# Rethinking Reward Model Evaluation: Are We Barking up the Wrong Tree?

*Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, Xing Yu, Xinyu Lu,*

*Ben He, Xianpei Han, Debing Zhang, Le Sun*

中国科学院大学 University of Chinese Academy of Sciences

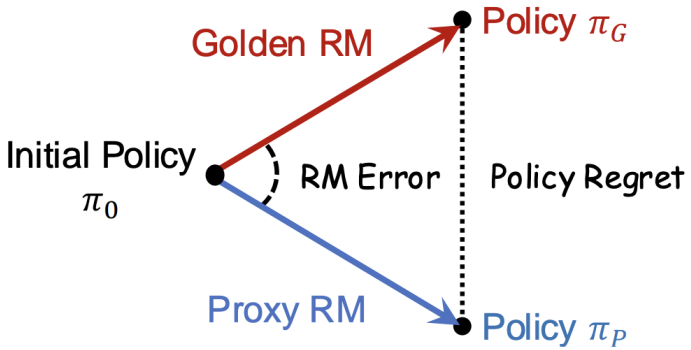
ISCAS 中国科学院软件研究所 Institute of Software Chinese Academy of Sciences

中文信息处理实验室-让机器理解语言 Chinese Information Processing Laboratory

小红书

# Introduction

**Reward Models (RMs) are crucial for aligning LLMs with human preferences in RLHF**




**Challenge: Imperfect RMs lead to performance deterioration in downstream policies**

- **RM Error:** Discrepancies between proxy RM and true human preferences, leading to misaligned optimization targets
- **Policy Regret:** Performance gap between policies optimized toward proxy RM versus ideal reward function

# *Introduction*

**Current Evaluation Practice**

- Measuring accuracy on validation sets
  - Simple and widely adopted
  - Remains unclear how it reflect downstream performance
- Evaluation on downstream tasks
  - Costly and time-consuming
  - Cannot distinguish problems between RM or R

**Research Questions**

- Does the RM error measured by accuracy correlate with policy regret?
- How to better measure RM error for policy regret prediction?
- What's the relationship between RM error and policy regret?

# *Experiment Framework*

**Synthetic Setting**:

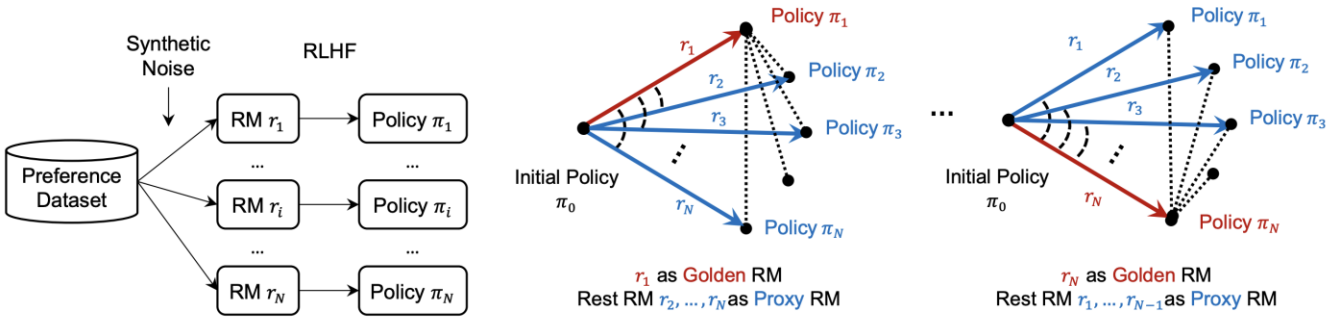
Use one reward model as "golden standard" to replace human preferences

