

# Can RLHF be More Efficient with Imperfect Reward Models? A Policy Coverage Perspective

ETH zürich

Google

Jiawei Huang<sup>1</sup> Bingcong Li<sup>1</sup> Christoph Dann<sup>2</sup> Niao He<sup>1</sup>

Contact Email: jiawei.huang@inf.ethz.ch

<sup>1</sup>ETH Zurich <sup>2</sup>Google Research



## Motivation

- Sample efficiency (# of human annotations) is crucial in online RLHF.
- Previous works focus on strategic exploration, while we study from a different perspective—**transfer learning**.
- Rich scenarios with imperfect but related source rewards available:
  - Reward models from relevant tasks
  - Easy-to-access evaluation metric other than human feedback
  - Guidance from advanced LLMs

**Key Question: How to improve sample efficiency in online RLHF by leveraging those source reward models?**

## Setting and Assumptions

### A Contextual Bandit Framework

- $\mathcal{S}$ : prompt space;  $\mathcal{A}$ : response space,
- $\pi: \mathcal{S} \rightarrow \Delta(\mathcal{A})$ : LLM as a policy. W.L.O.G.,  $\pi(\cdot|s) > 0$  everywhere.
- $r^*$ : unknown true (human intrinsic) reward model,
- Learning objective:

$$\pi_{r^*}^* \leftarrow \arg \max_{\pi} J_{\beta}(\pi) := \mathbb{E}_{s \sim \rho, a \sim \pi} [r^*(s, a)] - \beta \text{KL}(\pi \| \pi_{\text{ref}}), \quad (1)$$

with  $\rho$  as prompt distribution,  $\pi_{\text{ref}}$  as the reference policy, and yields:

$$\pi_{r^*}^*(a|s) \propto \pi_{\text{ref}}(a|s) \exp\left(\frac{r^*(s, a)}{\beta}\right), \quad (2)$$

- Bradley-Terry preference model ( $\sigma$  denotes sigmoid function)

$$\mathbb{P}_{r^*}(\mathbb{I}[a \succ \tilde{a}] | s, a, \tilde{a}) = \sigma(r(s, a) - r(s, \tilde{a})).$$

### Standard Assumptions

- **Bounded rewards:**  $r^* \in [0, R]$ ,
- **Function approximation:** A policy class  $\Pi$  is available.
  - (i)  $\pi_{r^*}^* \in \Pi$ . (ii)  $\forall \pi \in \Pi, \|\log \frac{\pi}{\pi_{\text{ref}}}\|_{\infty} \leq \frac{R}{\beta}$ .

### Online RLHF with Reward Transfer Setup

- **Online human feedback:** Query to  $\mathbb{P}_{r^*}(\cdot | s, a, \tilde{a})$  with arbitrary  $s, a, \tilde{a}$ .
- **Source reward models:**  $W$  source RMs  $r^1, \dots, r^W \in [0, R]$  available,
  - no prior knowledge on their quality
  - due to Eq. (2), any LLM policy can be converted as a RM.

## Blessing of Regularization: A Policy Coverage Perspective

**Transfer RL has been studied for decades.  
But the KL-regularization in Eq. (1) makes something different!**

**Policy Coverage:** The coverage coefficient of policy  $\tilde{\pi}$  by another policy  $\pi$ :

$$\text{Cov}^{\tilde{\pi}|\pi} := \mathbb{E}_{s \sim \rho, a \sim \tilde{\pi}} \left[ \frac{\tilde{\pi}(a|s)}{\pi(a|s)} \right].$$

### Why Policy Coverage Perspective?

- It serves as fundamental complexity measure in both online [1] and offline [2] RLHF.
- Optimization and exploration on policy (LLM) space is more efficient in RLHF

**Key Lemma:** special structure due to KL regularization ( $\beta > 0$ )

**Lemma 3.1:** For any  $\pi \in \text{conv}(\Pi) \cup \{\pi_{r^w}^*\}_{w=1}^W$ ,

$$\text{Cov}^{\pi_{r^*}^*|\pi} = 1 + O\left(e^{-\frac{2R}{\beta}}\right) \cdot \frac{J_{\beta}(\pi_{r^*}^*) - J_{\beta}(\pi)}{\beta}.$$

