Multilingual Rubric-Agnostic Reward Reasoning Models

David Anugraha[1], Shou-Yi Hung[2], Zilu Tang[3], Annie En-Shiun Lee[2,4], Derry Tanti Wijaya[3,5], Genta Indra Winata[6]

[1]Stanford University, [2]University of Toronto, [3]Boston University,
[4]Ontario Tech University, [5]Monash University Indonesia, [6]Capital One

Contact: davidanu@stanford.edu, genta.winata@capitalone.com

Overview

Introducing **mR3**!

A **multilingual rubric-agnostic reasoning reward model** designed for consistent and interpretable multilingual evaluation.

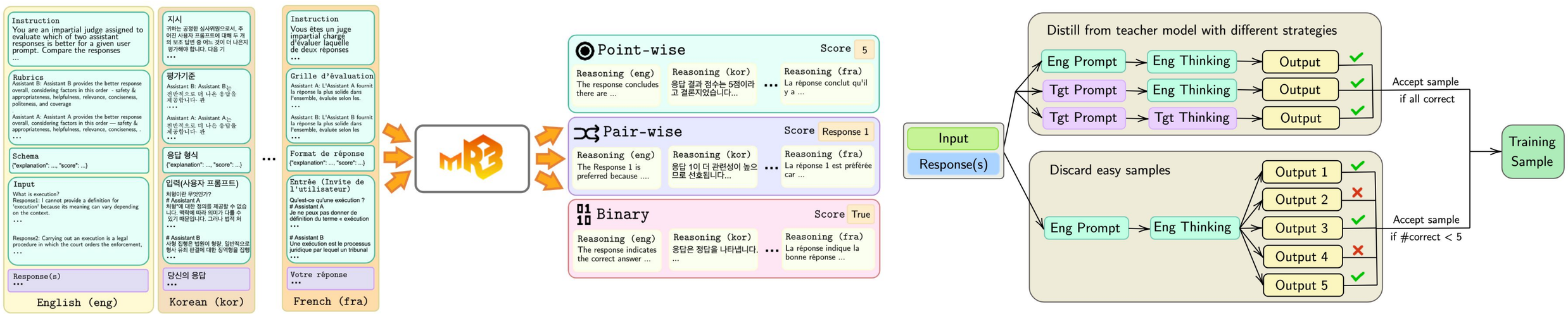
🌍 **Language Diversity:** Trained on **72** languages, the broadest language coverage to date.

🧠 **Non-English Reasoning:** Supports non-English reasoning via language forcing.

🧩 **Rubric Agnostic & Controllable:** Generalizes across rubrics and customizable for new use cases.

Method	# Lang	Data	Model Size (B)	Tasks: Point-wise	Tasks: Pair-wise	Tasks: Binary	Rubrics Customizable	Access*
ArmoRM (Wang et al., 2024a)	1	~974.4k	8	✓	-	-	-	✓
CLoud (Ankner et al., 2024)	1	~280k	8, 70	✓	-	-	-	✓
GenRM (Zhang et al., 2024)	1	~157.2k	2, 7, 9, ?	✓	-	✓	-	-
JudgeLRM (Chen et al., 2025a)	1	100K	3, 7	✓	✓	-	✓	✓
Prometheus1 (Kim et al., 2023)	1	100k	7, 13	✓	✓	-	✓	✓
Prometheus2 (Kim et al., 2024)	1	300k	7, 8X7	✓	✓	-	✓	✓
m-Prometheus (Pombal et al., 2025)	6	480k	4, 8, 14	✓	✓	-	✓	✓
Self-Taught (Wang et al., 2024b)	1	?	70	-	✓	-	✓	✓
Nemotron-English (Wang et al., 2025c)	1	22.4k	32, 70	✓	✓	-	✓	✓
Nemotron-Multilingual (Wang et al., 2025c)	13	40.5k	49, 70	✓	✓	-	✓	✓
SynRM (Ye et al., 2024)	1	5k	7, 35	-	✓	-	-	-
UniEval (Zhong et al., 2022)	1	~185.5k	1	-	-	✓	✓	✓
G-Eval (Liu et al., 2023)	?	?	?	✓	✓	✓	✓	-
Hercule (Doddapaneni et al., 2024)	6	100k	3, 7, 8	✓	-	-	✓	-
FLAMe (Vu et al., 2024)	1	5M+	24	✓	✓	✓	✓	-
RM-R1 (Chen et al., 2025b)	1	~100k	7, 14, 32	-	✓	-	✓	✓
R3 (Anugraha et al., 2025)	1	{4k, 14k}	4, 8, 14	✓	✓	✓	✓	✓
MR3	**72**	100K	4,8,14	✓	✓	✓	✓	✓

