Reward Reasoning Model

Jiaxin Guo*[12], Zewen Chi*[1], Li Dong*[1]
Qingxiu Dong[13], Xun Wu[1], Shaohan Huang[1], Furu Wei[1]⋄
[1] Microsoft Research [2] Tsinghua University [3] Peking University
https://aka.ms/GeneralAI

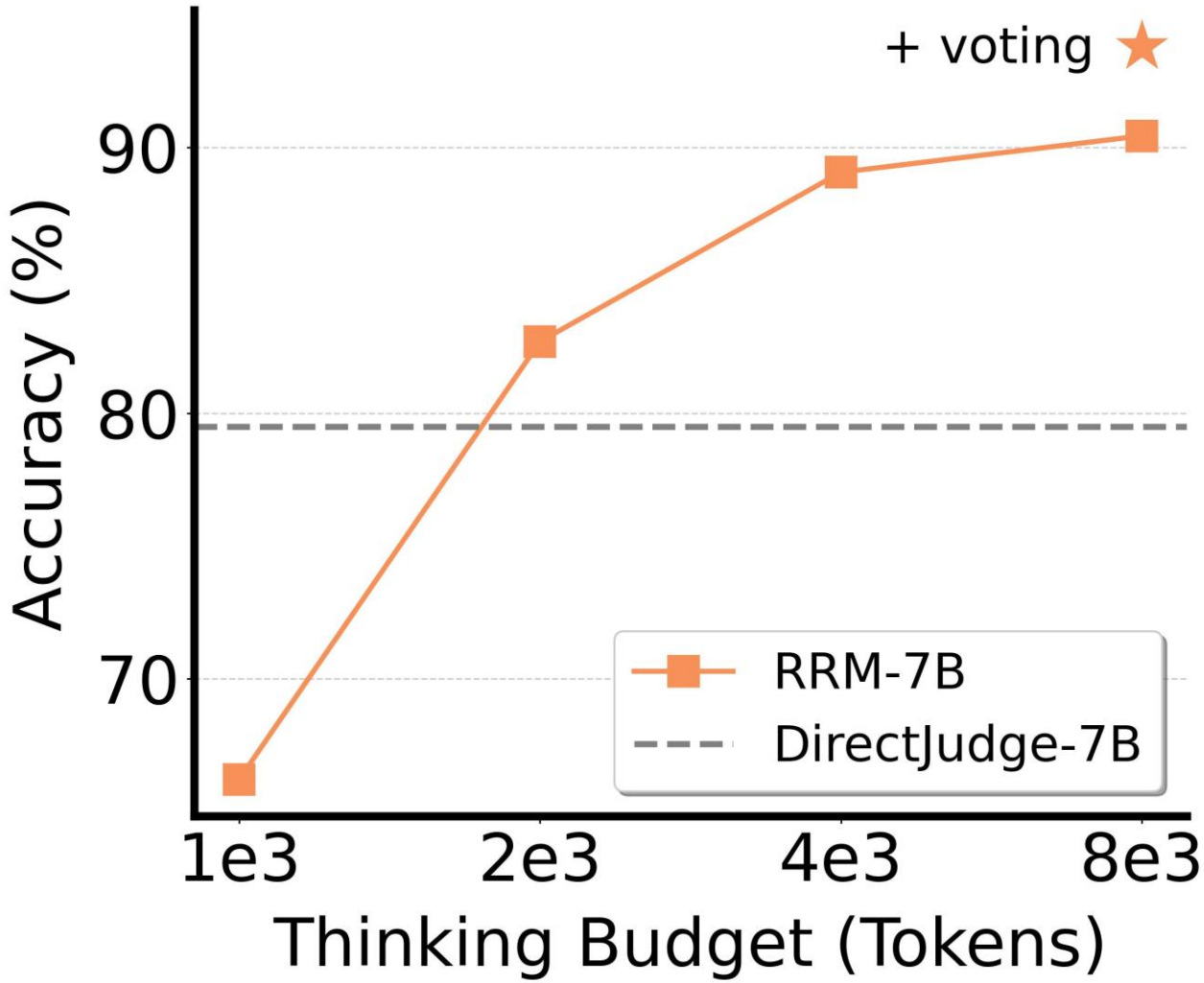
Presenter: Jiaxin Guo
Tsinghua University
https://xinyuerufei.github.io

1. Background and Motivation

Challenge: How can a reward model allocate more compute to harder tasks?