**Proxy-Golden RM Pairs**:

Create N=10 reward models by flipping different percentages of preference pairs

**Policy Optimization**:

Best-of-n sampling (BoN) and PPO algorithms



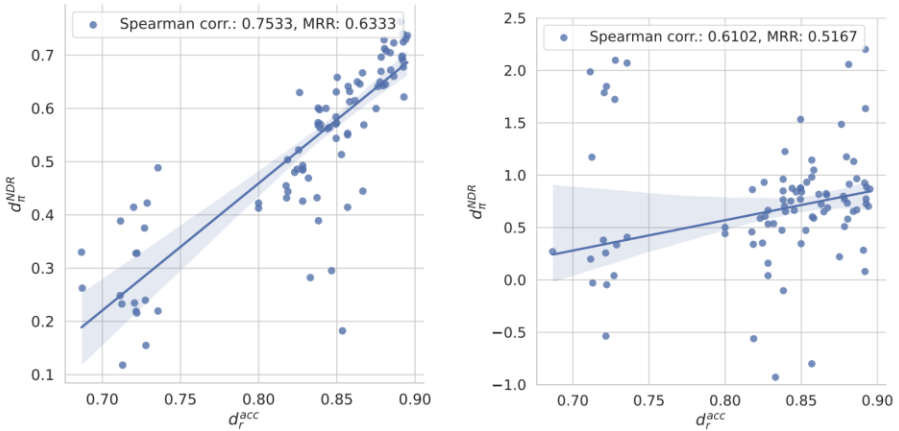

(a) Synthetic pipeline of Reinforcement Learning from Human Feedback.

(b) Proxy-golden RM pairs collection for evaluating the correlation between the RM error and policy regret.

# RQ1 - Does RM accuracy correlate with policy regret?

***Key Finding 1:*** **RM evaluation accuracy is positively related to policy regret, but even with similar accuracy, policies can exhibit different levels of regret**

- Positive but weak correlation between accuracy and policy regret
- Similar accuracy can lead to very different downstream performance




(a) Trend between accuracy $d_r^{acc}$ and the policy regret $d_\pi^{NDR}$ under BoN setting.

(b) Trend between the accuracy $d_r^{acc}$ and the policy regret $d_\pi^{NDR}$ under PPO setting.

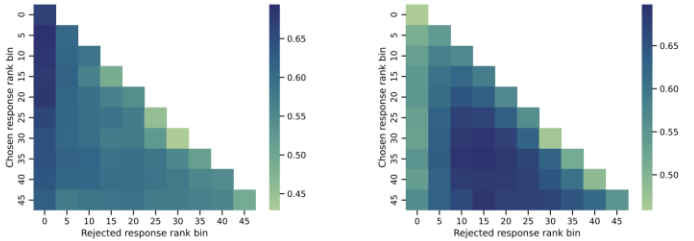
# RQ2 – How to better measure RM error

***Key Finding 2:*** **The rank of responses affects correlation more than the model used to sample them**

- Sampling from different models doesn't consistently improve correlation
- The rank of chosen/rejected samples matters more for prediction

Table 2: The correlation between policy regret and accuracy on datasets with responses sampled from different models. The **Origin** represents the result on the RewardBench dataset. The highest results are highlighted in **bold**. The model used for downstream optimization is suffixed with *.

| Model | BoN | | PPO | |
|---|---|---|---|---|
| | Spearman corr. | MRR | Spearman corr. | MRR |
| Mistral-7B | 0.680±0.042 | 0.579±0.079 | 0.580±0.037 | 0.573±0.073 |
| Llama3-8B* | 0.644±0.038 | 0.598±0.080 | **0.648±0.047** | 0.551±0.061 |
| Qwen2-7B | 0.680±0.030 | 0.529±0.074 | 0.603±0.039 | **0.599±0.060** |
| Vicuna-7b | 0.703±0.035 | **0.658±0.075** | 0.527±0.036 | 0.517±0.068 |
| Origin | **0.753** | 0.633 | 0.610 | 0.517 |




(a) The correlation trend on BoN. (b) The correlation trend on PPO.

Figure 4: The correlation between policy regret and accuracy on datasets constructed from responses. Each grid on the x and y axes represents a sampled bin, with the values below indicating the corresponding rank. For example, the 5 on the y-axis indicates that the chosen responses are sampled from those ranked between 5 and 10.