Methodology



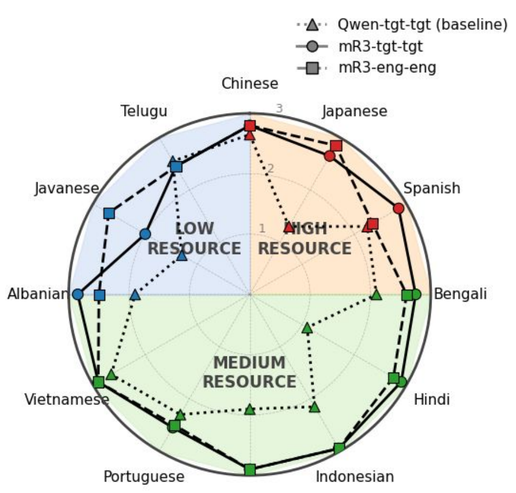
Results Summary

Model	m-RewardBench Acc. 23 langs	RewardBench Acc. 1 lang	MM-Eval Acc. 18 langs	IndoPref Acc. 1 lang
Base Models				
QWEN3-4B	84.51 ± 0.08	88.04 ± 0.37	80.07 ± 0.52	68.80 ± 0.32
QWEN3-8B	86.57 ± 0.06	88.77 ± 0.13	81.95 ± 0.31	72.30 ± 0.45
QWEN3-14B	88.46 ± 0.12	89.72 ± 0.42	84.33 ± 0.35	73.22 ± 0.25
GPT-OSS-20B	86.66 ± 0.30	87.81 ± 0.39	82.03 ± 1.25	69.71 ± 0.43
GPT-OSS-120B	89.05 ± 0.06	90.30 ± 0.50	85.01 ± 0.24	72.15 ± 0.15
DEEPSEEK-R1-14B	68.53 ± 1.62	70.91 ± 1.48	58.77 ± 2.10	55.19 ± 2.73
QWEN2.5-14B-INSTRUCT	77.21 ± 0.06	80.64 ± 0.28	78.00 ± 0.35	70.04 ± 0.26
Existing Reward Models				
PROMETHEUS-7B-V2.0	67.31	72.05	60.90	57.41 ± 0.72
PROMETHEUS-8X7B-V2.0	75.15	74.06	64.34	58.38 ± 0.70
M-PROMETHEUS-7B	77.54	76.84	69.66	60.08 ± 0.66
M-PROMETHEUS-14B	79.51	79.67	77.26	48.16 ± 0.11
R3-QWEN3-14B-LORA-4K	88.07 ± 0.13	**91.00** ± 0.40	84.04 ± 0.34	72.65 ± 0.77
R3-QWEN3-8B-14K	85.86 ± 0.26	88.80 ± 0.09	80.03 ± 0.59	71.60 ± 0.75
R3-QWEN3-4B-14K	84.64 ± 0.20	87.50 ± 0.27	79.37 ± 0.40	70.83 ± 0.84
RM-R1-14B	85.49 ± 0.55	88.51	74.12 ± 1.27	66.42 ± 1.72
RM-R1-32B	87.98 ± 0.28	90.89	80.62 ± 0.67	69.33 ± 0.56
NEMOTRON-49B-EN-THINKING	88.25 ± 0.02	88.72 ± 0.28	75.47 ± 0.11	69.59 ± 0.26
NEMOTRON-MULTILINGUAL-49B-EN-THINKING	89.03 ± 0.03	89.62 ± 0.06	76.27 ± 0.05	68.40 ± 0.06
MR3 Models (Ours)				
MR3-QWEN3-4B	87.61 ± 0.17	89.74 ± 0.52	82.62 ± 0.51	72.22 ± 0.25
MR3-QWEN3-8B	88.44 ± 0.07	90.50 ± 0.25	84.84 ± 0.37	72.86 ± 0.16
MR3-QWEN3-14B	**89.18** ± 0.08	90.79 ± 0.25	**86.05** ± 0.18	**74.14** ± 0.26
MR3-DEEPSEEK-R1-14B	87.12 ± 0.37	88.73 ± 0.95	81.85 ± 0.38	70.11 ± 0.98
MR3-QWEN2.5-14B-INSTRUCT	85.41 ± 0.49	88.21 ± 0.51	81.51 ± 0.64	68.10 ± 0.55