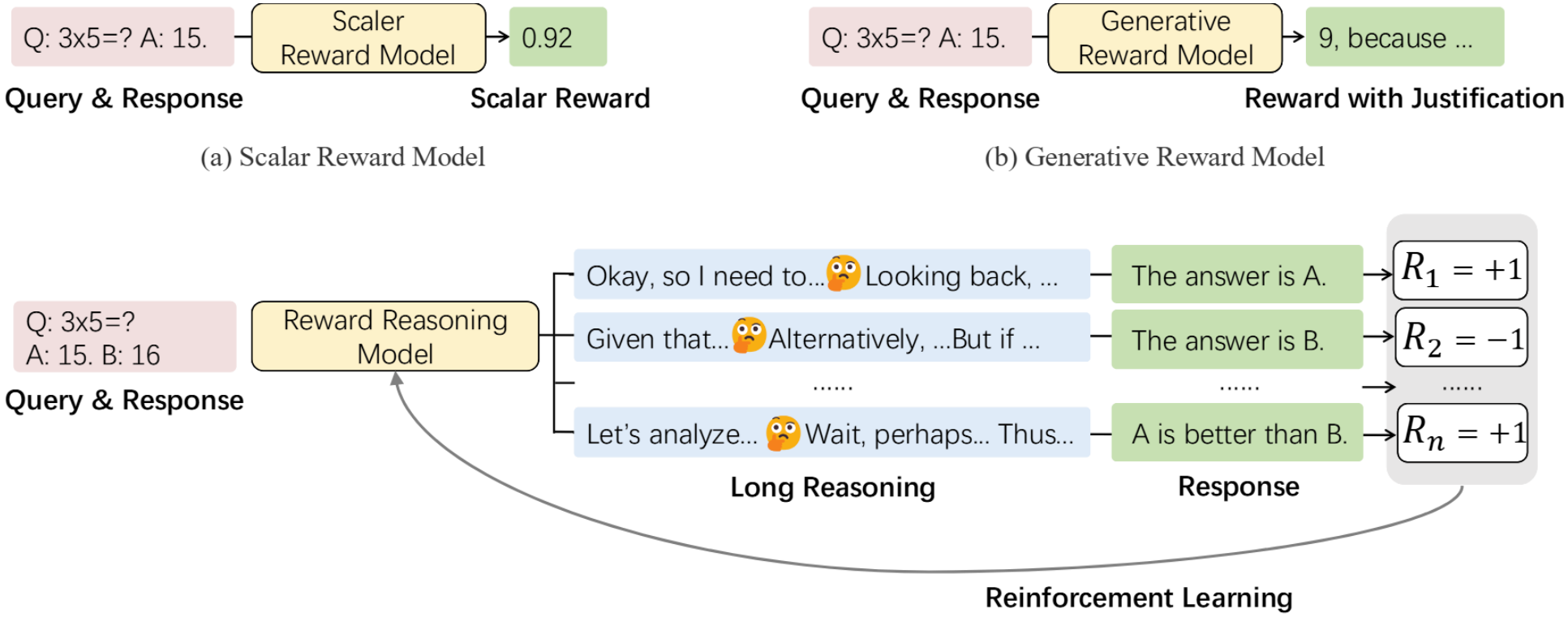
- Current reward models **lack adaptive compute scaling**, using nearly uniform computation for all inputs.

- This inflexibility **limits their reasoning capability** on complex or multi-step problems.



