Fixing the Broken Compass: Diagnosing and Improving Inference-Time Reward Modeling

Jiachun Li[1,2], Pengfei Cao[1,2], Chenhao Wang[1,2], Zhuoran Jin[1,2], Yubo Chen[1,2], Kang Liu[1,2], Jun Zhao[1,2]

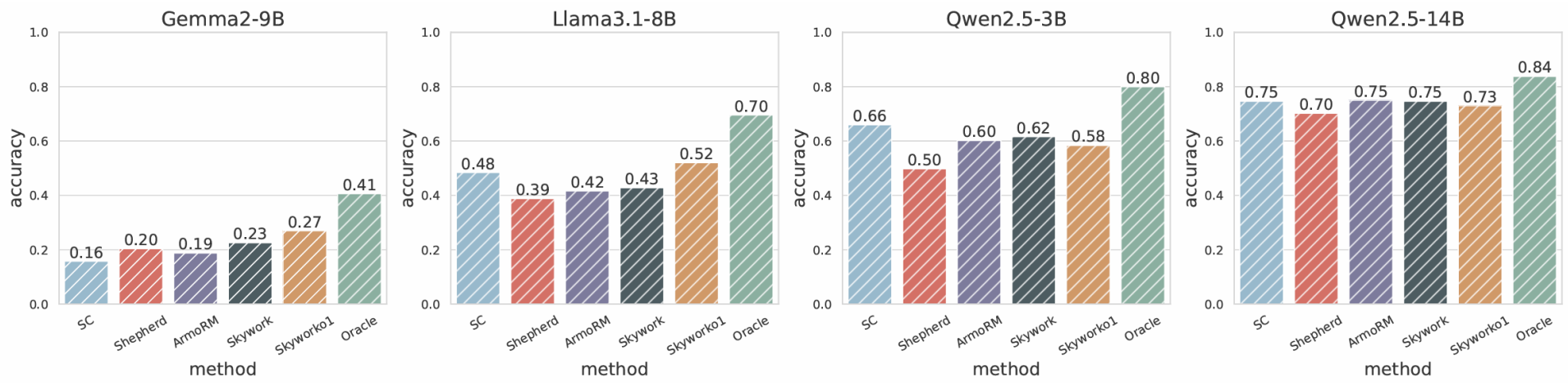
[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] National Laboratory of Pattern Recognition, Institute of Automation, CAS

中国科学院 CHINESE ACADEMY OF SCIENCES

CASIA 中国科学院自动化研究所 INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES

Motivation

- **Inference-time scaling paradigm is powerful**
  - Existing work focuses on **training-time optimization** (e.g. GRPO)
  - Reward model at inference is **underexplored & problematic** (e.g. MCTS)



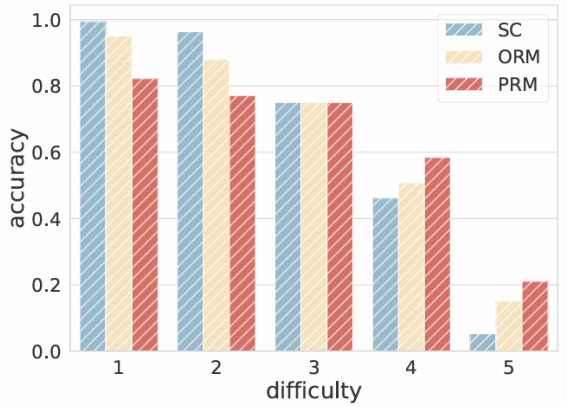
RM-based BoN sometimes performs worth than Self-Consitency

Problem: Reward Model is Broken

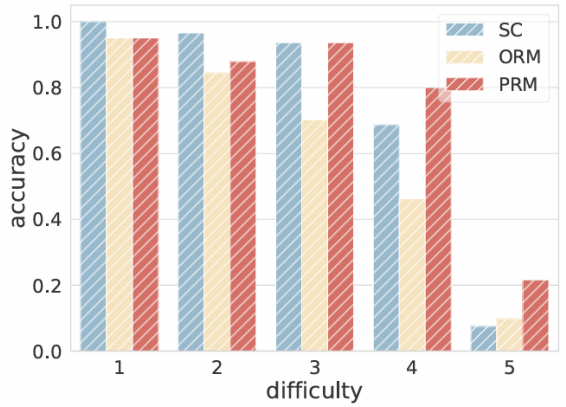
- **What's wrong with Reward Models?**
  - **Worse on easy questions**
    - We find that the introduction of the RM can hinder the LLM's reasoning performance on simple problems
  - **Worse with more rollouts**
    - We find that RMs struggle to correctly score incorrect responses with low occurrence frequencies, making it difficult to distinguish incorrect responses from correct ones as n grows
  - **Sensitive to high diversity**
    - We find that during inference, it is essential to constrain the diversity of the sampling distribution to maintain the optimal performance of the RM

