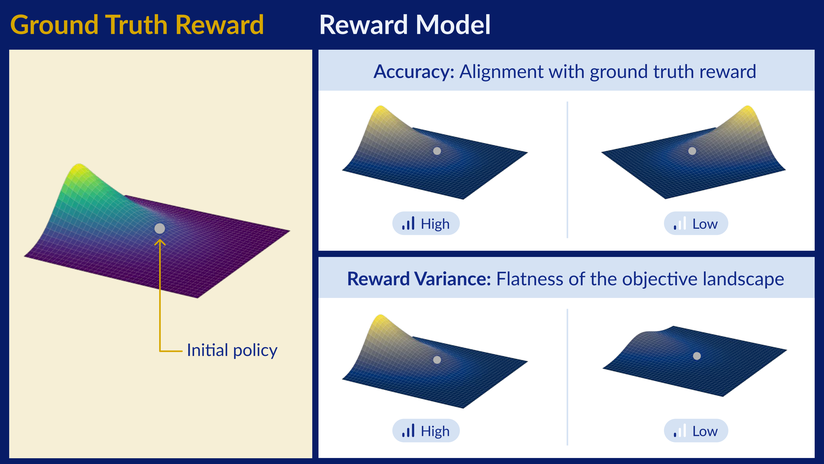
What makes a good reward model for RLHF?



Aside from being accurate, it needs to induce **sufficient reward variance!**

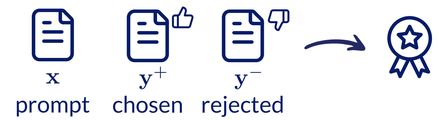
What Makes a Reward Model a Good Teacher? An Optimization Perspective

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason D. Lee, Sanjeev Arora

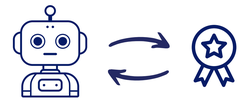
Princeton Language and Intelligence, Princeton University

Reinforcement Learning from Human Feedback (RLHF)

- Train **reward model (RM)** to approximate **ground truth reward**



- **Align language model (LM)** by maximizing learned reward **via policy gradient**



Accuracy

RMs are typically evaluated via **accuracy** (e.g., RewardBench; Lambert et al. 2024)

Is $r(\mathbf{x}, \mathbf{y}^+) > r(\mathbf{x}, \mathbf{y}^-)$? +1 Yes / No 0

Q: Are more accurate reward models (RMs) necessarily better teachers? ✗

1 Main Result

Low reward variance leads to slow optimization, even if the RM is highly accurate

$$\mathrm{Var}_{\mathbf{y} \sim \pi_\theta(\cdot|\mathbf{x})}[r(\mathbf{x}, \mathbf{y})]$$

Notation

r - RM $\quad \pi_\theta$ - LM $\quad \mathcal{S}$ - training prompts

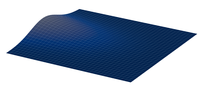
Theorem (low reward variance → slow optimization)

The time it takes for the expected RM and ground truth rewards to increase by any additive constant is:

$$\Omega\left(\mathbb{E}_{\mathbf{x} \sim \mathcal{S}}\left[\mathrm{Var}_{\mathbf{y} \sim \pi_{\theta_{\text{init}}}(\cdot|\mathbf{x})}[r(\mathbf{x}, \mathbf{y})]\right]^{-\frac{1}{3}}\right)$$

*Result applies to any RL setting with softmax policies

Proof Idea: Low reward variance causes a flat objective landscape for softmax policies (includes LMs as a special case)



2 Implication I

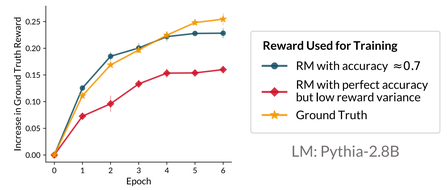
More accurate RMs are not necessarily better teachers for RLHF

Theorem (more accurate RMs are not always better)

For any LM, there exist a **perfectly accurate** r_{per} and **relatively inaccurate** $\bar{r}$ s.t. ground truth reward increases **arbitrarily slower** when using r_{per}

*Same result holds with almost any accuracy values

Experiments: RLHF on UltraFeedback using different RMs



Result: Even perfectly acc RMs can underperform less acc RMs due to low reward variance!

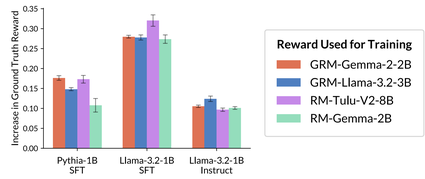
3 Implication II

Existing RM benchmarks suffer from **fundamental limitations**

Theorem (for different LMs different RMs are better)

The same RM can induce high reward variance and work well for one LM, but induce low reward variance and work poorly for another LM

Experiments: RLHF on UltraFeedback using different LMs and RMs



Result: For different LMs different RMs are better

↓

Benchmarks evaluating RMs in isolation of the LM being aligned are fundamentally limited