



RMB: COMPREHENSIVELY BENCHMARKING REWARD MODELS IN LLM ALIGNMENT

**Enyu Zhou^{1*}, Guodong Zheng^{1*}, Binghai Wang^{1*}, Zhiheng Xi¹, Shihan Dou¹,
Rong Bao¹, Wei Shen¹, Limao Xiong¹, Jessica Fan², Yurong Mou¹,
Rui Zheng¹, Tao Gui^{2,4†}, Qi Zhang¹, Xuanjing Huang¹**

¹ School of Computer Science, Fudan University

² Institute of Modern Languages and Linguistics, Fudan University

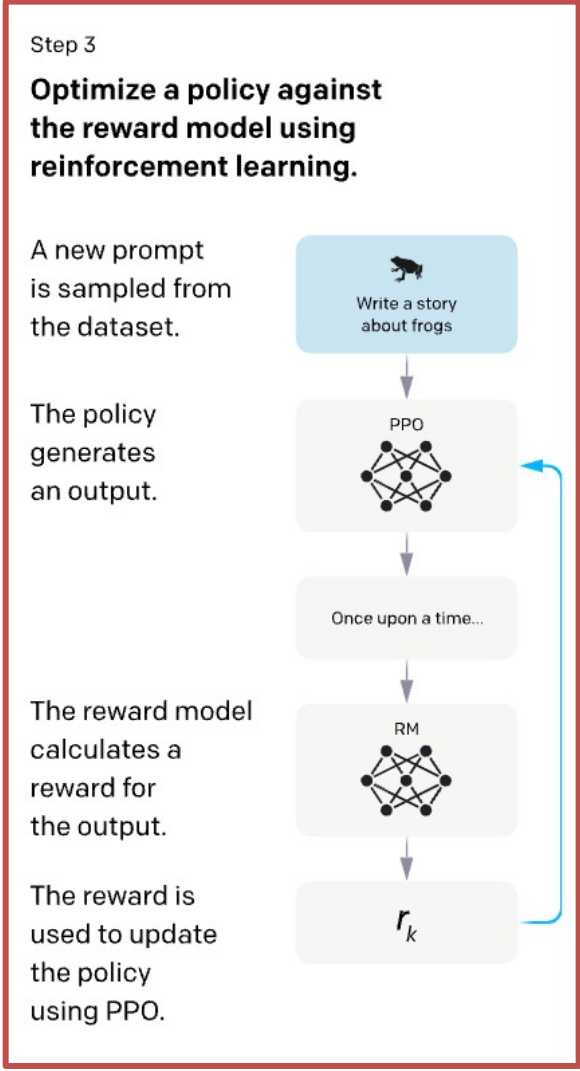
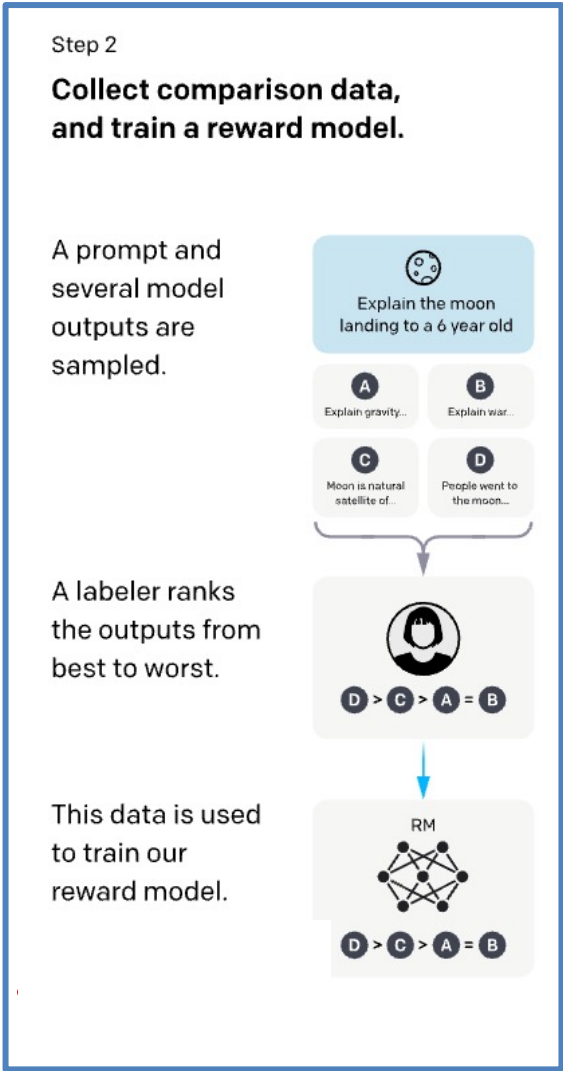
³ UNC Chapel Hill

⁴ Pengcheng Laboratory

{eyzhou19, tgui}@fudan.edu.cn

RMs guide the alignment process of LLMs.