Finding 1 (Question Difficulty)

- **RM hurts simple questions**
  - BoN & MCTS < SC on easy questions
  - RM introduces noise on these questions

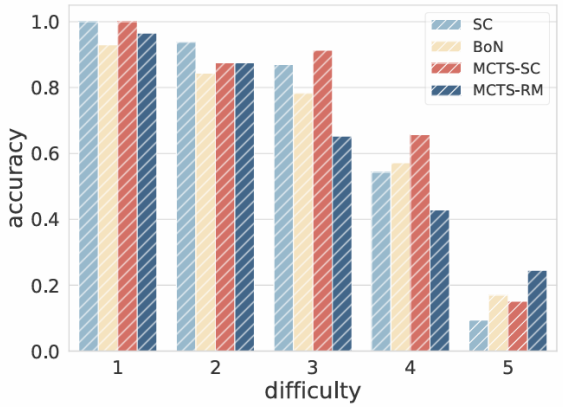


(a) Qwen2.5-3B

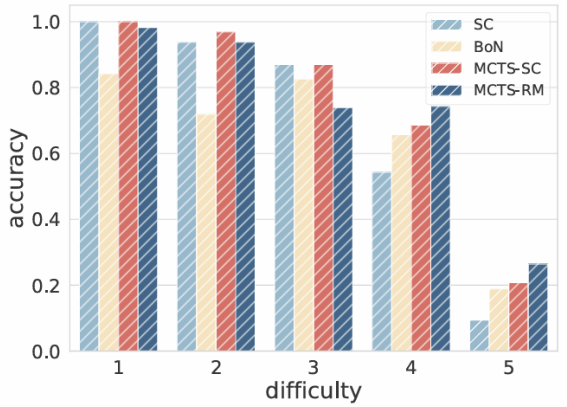


(b) Llama3.1-8B

Figure 2: Performance of BoN inference across different question difficulty levels.



(a) ORM



(b) PRM

Figure 3: Performance of MCTS inference across different question difficulty levels.

Finding 2 (Sampling Number)

RM faces reverse long-tail issue

- Rare wrong answers tends to get high scores (Figure 5)
- With n increases, RM makes more incorrect selection compared to SC (Figure 4)

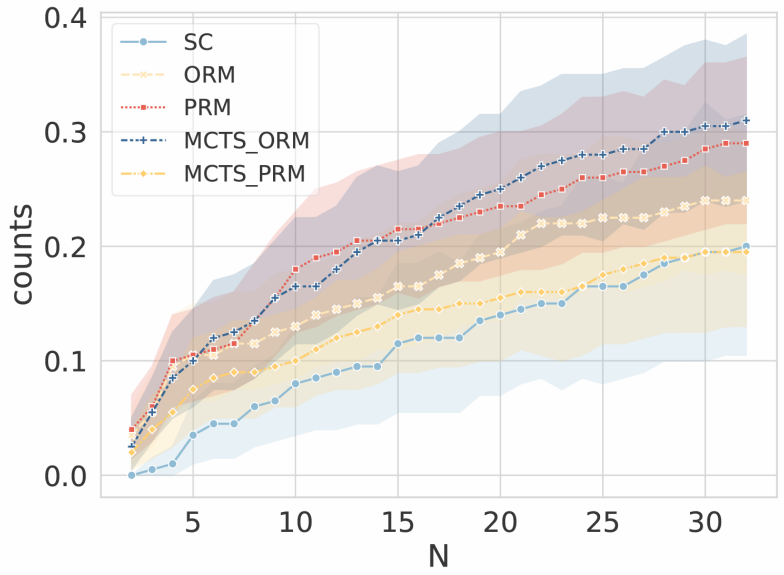


Figure 4: The number of times the model's selection changes from correct to incorrect.

