Back to Blackwell: Closing the Loop on Intransitivity in Multi-Objective Preference Fine-Tuning

Jiahao Zhang* [1] Lujing Zhang* [1] Keltin Grimes [1] Zhuohao Yu [1] Gokul Swamy [1] Zhiwei Steven Wu [1]

[1]Carnegie Mellon University

🔑 Problem: Intransitivity in Preference Feedback

Intransitive (i.e., $A \succ B, B \succ C \implies A \succ C$) preferences are a key problem for the standard RLHF pipeline as *no* reward can rationalize them. They are a consequence of two factors:

1. When performing RLHF at scale, we collect conflicting data from diverse annotators.

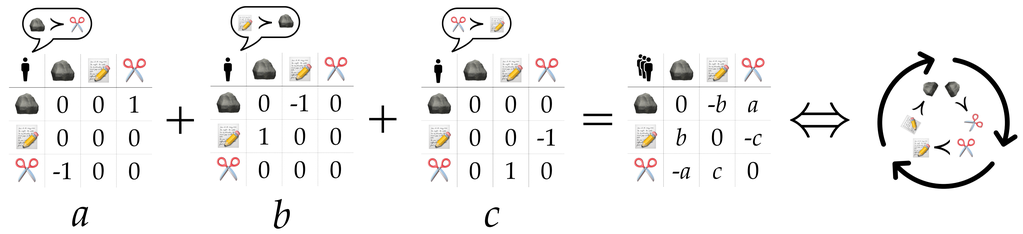


Figure 1. While each annotator has rational preferences (e.g. 🪨 ≻ ✂️), aggregating them together leads to cyclic / intransitive preferences. As we increase the size of the choice space (e.g. all completions to a prompt), the likelihood of cyclic preferences skyrockets 🚀.

2. A single rater may compare different pairs of options based on different **rubrics**, leading to cycles (e.g., 📝 ≻ 🪨 to send a letter, 🪨 ≻ ✂️ as a doorstop, ✂️ ≻ 📝 for cutting).

In either setting, reward-based RLHF (reward model + PPO, DPO) will provably have to ignore some aspect of human preferences – either a subgroup or criteria.

🔑 **Goal:** A scalable preference fine-tuning algorithm to handle **both** sources of intransitivity.

Back to Blackwell: A New Solution Concept for RLHF

Given policy class $\pi \in \Pi$, where $\Pi \subseteq \{\mathcal{X} \to \Delta(\mathcal{Y})\}$ and a m-dimensional preference oracle $\mathcal{P} : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0,1]^m$, we define the ***Maximum Entropy Blackwell Winner*** as:

$$\pi_{\text{MEBW}} = \arg\max_{\pi \in \Pi} \min_{w \in (\Delta_m)^{\mathcal{X}}} \min_{\pi'} \mathbb{E}_x[\langle w(x), \mathcal{P}(x, \pi, \pi') \rangle] + \beta \mathbb{E}_x[\mathbb{D}_{\text{KL}}(\pi(\cdot \mid x) \| \pi'_{\text{base}}(\cdot \mid x))]$$

✓ **Pairwise Preferences:** no assumption of a single, shared population-level reward function.

✓ **Multi-Objective:** can train policies that make nuanced trade-offs / thresholds.

✓ **Configurable Pessimism:** Scaling β controls how "local" the equilibrium is.

✗ **Adversarial training:** Infeasible with LLMs. No Minimax Theorem $\implies$ no Self-Play.

🔑 **Question:** How do we compute MEBWs at the scale of large language models?

🔑 Solution: PROSPER: A Scalable Algorithm for MEBWs

🔑 **Idea 1:** Use the closed-form for maximum entropy RL to eliminate the inner optimization.

$$\pi_{\text{MEBW}} = \arg\max_{\pi \in \Pi} \mathbb{E}_x \left[\min_{i \in [m]} h_i(\pi, x) \right],$$

where $h_j(\pi, x) = -\beta \log \mathbb{E}_{y' \sim \pi_{\text{base}}(\cdot|x)} \left[\exp\left(-\mathbb{E}_{y \sim \pi}[\mathcal{P}_i(x, y, y')]/\beta \right) \right]$ is the partition fn.

🔑 **Idea 2:** As h_j is concave in π, we can reduce solving the above to online mirror descent.

Practically, as in REBEL, we iteratively solve the following square loss regression problem:

PROSPER (PReference Optimization with a Single Player over Entire Rubrics)

For each iteration $t \in [T]$, solve square-loss regression problem:

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \mathbb{E}_{x \sim \rho, y \sim \pi_t(\cdot|x), y' \sim \mu(\cdot|x)} \left(\frac{1}{\eta} \left(\ln \frac{\pi_\theta(y|x)}{\pi_{\theta_t}(y|x)} - \ln \frac{\pi_\theta(y'|x)}{\pi_{\theta_t}(y'|x)} \right) - (g_t(x,y) - g_t(x,y')) \right)^2$$

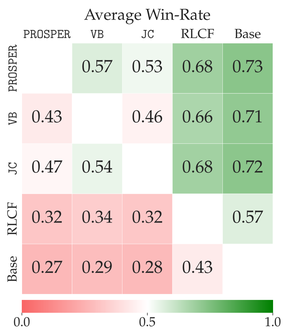
where $g_t(x,z) = \min_{i \in [m]} \frac{\mathbb{E}_{y' \sim \pi_{\text{base}}(\cdot|x)}[P_i(x,z,y') \exp(-\mathbb{E}_{y \sim \pi(\cdot|x)}[P_i(y,y')]/\beta)]}{\mathbb{E}_{y' \sim \pi_{\text{base}}(\cdot|x)}[\exp(-\mathbb{E}_{y \sim \pi(\cdot|x)}[P_i(y,y')]/\beta)]}$ is the gradient of the above.

In our paper, we prove that after running PROSPER for T iterations, $V(\pi^\star) - V(\hat{\pi}) \leq O(\sqrt{1/T} + \sqrt{C_{\mu \to \pi^\star} \epsilon})$, where ϵ is the regression training error and $C_{\mu \to \pi} = \max_{x,y} \frac{\pi(y|x)}{\mu(y|x)}$.

Experimental Results on LLM-as-a-judge

We finetune the 3B and 7B parameter Qwen2.5-Instruct models on WildChecklists, which consists of prompts and associated rubrics. We use Qwen3-14B as preference oracle.

1. Baseline: RLCF (RL from Checklist Feedback, reward-based)
2. Ablations: PROSPER-VB (✗ adversarially chosen competitor) and PROSPER-JC (✗ adversarially weighted objectives)



	Arena-Hard		AlpacaEval	
	Vanilla	Style-Controlled	Vanilla	Length-Controlled
Qwen2.5-7B-Instruct	42.4	44.2	37.1	25.32
+ RLCF	42.6	45.0	41.4	17.24
+ PROSPER-JC	44.2	42.0	<u>55.3</u>	**38.21**
+ PROSPER-VB	<u>47.6</u>	<u>45.5</u>	51.2	33.64
+ PROSPER	**49.2**	**46.1**	**55.4**	<u>37.61</u>

Figure 2. From left to right: (i) win-rate matrix on held-out prompts, (ii) results on instruction following and general chat benchmarks.

Feel free to reach out to jiahaozhang@cmu.edu with questions!