### Interpretation

- $\text{Cov}^{\pi_{r^*}^*|\pi}$  can be identified by  $\pi$ 's value gap
  - vastly distinguished from pure reward maximization
- KL-reg "reconciles" exploration and exploitation
  - **exploiting policies with high policy value coincides with exploration!**
- **Theorem 3.2** [Informal]: Offline learning on the online dataset collected by any no-regret algorithm yields a policy  $\pi_{\text{OFF}}$  converges to  $\pi_{r^*}^*$  at rate of  $\tilde{O}(T^{-1/2})$ ,
  - **no dependence on complexity of  $\Pi$ !**
  - **faster** than the convergence rates in existing online RLHF literature [1].

**New Insights for Transfer Learning—Find and transfer from  $\pi$  with the lowest  $\text{Cov}^{\pi_{r^*}^*|\pi}$**

- **Principle 1:** Transfer from the policy with the highest policy value.
- **Principle 2:** Keep tracking  $\pi_{\text{OFF}}$  and treat it as a transfer candidate.
  - we call this "self-transfer learning", and call such a  $\pi_{\text{OFF}}$  "self-transfer policy".

## References

- [1] Xie et al., Exploratory preference optimization: Harnessing implicit q\*-approximation for sample-efficient rlhf.  
[2] Liu et al., Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer.