2. Methods

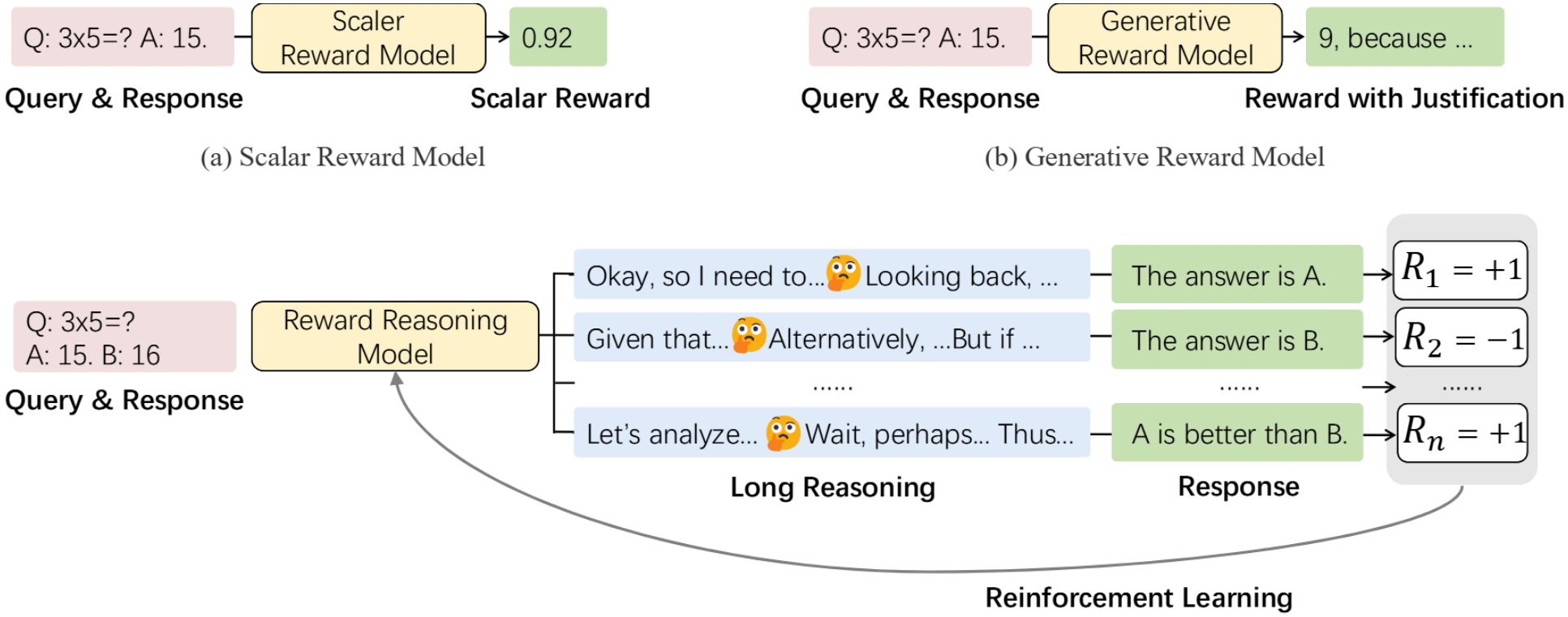
- Input: (Query, Response 1, Response 2)
- Output: Autoregressive text containing
  → a reasoning process
  → a final decision in the format \boxed{Assistant 1} or \boxed{Assistant 2}



2. Methods

Training via Reinforcement Learning

$$R = \begin{cases} +1, & \textit{if RRM selects correct response} \\ -1, & \textit{otherwise} \end{cases}$$



3. Results and Analysis

Results on Reward Benchmarks

Table 1: Evaluation results on RewardBench benchmark and PandaLM Test. **Bold** numbers indicate the best performance, Underlined numbers indicate the second best.

Models	RewardBench					PandaLM Test	
	Chat	Chat Hard	Safety	Reasoning	Overall	Agreement	F1
Skywork-Reward-Gemma-2-27B-v0.2 [40]	96.1	89.9	**93.0**	98.1	**94.3**	76.6	76.4
JudgeLM-7B [82]	87.3	43.6	74.5	48.7	63.5	65.1	61.9
JudgeLM-33B [82]	92.7	54.2	85.8	58.3	72.3	75.2	69.7
Claude-3.5-Sonnet-20240620 [35]	96.4	74.0	81.6	84.7	84.2	-	-
DeepSeek-R1 [42, 12]	**97.1**	73.7	73.3	95.6	84.9	78.7	72.5
DeepSeek-GRM-27B [42]	94.1	78.3	88.0	83.8	86.0	-	-
GPT-4-0125-preview [35]	95.3	74.3	87.6	86.9	86.0	66.5	61.8
GPT-4o-0806 [35]	96.1	76.1	86.6	88.1	86.7	-	-
RM-R1-DeepSeek-Distilled-Qwen-7B [14]	88.9	66.2	78.4	87.0	80.1	-	-
RM-R1-DeepSeek-Distilled-Qwen-14B [14]	91.3	**91.3**	79.4	95.5	88.9	-	-
RM-R1-DeepSeek-Distilled-Qwen-32B [14]	95.3	80.3	91.1	96.8	90.9	-	-
DirectJudge-7B	86.0	69.7	85.5	79.5	80.2	70.3	70.2
DirectJudge-32B	96.1	85.1	89.5	90.9	90.4	76.7	77.4
RRM-7B	87.7	70.4	80.7	90.0	82.2	72.9	71.1
RRM-7B (voting@16)	92.1	71.5	81.3	93.8	84.8	75.9	77.8
RRM-32B	94.7	81.1	90.7	98.3	91.2	78.8	79.0
RRM-32B (voting@16)	96.1	81.4	91.6	**98.6**	91.9	**80.2**	**81.9**