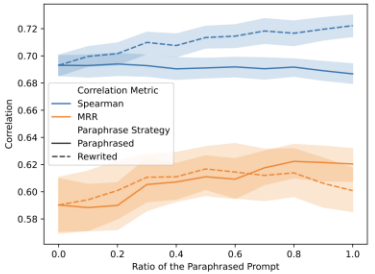
# RQ2 – How to better measure RM error

***Key Finding 3:*** **Prompt differences between RM test dataset and downstream test dataset weaken correlation**

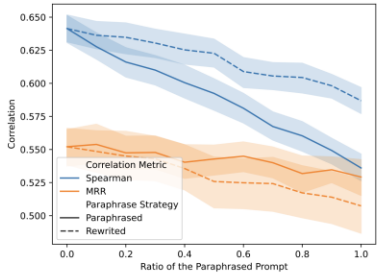
- Accuracy in each category aligns more closely with regret in that category for BoN.
- Paraphrasing prompts reduces correlation, particularly for PPO optimization

Table 3: The correlation between policy regret and accuracy on the test dataset composed of prompts from different categories evaluated by the Spearman coefficient. The highest result in each column is **bolded**, and the highest in each row is underlined.

| | Regret | Accuracy | | | | |
|---|---|---|---|---|---|---|
| | | Chat | ChatHard | Code | Math | Safety |
| BoN | Chat | **0.529±0.082** | 0.682±0.058 | 0.573±0.050 | 0.557±0.052 | 0.657±0.044 |
| | ChatHard | 0.493±0.089 | 0.682±0.053 | 0.583±0.038 | 0.538±0.066 | 0.655±0.051 |
| | Code | 0.504±0.095 | 0.634±0.059 | **0.717±0.043** | 0.456±0.060 | 0.646±0.053 |
| | Math | 0.288±0.121 | 0.343±0.080 | 0.244±0.048 | **0.610±0.058** | 0.282±0.059 |
| | Safety | 0.515±0.093 | **0.705±0.057** | 0.521±0.047 | 0.497±0.067 | **0.687±0.049** |
| PPO | Chat | 0.349±0.105 | 0.500±0.086 | 0.441±0.054 | 0.192±0.083 | **0.611±0.061** |
| | ChatHard | 0.314±0.115 | 0.484±0.083 | 0.434±0.057 | 0.202±0.087 | 0.576±0.073 |
| | Code | 0.384±0.092 | 0.450±0.080 | **0.527±0.051** | **0.371±0.060** | 0.459±0.053 |
| | Math | 0.185±0.091 | 0.275±0.064 | 0.238±0.050 | 0.145±0.057 | 0.313±0.054 |
| | Safety | **0.359±0.118** | **0.521±0.062** | 0.442±0.061 | 0.332±0.069 | 0.533±0.053 |




(a) The correlation trend on BoN.



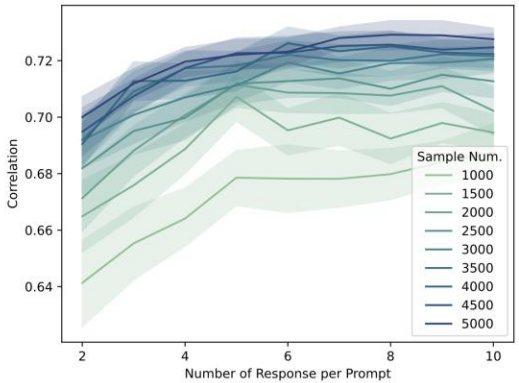

(b) The correlation trend on PPO.

Figure 5: The correlation between policy regret and accuracy on datasets with different ratios of paraphrased prompts evaluated by the Spearman coefficient and MRR. Different line styles are used to represent datasets paraphrased by different strategies.