## TPO: A Transfer Learning Algorithm with Provable Benefits

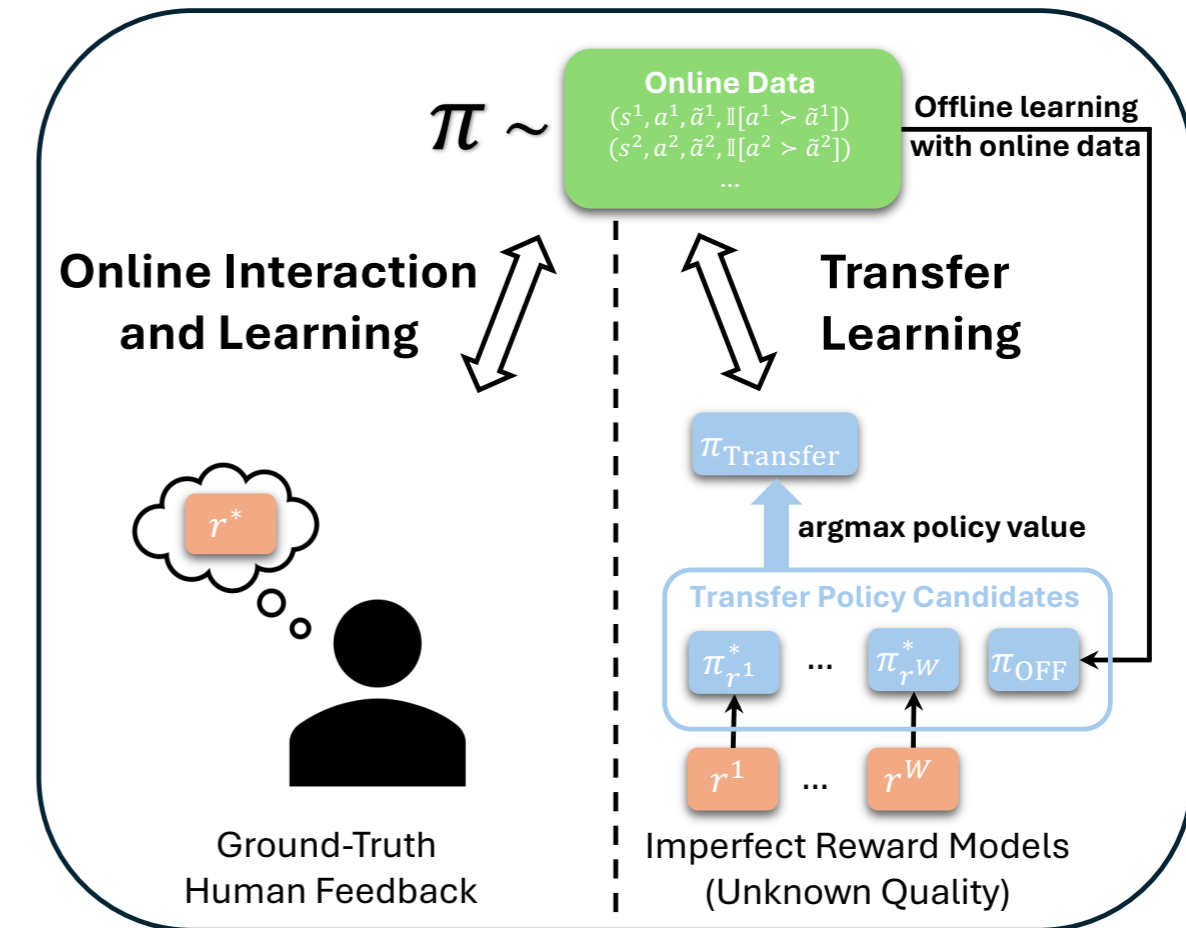


Figure 1. Illustration of TPO (Transfer Policy Optimization).  $\pi_{r^w}^*$  denotes the optimal policy w.r.t.  $r^w$ .

**Theoretical Guarantees:** Define  $\Delta_{\min} := \min_{w \in [W]} J_{\beta}(\pi_{r^w}^*) - J_{\beta}(\pi_{r^*}^*)$ .

- When  $T \leq \tilde{O}\left(\frac{1}{\Delta_{\min}^2}\right)$ ,  $\text{Reg}(T) = \tilde{O}(\sqrt{WT})$  – Reduce dependence on complexity of  $\Pi$  to  $W$
- When  $T > \tilde{O}\left(\frac{1}{\Delta_{\min}^2}\right)$ ,  $\text{Reg}(T) = \tilde{O}(\sqrt{T})$  – No dependence on complexity of  $\Pi$

## Empirical TPO: From Theory to Practice

- TPO estimates policy value to identify the one cover  $\pi_{r^*}^*$  the best.
- However, value estimation is computationally expensive.
- Is there a more accessible indicator for  $\text{Cov}^{\pi_{r^*}^*|\pi}$ ? **Yes, the win rates!**
- **Lemma 5.1** A lower bound for  $\text{Cov}^{\pi_{r^*}^*|\pi}$  given an arbitrary comparator  $\pi_{\text{Comp}}$ :

$$\text{Cov}^{\pi_{r^*}^*|\pi} \geq \max_{\gamma > 0} \left( \sqrt{(\gamma + 2 \cdot \mathbb{P}_{r^*}(\pi \succ \pi_{\text{Comp}})) \log \frac{1 + \gamma}{\gamma}} + \sqrt{\frac{J_{\beta}(\pi_{r^*}^*) - J_{\beta}(\pi_{\text{Comp}})}{2\beta}} \right) - 1$$

- **Inspired empirical algorithm design:**

- Transfer policy selection as a Multi-Armed Bandit problem.
  - o Selecting policy with high win rate by UCB.
- Compute  $\pi_{\text{Online}}$  by any online method (e.g. iterative DPO, XPO),
  - o Take  $\pi_{\text{Online}}$  as the comparator  $\pi_{\text{Comp}}$ , which continuously improves
  - o **Transfer from the expert until beat it**
- Scalable in practice!