3. Results and Analysis

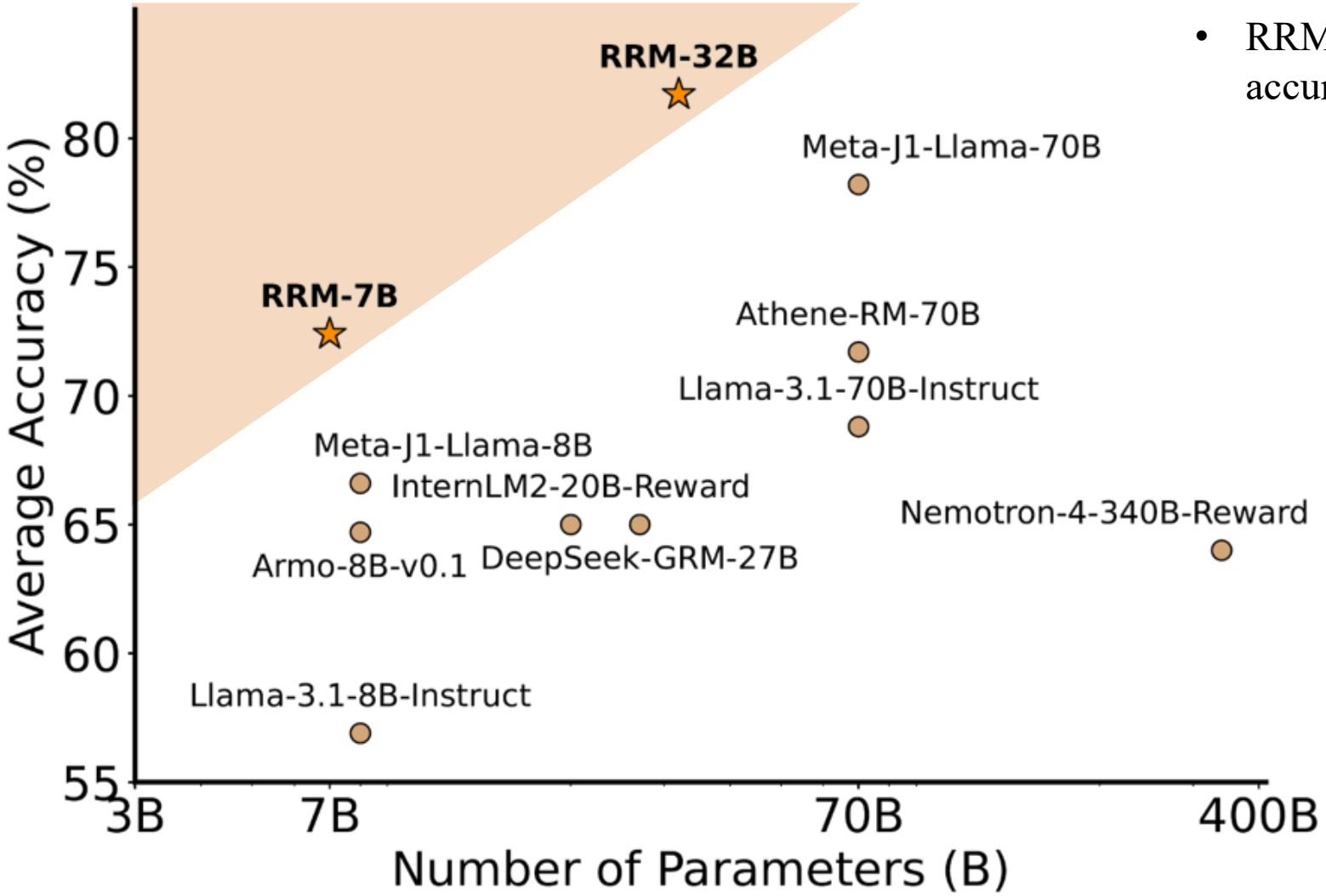
Results on Reward Benchmarks

Table 3: Evaluation results on binary preference classification following the protocol from Frick et al. [18]. For each benchmark, we report accuracy over a single random permutation of paired responses.