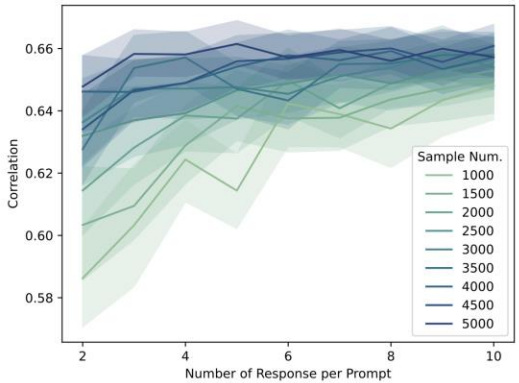
# RQ2 – How to better measure RM error

***Key Finding 4:*** **Increasing responses per prompt enhances correlation**

- More responses per prompt consistently achieves higher correlation
- With fixed budget, expanding responses is more effective than adding prompts




(a) The correlation trend on BoN.



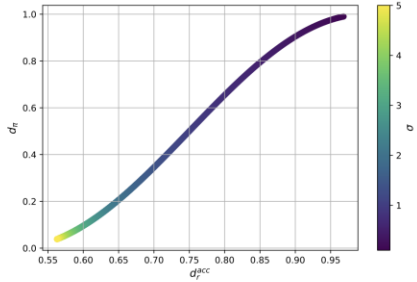

(b) The correlation trend on PPO.

Figure 6: The Spearman coefficient between accuracy and policy regret with test datasets having the same number of samples but varying numbers of responses per prompt.

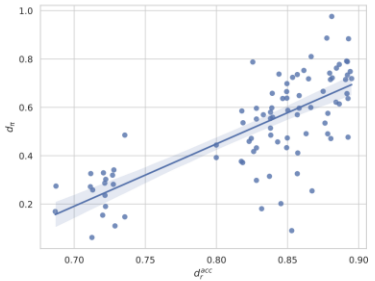
# RQ2 – How to better measure RM error

***Key Finding 5:*** **Accuracy alone can be insufficient to capture potential RM overoptimization**

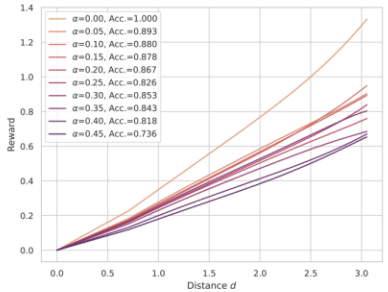
- RMs with similar accuracy can behave quite differently in terms of overoptimization
- Different Goodhart's effects beyond the Regressional type may be at play
- Suggests limitations of accuracy as a sole predictor



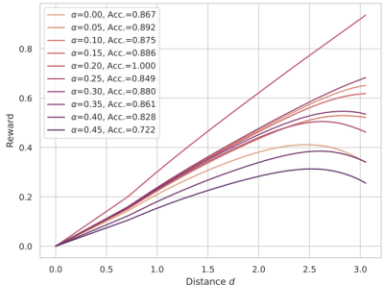

(a) Expected trend of accuracy $d_r^{acc}$ and the degree of overoptimization $d_\pi$ by varying the noise $\sigma$.




(b) The trend of the accuracy $d_r^{acc}$ and the estimated degree of overoptimization $d_\pi$ under BoN.




(a) RM with $\alpha = 0$ as golden reward model.




(b) RM with $\alpha = 0.15$ as golden reward model.

# *Conclusion*

**Accuracy is useful but limited:**

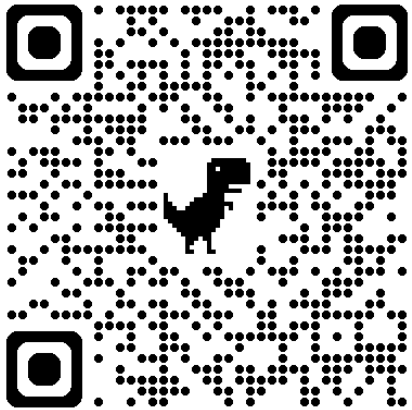
- Provides a basic signal but insufficient for full prediction

**Test dataset design matters**

- More responses per prompt are beneficial
- Prompt and response distributions affect correlation

**Beyond accuracy**

- Need to consider overoptimization patterns



↑ Scan to Read Paper

# THANKS

wenxueru2022@iscas.ac.cn