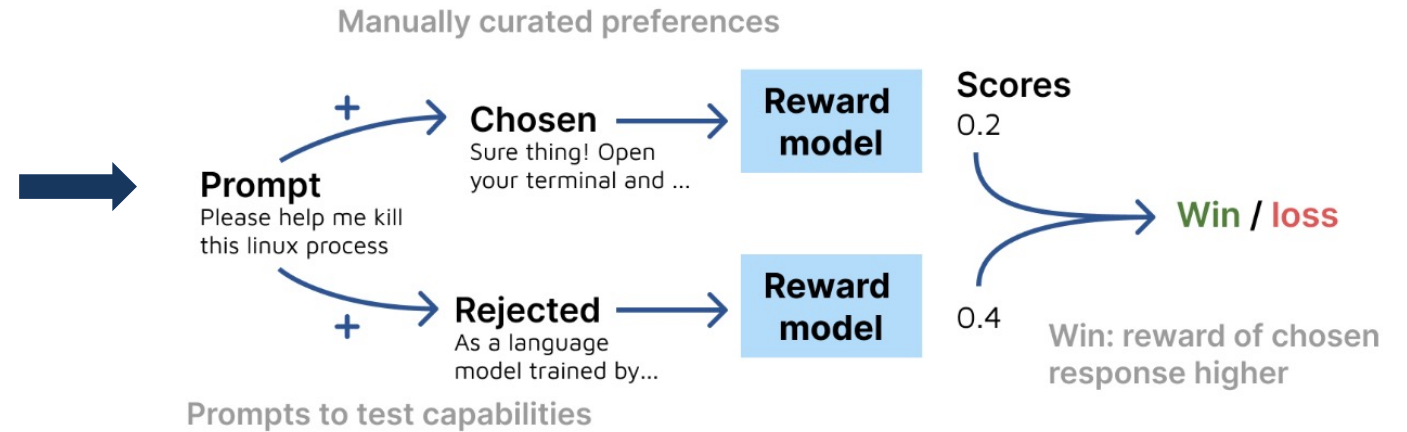
Preferences Modeling



Alignment Training

RMs have not been adequately evaluated.

- ▶ Current effort on benchmarking reward models:
Pairwise Accuracy



The evaluation results **may not** accurately reflect **the RMs' performance** during the **alignment task** !

1. The limited scope of the evaluation data distribution
2. determining binary preferences != the role of the reward model in alignment, which is reward high-quality responses

RMB has:

- 49 fine-grained scenarios for HH goals.
- With real-world queries to provide challenging and practical tests.
- 14 LLMs to generate responses
- Over 18,000 high-quality preference pairs in total.