Models	MMLU-Pro	MATH	GPQA	Overall
Skywork-Reward-Gemma-2-27B [67]	55.0	46.2	44.7	48.6
Gemma-2-27B [42]	66.2	66.4	51.9	61.5
DeepSeek-GRM-27B (voting@32) [42]	65.5	69.4	56.0	63.6
DeepSeek-GRM-27B (MetaRM) (voting@32) [42]	68.1	70.0	56.9	65.0
Llama-3.1-8B-Instruct [67]	56.3	62.9	51.4	56.9
Llama-3.1-70B-Instruct [67]	72.1	73.1	61.2	68.8
J1-Llama-8B (SC@32) [67]	67.5	76.6	55.7	66.7
J1-Llama-70B (SC@32) [67]	79.9	88.1	66.5	78.2
RRM-7B	66.5	88.0	57.9	70.3
RRM-7B (voting@5)	68.3	90.5	58.3	72.4
RRM-32B	80.5	94.3	67.4	80.7
RRM-32B (voting@5)	**81.3**	**95.4**	**68.4**	**81.7**

3. Results and Analysis

Frontier Reward Modeling Performance

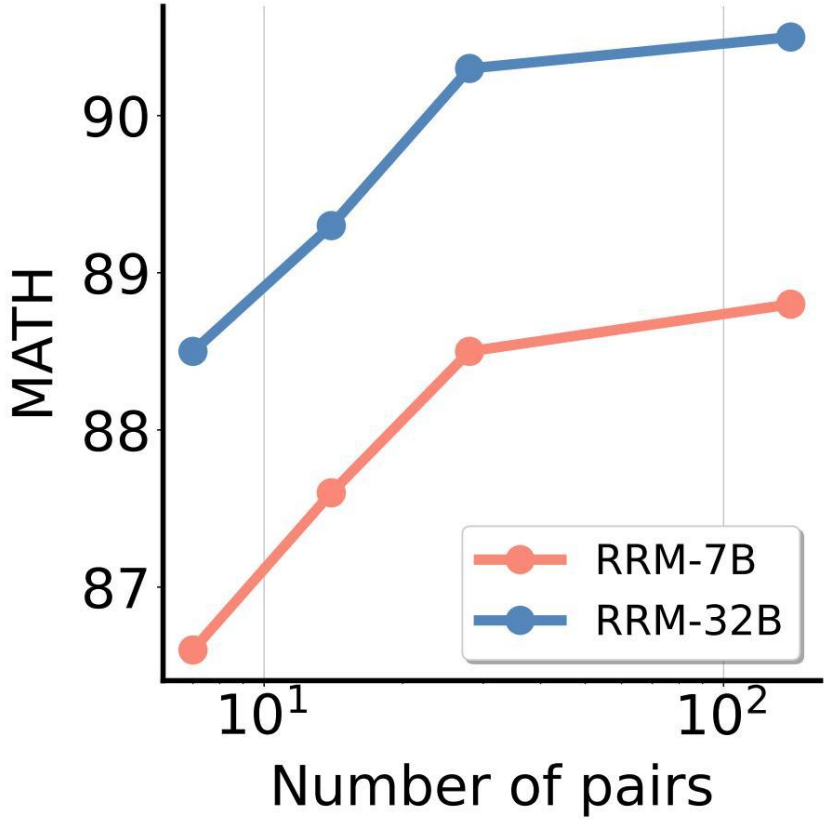


- RRMs dominate the top-left region, showing higher accuracy with much smaller model sizes.