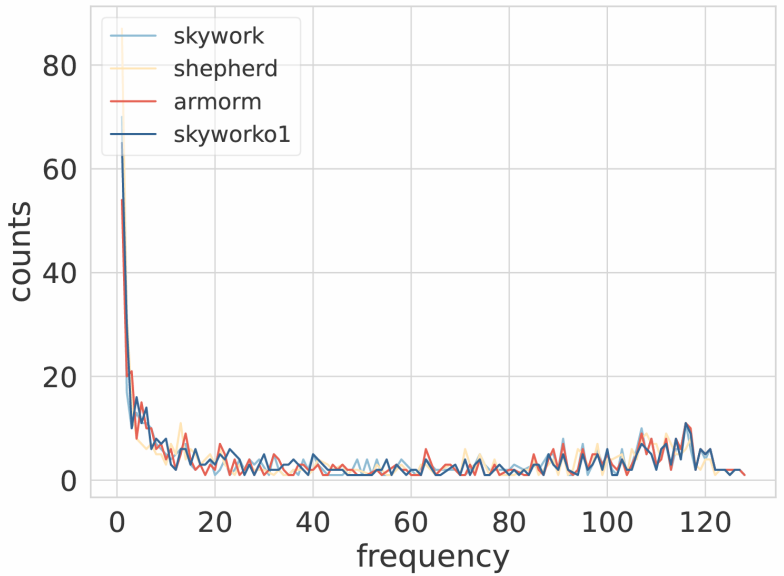


Figure 5: Frequency statistics of the highest-scored negative responses in BoN.

Finding 3 (Sampling Diversity)

- **RM can not distinguish high diversity rollouts**

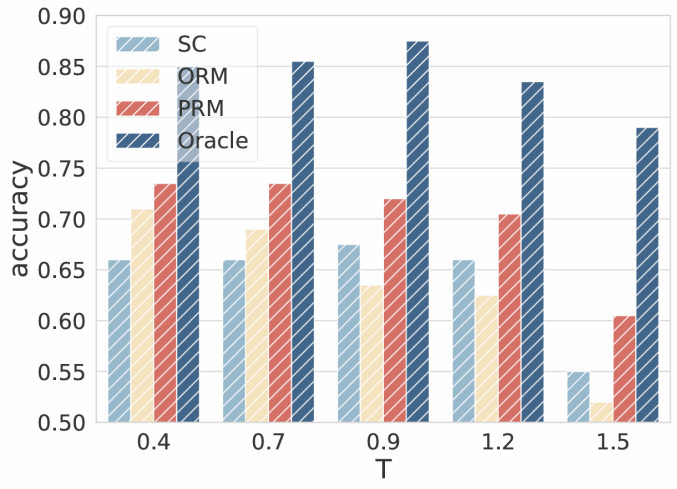
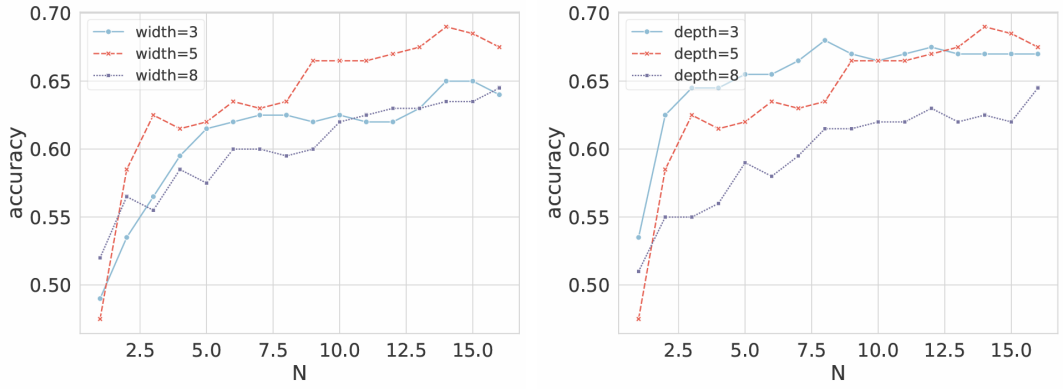


Figure 6: BoN performance across different temperatures (Qwen2.5-3B).