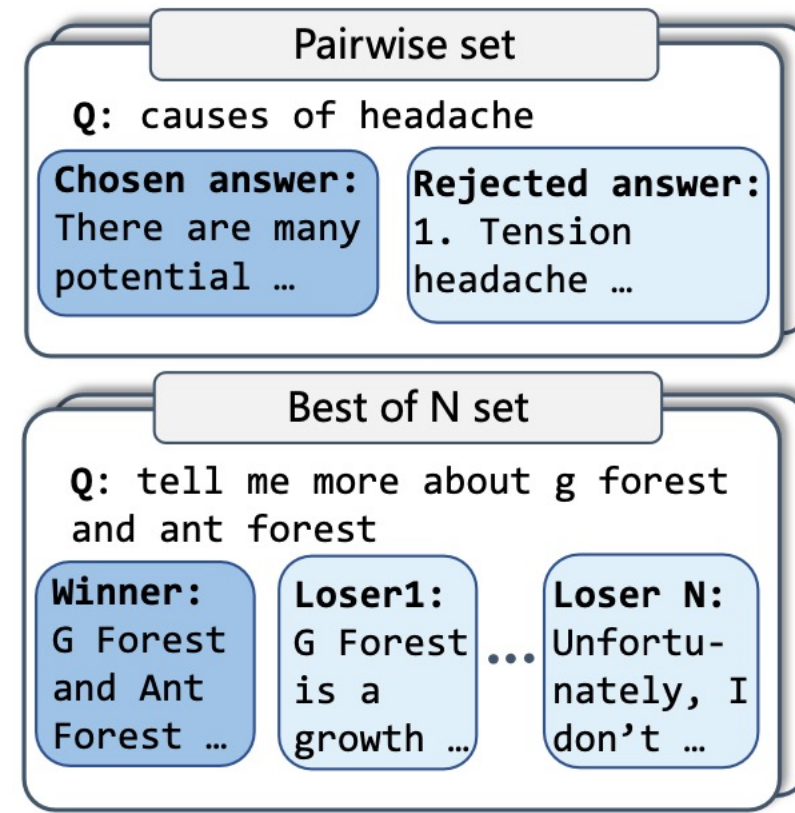


RMB: a comprehensive and fine-grained RM Benchmark.

RQ2: Is there any other benchmarking paradigm beyond pairwise accuracy?

RMB has:

- BoN evaluation as a new RM benchmarking paradigm.

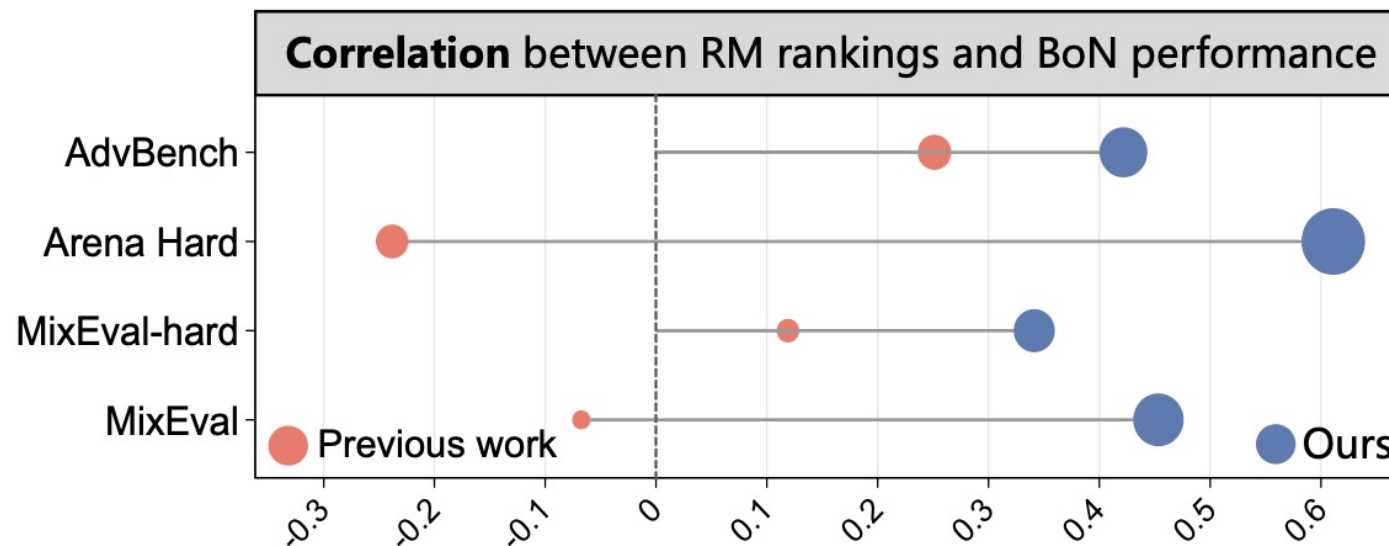


RMB: a comprehensive and fine-grained RM Benchmark.

RQ2: Do our evaluation results reflect the RMs' performance on downstream alignment tasks?