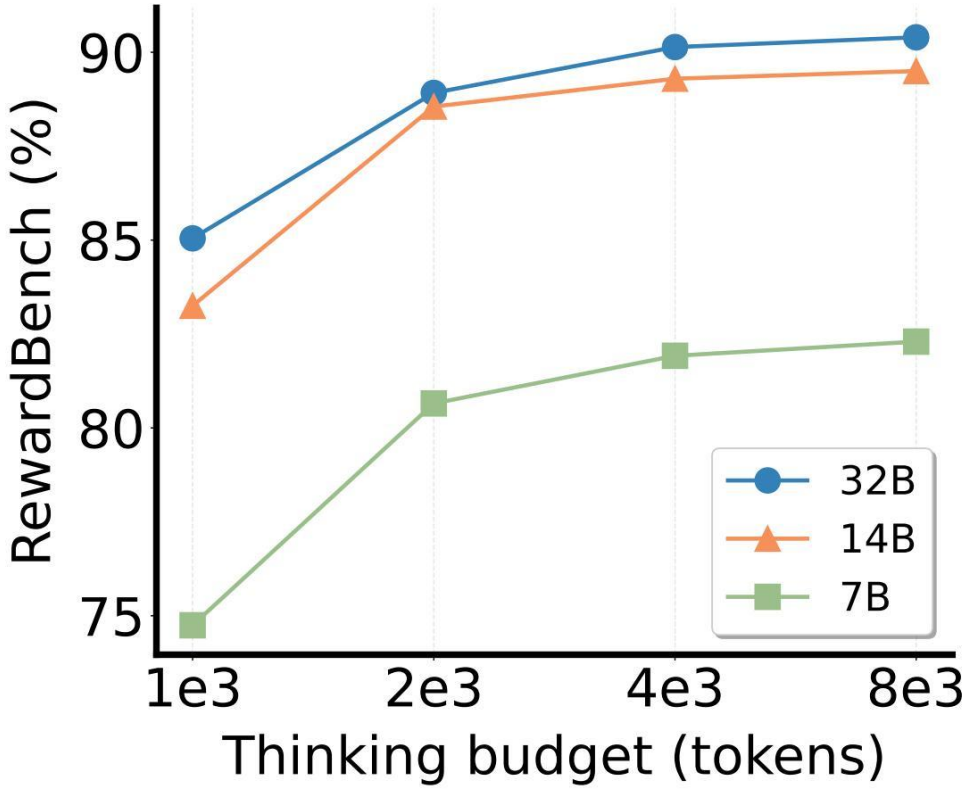
3. Results and Analysis

Test-Time Scaling (TTS)

- Parallel TTS: RRMs effectively utilize increased computational budgets—more comparisons yield consistent gains.



- Sequential TTS: RRMs achieve higher accuracy with longer reasoning horizons, effectively leveraging extended token budgets for better rewards.



3. Results and Analysis

Adaptive Compute Allocation

- RRMs dynamically adjust reasoning length based on task complexity. Table shows average number of tokens before the </think> token on RewardBench subtypes.)

Task Subtype	Average Thinking Length (Tokens)
Chat	421.97
Chat Hard	363.71
Safety	319.62
Reasoning	**848.13**

3. Results and Analysis

Direct Preference Optimization with RRM Annotators

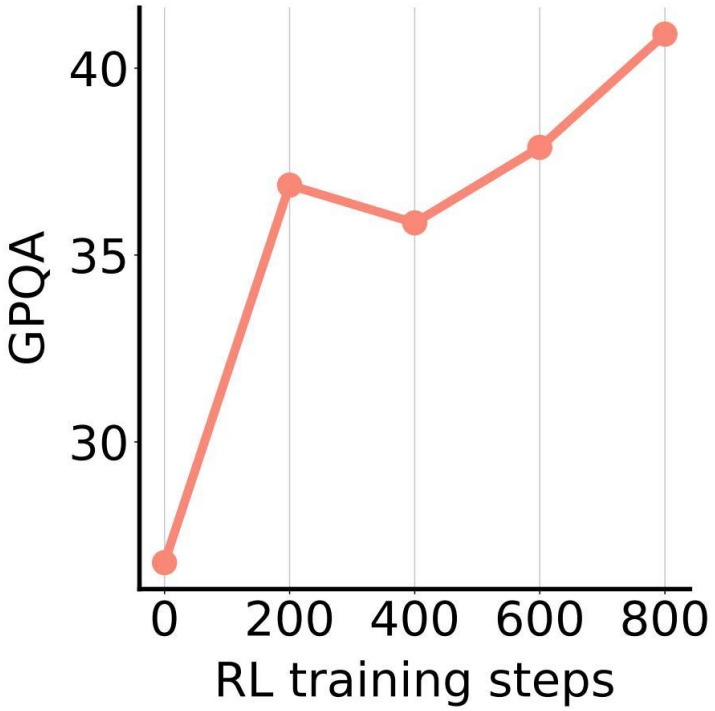
- RRMs can serve as high-quality preference annotators for post-training. Table shows performance of DPO post-trained Qwen2.5-7B models on Arena-Hard.

