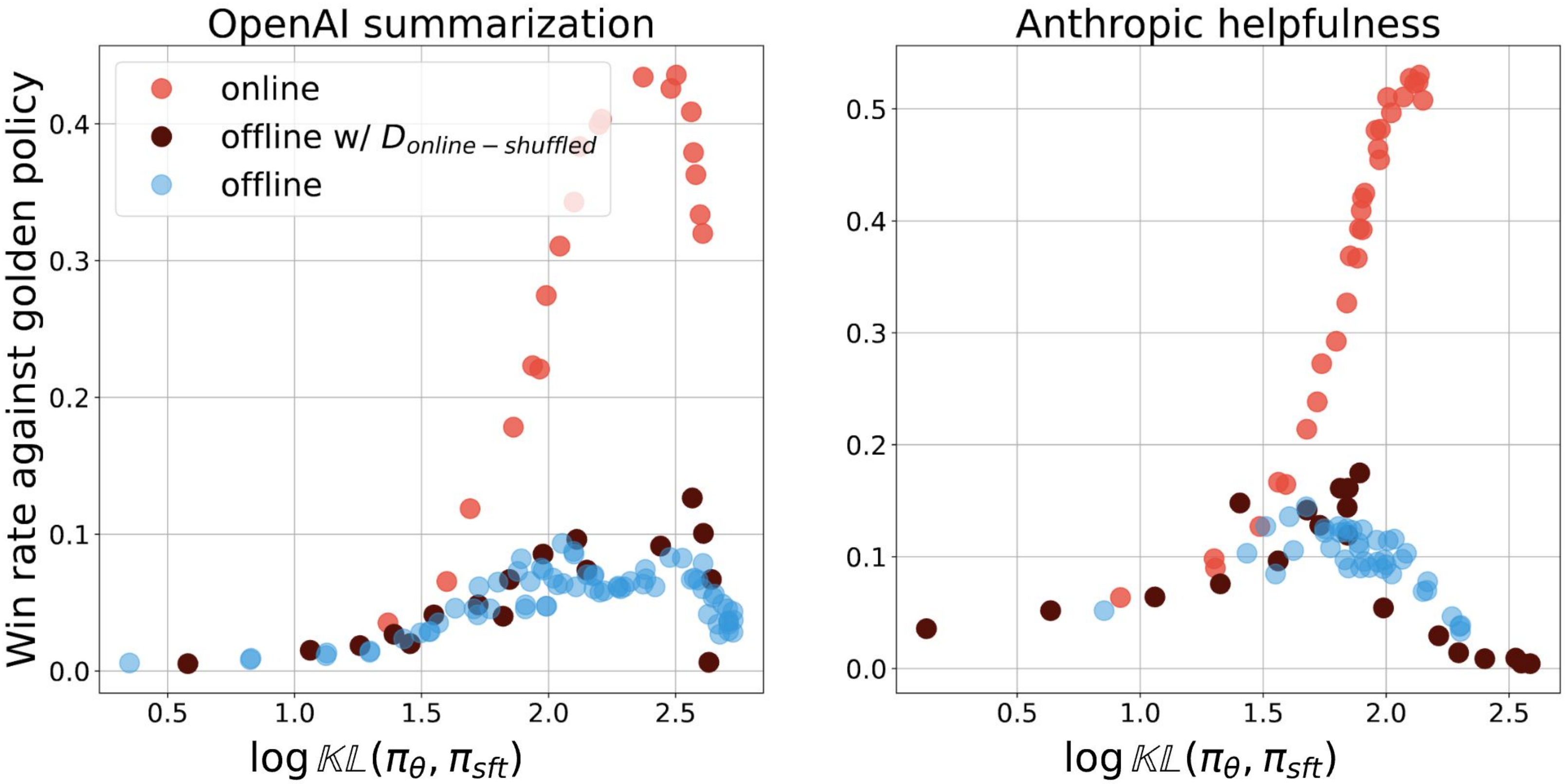
Closing the performance gap

1. Data coverage
2. Sub-optimal offline dataset
3. Loss function formulation
4. Model scale
5. ...

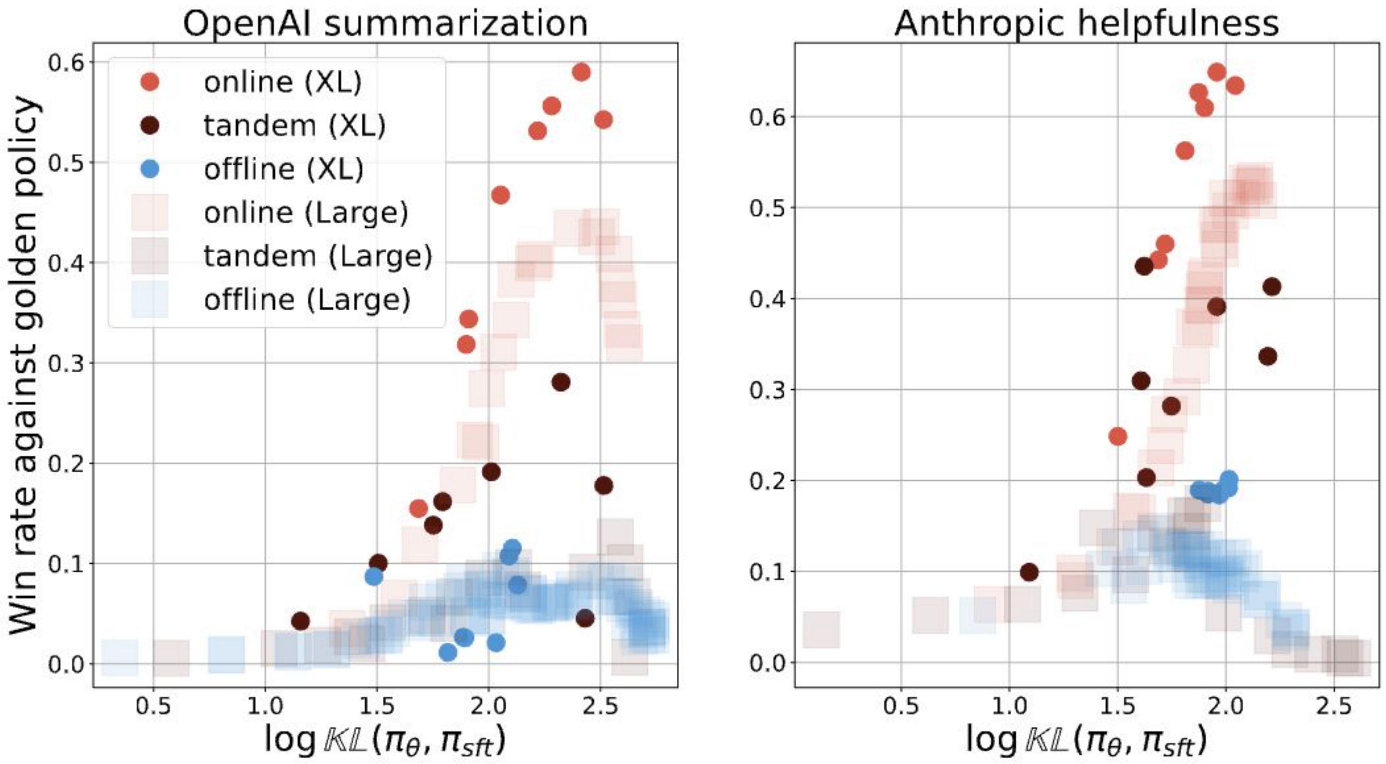
Hypothesis 1: Data Coverage



Hypothesis 1: Data Coverage



Hypothesis 4: Model Scale



TL;DR

1. Empirically, on-policy data (in some form) leads to better performance
2. Many ways to get this kind of data
   a. Online RLHF
   b. Iterative (offline RLHF)