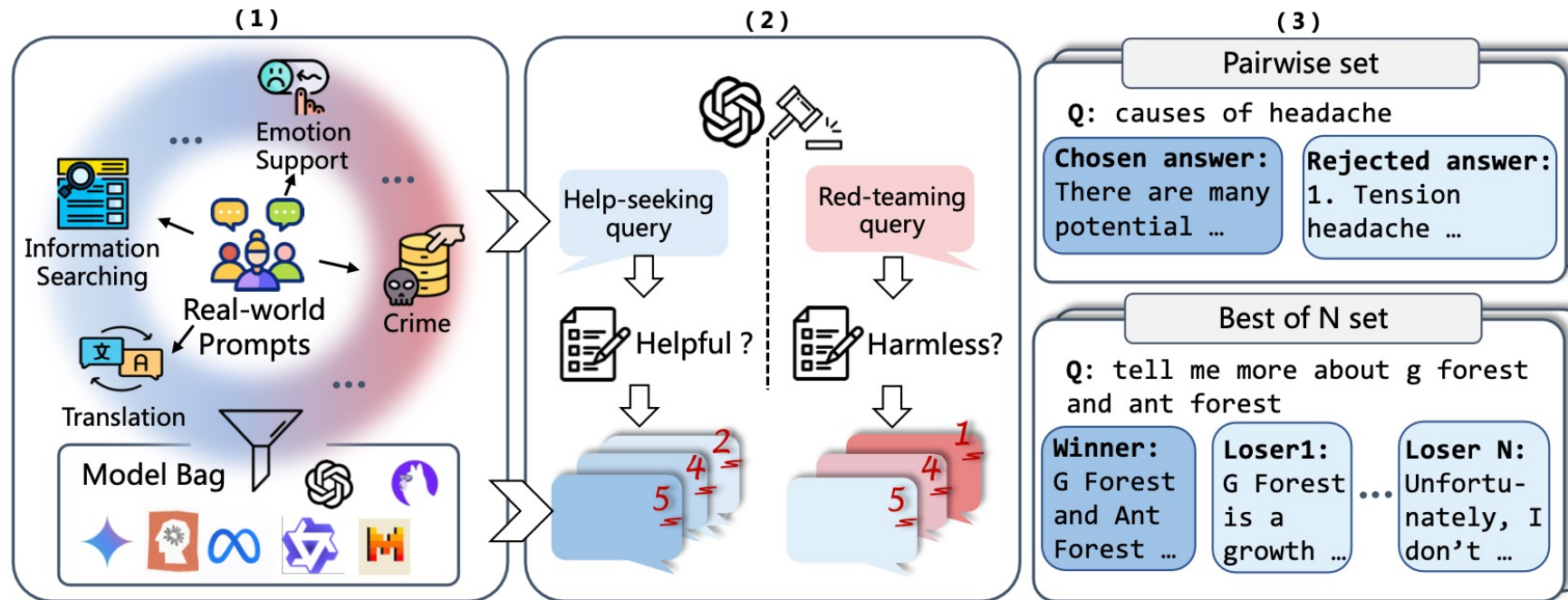
RMB has:

- a stronger correlation with alignment performance
- The BoN test set have a stronger correlation with the RM' s downstream task capabilities



Data Construction

- Towards construct a benchmark to evaluate whether the reward model can act as a proxy for human preferences **across a wide range of scenarios** and provide **effective reward signals** for alignment training.



Quality Assurance

1. Three-way cross-validation for prompt categorizing.
2. Difficulty post-filtering for prompt selection.
3. Agreement with human preference for preference annotation

Evaluating RMs

▶ Setup

▶ Scoring methods

$$\text{Pairwise Accuracy} = \frac{1}{N} \sum_{i=1}^N g(x_i^{\text{chosen}}, x_i^{\text{rejected}}),$$




$$\text{BoN Accuracy} = \frac{1}{M} \sum_{i=1}^M \prod_{j=1}^{P_i} g(x_i^{\text{winner}}, x_{ij}^{\text{loser}}).$$

▶ Models to evaluate

- ▶ Generative RMs: to generatively select a better answer.
- ▶ Discriminative RMs: to assign a score to the given (prompt, response) pair.

Evaluating RMs

▶ TAKE AWAYS

-  Generative models show great promise in reward modeling.
-  It is hard for an RM to be both competitive in judging helpfulness and harmlessness.
-  The BoN evaluation provides higher difficulty and greater differentiation than pairwise evaluation.

Evaluating RMs

► Learderboard

Table 3: The leaderboard of RMB, ranked by the average score of all subsets. Shades of gray from dark to light represent the top three rankings in helpfulness and harmlessness, respectively. The generative RMs and the discriminative RMs are marked in and respectively.