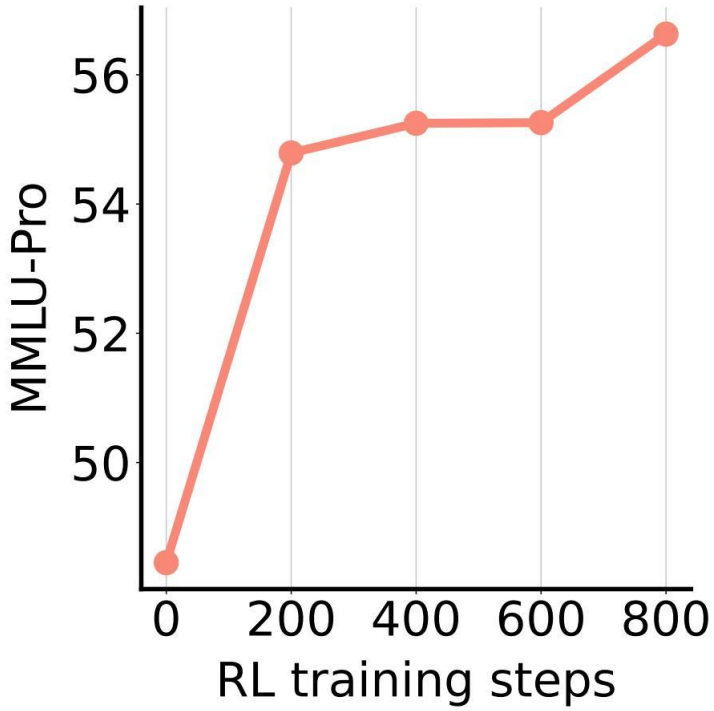
Model	Arena-Hard Score	95% CI
Before Post-Training		
Base Model	18.3	(-1.61, +1.66)
After DPO Post-Training (Preference Annotators)		
GPT-4o	51.9	(-2.96, +2.93)
RRM-7B	53.8	(-1.72, +1.85)
RRM-32B	**55.4**	(-2.60, +2.67)

3. Results and Analysis

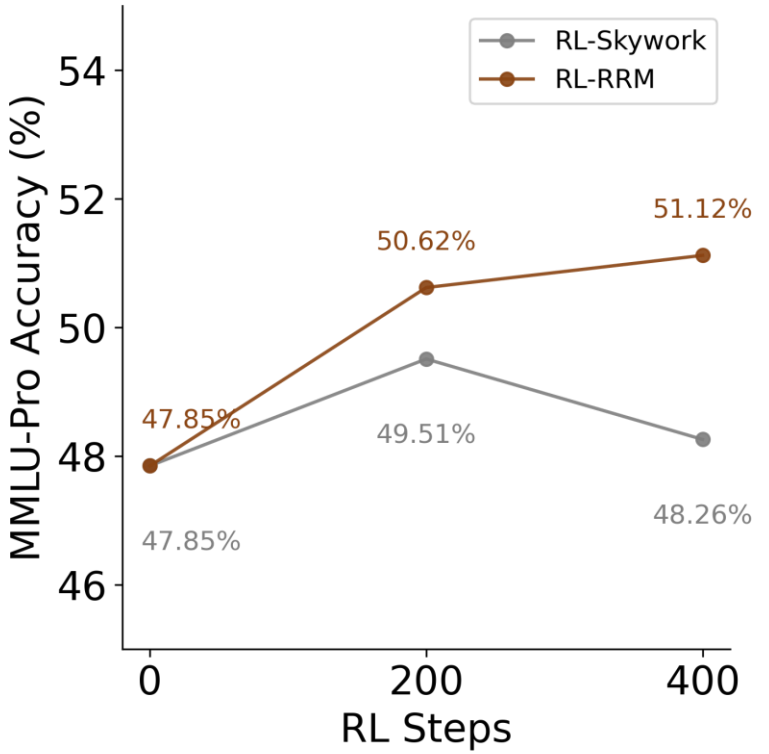
RL with RRM Rewards

- RRM-guided RL on **unlabeled** WebInstruct data steadily improves model performance throughout training **in general domain**.





- Compared with scalar-RM baselines (Skywork), RRM-guided RL achieves higher and more stable gains.



清華大學
Tsinghua University

Thanks for watching!