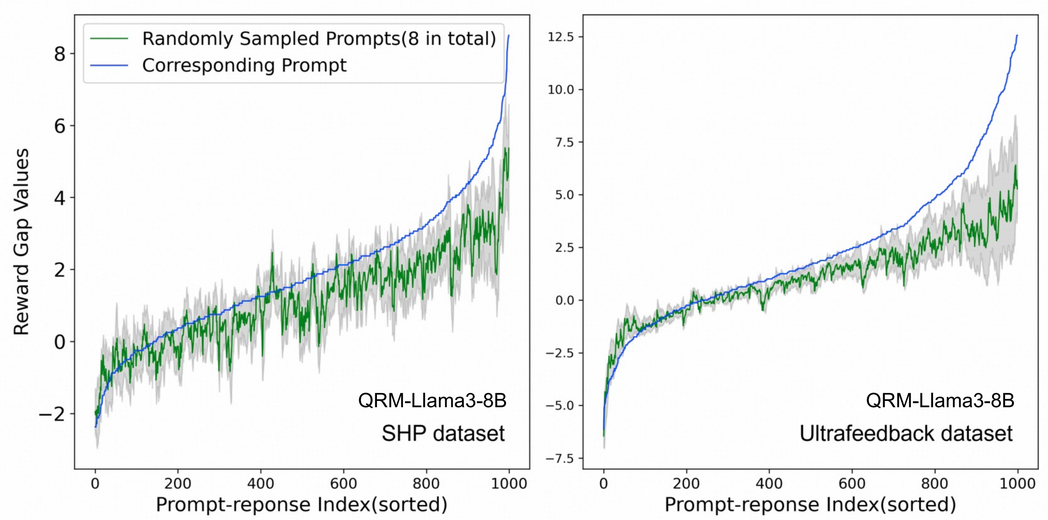
Information-Theoretic Reward Decomposition for Generalizable RLHF

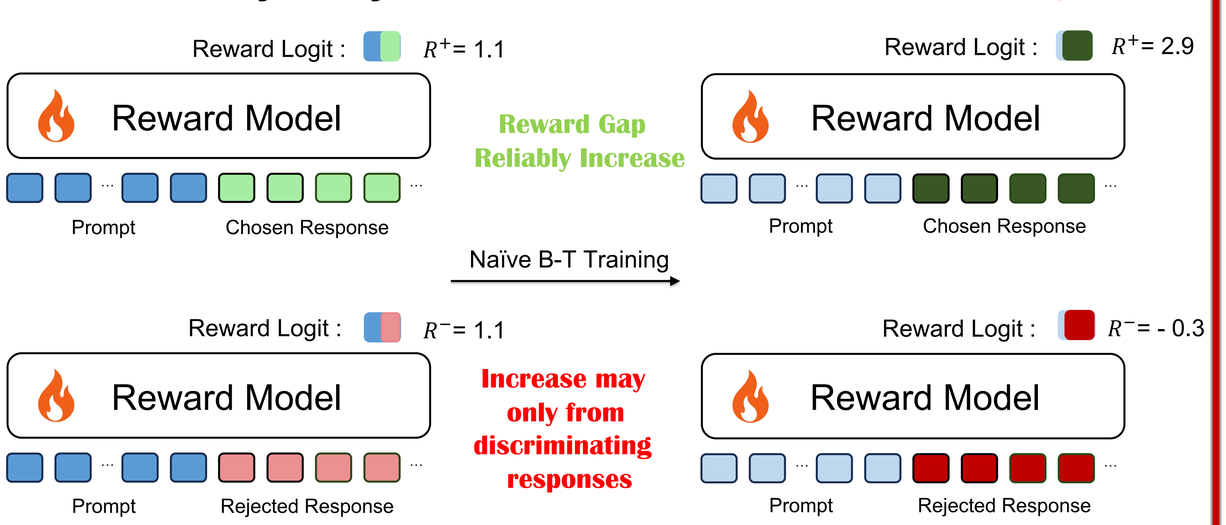
NIPS 2025 Liyuan Mao, Haoran Xu, Amy Zhang, Weinan Zhang, Chenjia Bai

Defect of Bradley-Terry Reward: $\Delta r \uparrow \neq$ Generalizability $\uparrow$

Prompt and Response can Ineuqally Influence Reward Value



Naïve Bradley-Terry RM can't Avoid the Dominance of Response



Why such reward model is not truly generalizable?

1. preference order between two responses may reverse under different prompts
2. response's dominant role can introduce spurious preference (e.g. response length, format)

Our Solution

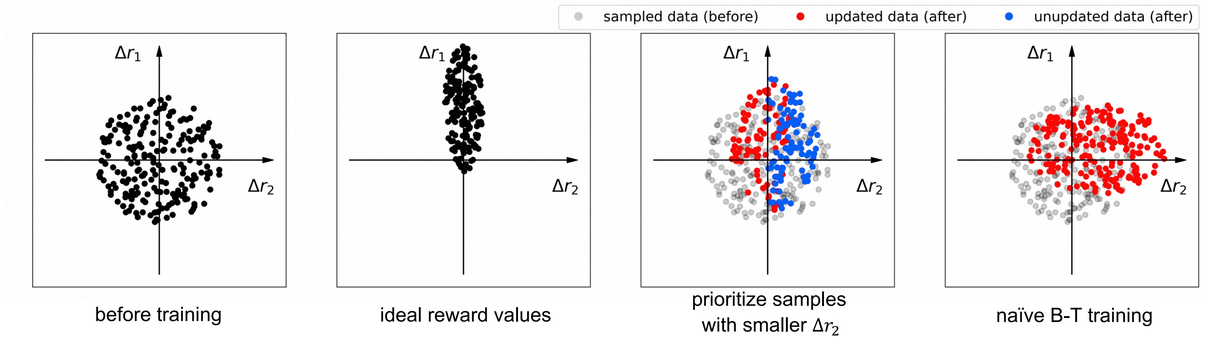
Decompose Reward Gap

$$\Delta r(x, y_+, y_-) = \Delta r_2(y_+, y_-) + \Delta r_1(x, y_+, y_-)$$

Prompt-free reward gap
(If the prompt is unknown, which response does the reward model prefer?)

Prompt-related reward gap
(If the prompt is specified, what extra preference information does it offer?)

Selectively Update with Extracted Δr_2



Always Learn More Generalizable Preference / Eliminate Existing Bias

Experiments on Customized Dataset (Based on SHP Dataset)