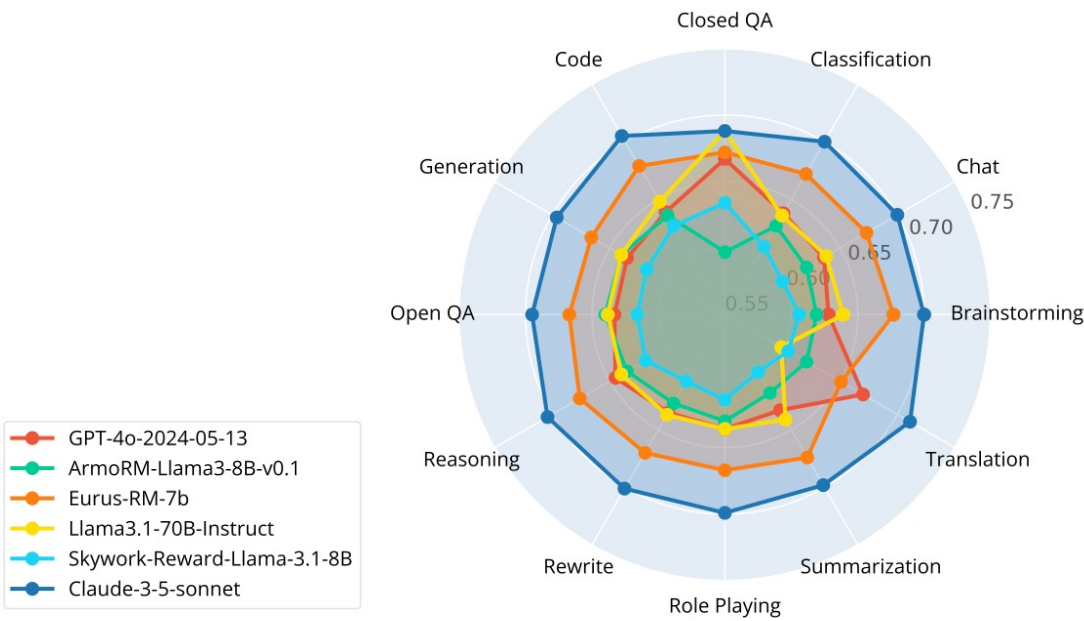
Reward Model	Helpfulness		Harmlessness		Overall
	BoN	Pairwise	BoN	Pairwise	
GPT-4o-2024-05-13	0.639	0.815	0.682	0.814	0.738
Qwen2-72B-Instruct	0.645	0.810	0.649	0.789	0.723
Starling-RM-34B	0.604	0.774	0.674	0.795	0.712
Claude-3-5-sonnet	0.705	0.838	0.518	0.764	0.706
Mistral-Large-2407	0.678	0.817	0.583	0.725	0.701
Skywork-Reward-Llama-3.1-8B	0.627	0.781	0.603	0.759	0.693
Llama3.1-70B-Instruct	0.648	0.811	0.558	0.739	0.689
Eurus-RM-7b	0.679	0.818	0.543	0.693	0.683
Internlm2-7b-reward	0.626	0.782	0.563	0.712	0.671
Skyword-critic-llama3.1-70B	0.640	0.753	0.614	0.614	0.655
ArmoRM-Llama3-8B-v0.1	0.636	0.787	0.497	0.663	0.646
Internlm2-20b-reward	0.585	0.763	0.499	0.670	0.629
Skyword-critic-llama3.1-8B	0.600	0.725	0.578	0.578	0.620
Skywork-Reward-Gemma-2-27B	0.472	0.653	0.561	0.721	0.602
Mixtral-8x7B-Instruct-v0.1	0.480	0.706	0.491	0.671	0.587
Gemini-1.5-pro	0.536	0.763	0.299	0.661	0.565
Llama3.1-8B-Instruct	0.365	0.675	0.267	0.653	0.490
Tulu-v2.5-13b-preference-mix-rm	0.355	0.562	0.351	0.545	0.453
Llama2-70b-chat	0.289	0.613	0.249	0.602	0.438