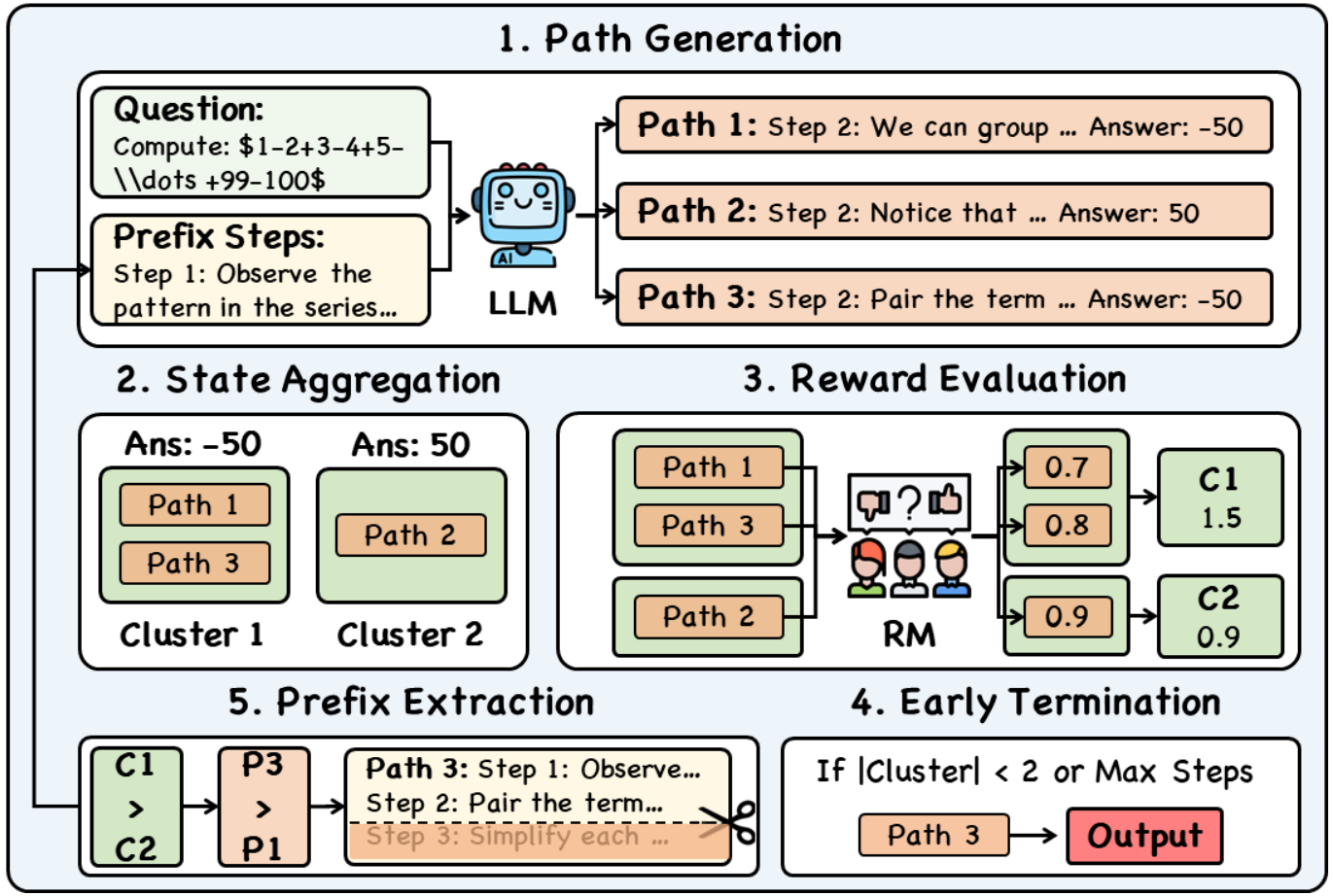
(a) Tree width (b) Tree depth

Figure 7: MCTS performance under different tree structures (ORM).

Our Method: CRISP

Main Process of Our Method

- Generate paths
- Cluster by answer
- Aggregate rewards
- Early stop
- Prefix update



Main Results

Overall Results

- +5% over RM methods
- +10% over R1 (non-math)

Methods		Qwen2.5-3B			Llama3.1-8B		
		GSM8K	*MATH*	*Olympiad*	*GSM8K*	*MATH*	*Olympiad*
CoT		0.78	0.46	0.24	0.85	0.38	0.11
Self-Consistency		0.83	0.64	0.31	0.91	0.57	0.16
Best-of-N	+ ORM	0.83	0.65	0.31	0.91	0.47	0.18
	+ PRM	0.87	0.61	0.34	0.95	0.62	0.23
BoN Weighted	+ ORM	0.83	0.67	0.31	0.89	0.53	0.20
	+ PRM	0.86	0.60	0.36	0.94	0.62	0.24
MCTS	+ ORM	0.92	0.67	0.34	0.90	0.43	0.13
	+ PRM	0.95	0.71	0.31	0.95	0.57	0.19
Beam Search		0.95	0.73	0.34	0.94	0.56	0.15
Ours	+ ORM	0.91	0.70	0.36	0.89	0.49	0.18
	+ PRM	**0.96**	**0.76**	**0.39**	**0.95**	**0.67**	**0.26**

Base Models	Methods	*Math*		*Commonsense*		*Social*		*Logical*	
		Acc	Length	Acc	Length	Acc	Length	Acc	Length
Qwen2.5-Math-1.5B	Chat	0.52	1470	0.40	1400	0.46	1204	0.40	2790
	R1-Distill	**0.79**	13421	0.47	6066	0.52	6407	0.35	12352
	Ours	0.59	943	**0.58**	1004	**0.61**	1144	**0.44**	1143
Qwen2.5-Math-7B	Chat	0.74	1855	0.58	1479	0.58	1388	0.49	2133
	R1-Distill	**0.88**	9626	0.65	3612	0.66	2920	0.50	6492
	Ours	0.79	987	**0.72**	1100	**0.66**	1059	**0.59**	2058

Thanks