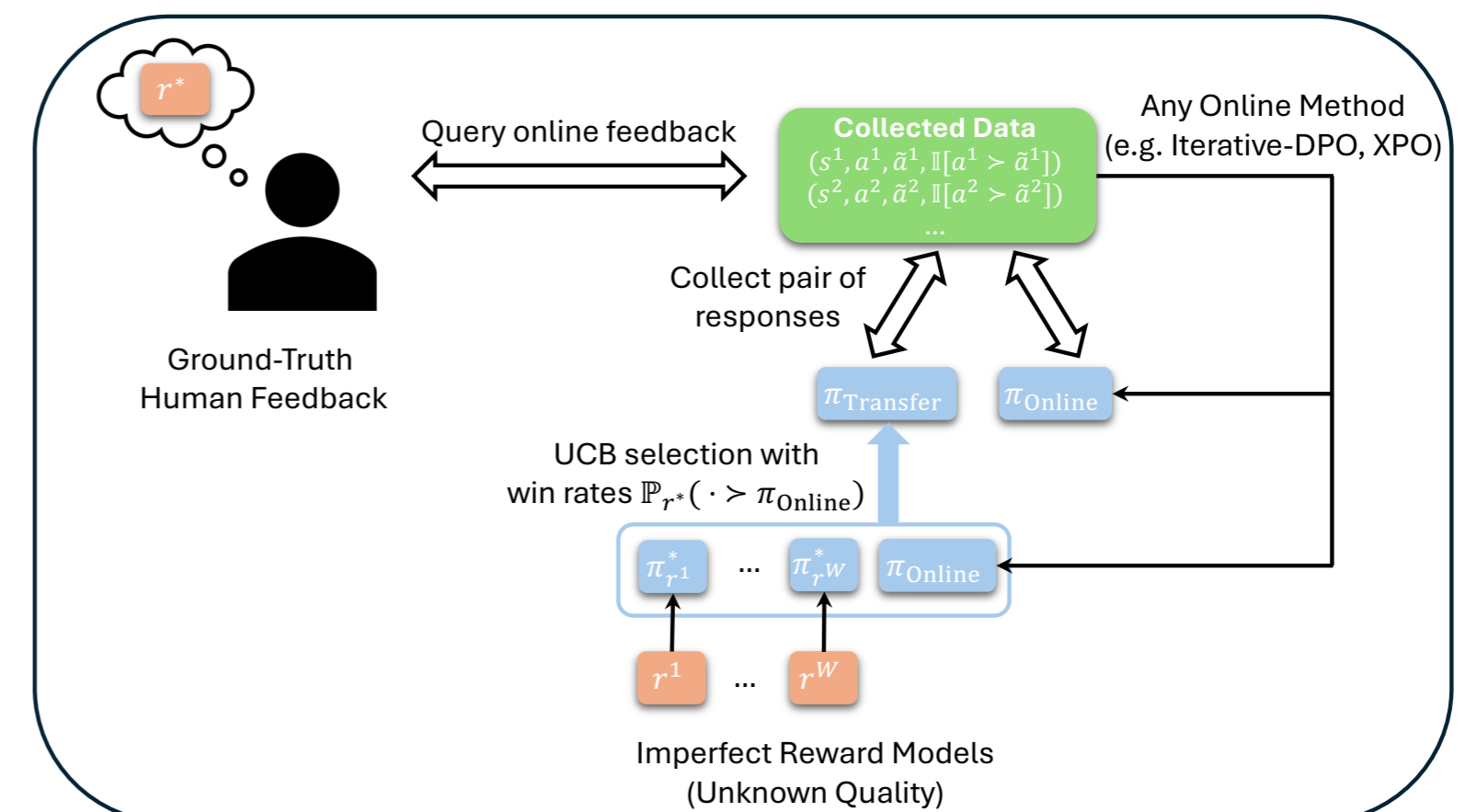


Figure 2. Illustration of empirical TPO

## Experiments in Summarization Tasks with T5

- Fine-tuning T5-small (80M) on XSum dataset.
- 4 source reward models:
  - 2 metrics of similarity with human summary: (a) ROUGE score (b) BERTScore
  - 2 advanced LLMs: (c) T5-Base (250M) (d) T5-Large (770M)
- Llama3-8B to simulate human feedback.

	Without Transfer	Purely Exploit ROUGE	Purely Exploit T5-Large
Iter 1	52.1 ± 1.2	53.1 ± 1.1	49.5 ± 0.9
Iter 2	53.3 ± 1.6	54.5 ± 1.3	49.1 ± 0.4
Iter 3	54.0 ± 1.2	53.3 ± 1.5	50.6 ± 0.3

Table 1. Win rates (%) of the policies trained by empirical TPO competed with 3 baselines.

### Interpretation

- Transfer learning makes online RLHF more efficient.
- Without prior knowledge, quickly adapt to the best source model (T5-Large) without being trapped by low-quality ones (ROUGE score).
- Switch back to online learning when source models are no longer helpful.