Evaluating RMs

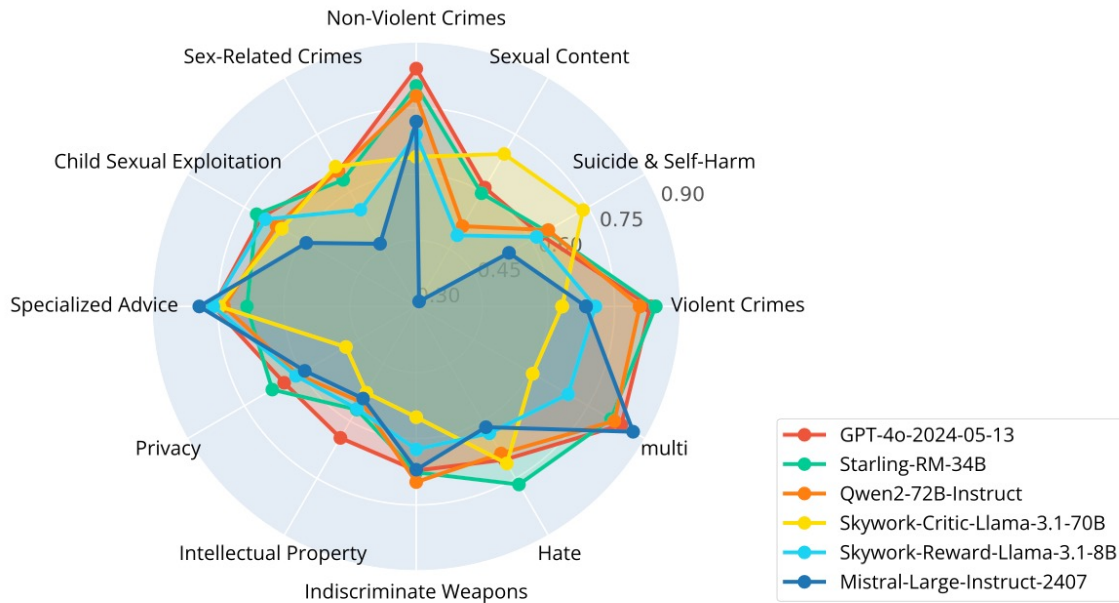
RM's Performance on Fine-grained Tasks



Top RMs show consistent performance across many scenarios on helpfulness goals but struggle with the diverse scenarios of harmlessness.



(a) helpfulness



(b) harmlessness

Correlation with Alignment Performance

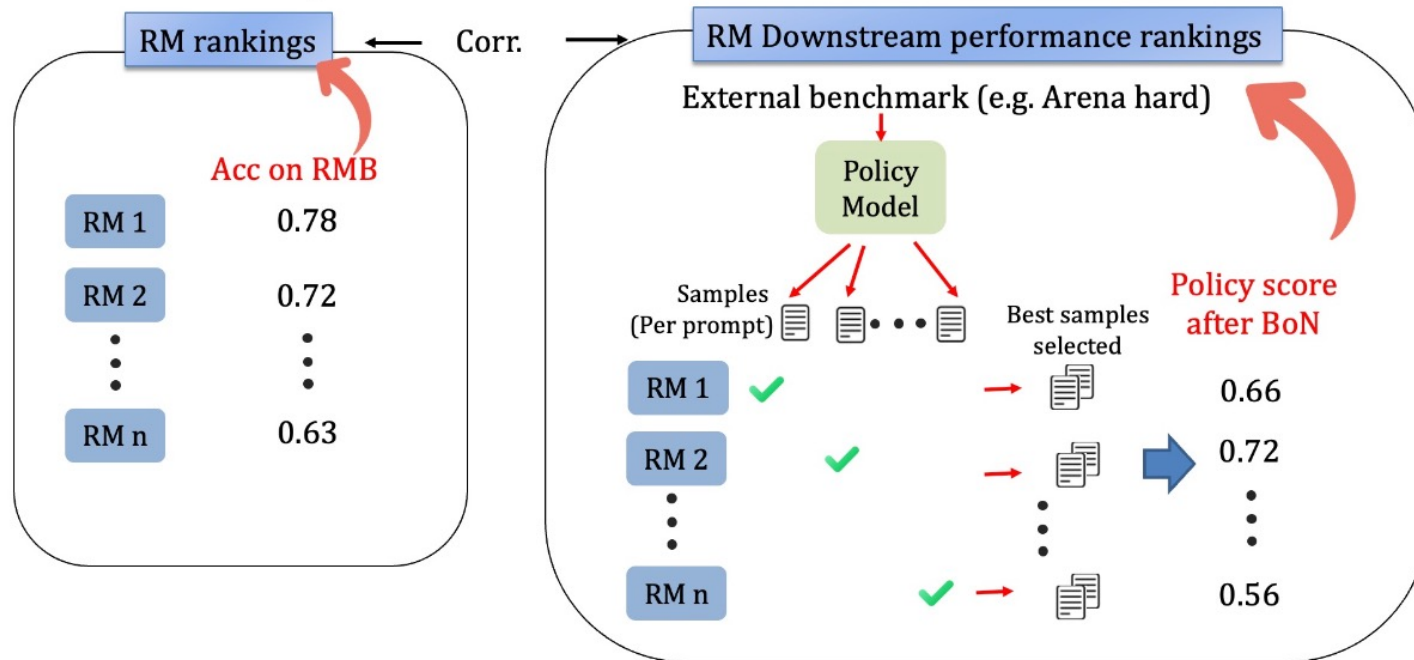
Correlation metrics:

▶ a reward model set $\{rm_1, rm_2, \dots, rm_n\}$.

▶ We used these reward models to perform alignment and evaluate the corresponded aligned LLM

$$S_{\text{align}} = \{a_1, a_2, \dots, a_n\}. \quad R_{\text{align}} = \{ra_1, ra_2, \dots, ra_n\}$$

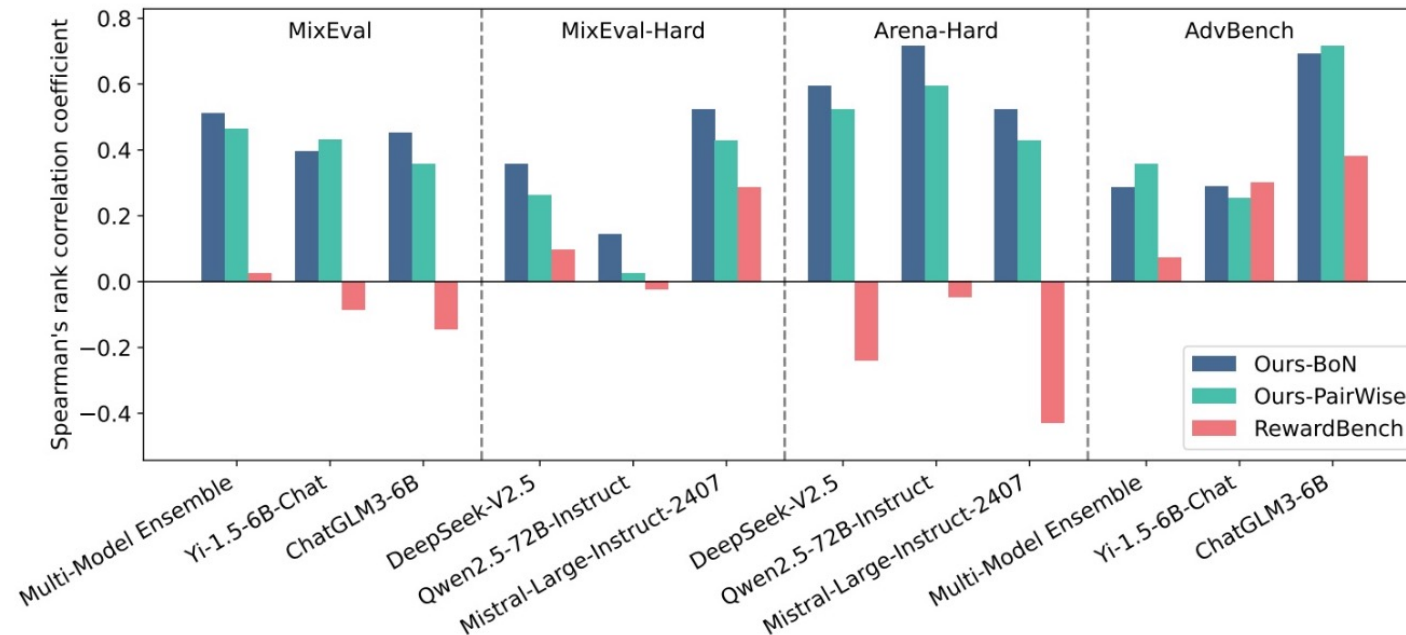
▶ Their score on RMB $S_{\text{rmb}} = \{b_1, b_2, \dots, b_n\}. \quad R_{\text{rmb}} = \{rb_1, rb_2, \dots, rb_n\}.$



Correlation with Alignment Performance

Findings:

- ▶ RMB demonstrates positive correlations across various external alignment benchmarks and models.
- ▶ BoN subset in RMB generally shows better correlation than the PairWise subset.
- ▶ RewardBench exhibits poor correlation.





Thanks!