Model	m-ArenaHard-v2.0 English-only WR (%) 1 lang	m-ArenaHard-v2.0 English-only 95% CI 1 lang	m-ArenaHard-v2.0 Overall WR (%) 23 langs	m-ArenaHard-v2.0 Overall 95% CI 23 langs	INCLUDE Acc. 44 langs	M-IMO Acc. 38 langs	MCLM MT-AIME2024 Acc. 55 langs	MCLM MT-MATH100 Acc. 55 langs
QWEN3-30B-A3B-INSTRUCT-2507	49.1	[45.3, 52.6]	39.1	[38.4, 39.9]	64.96 ± 0.08	40.22 ± 0.90	60.79 ± 0.59	90.47 ± 0.33
+ DPO w/ Nemotron-Multilingual-49B	56.2	[52.5, 59.9]	47.0	[46.3, 47.8]	66.09 ± 0.67	42.43 ± 1.11	63.35 ± 1.02	90.45 ± 0.17
+ DPO w/ mR3-Qwen3-14B (Ours)	57.3	[53.5, 61.1]	45.2	[44.4, 45.9]	68.75 ± 0.20	44.02 ± 0.86	65.90 ± 0.52	92.08 ± 0.24

[Takeaway 3]

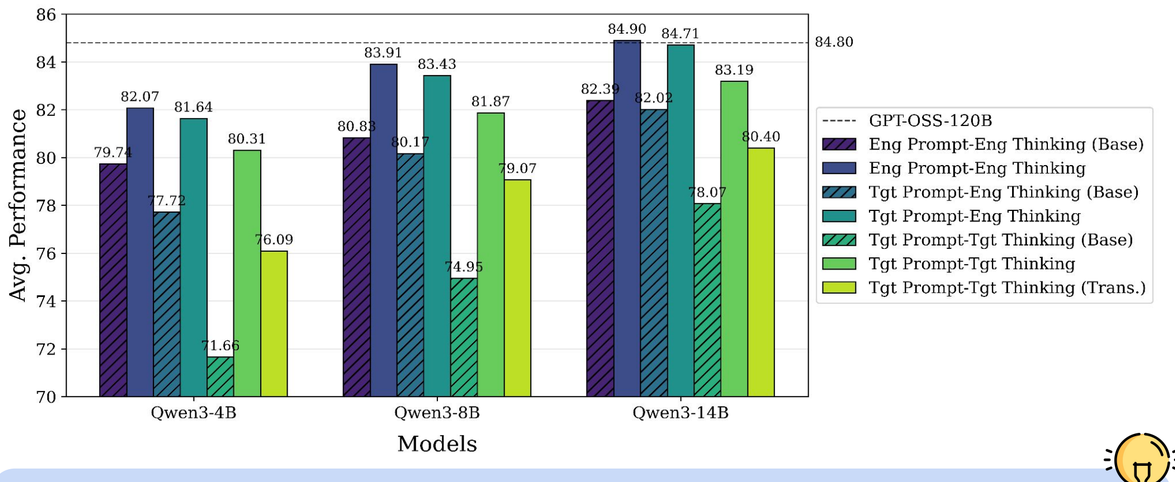
Using **mR3** as a reward signal for post-training yields a stronger policy model on mulitlingual benchmarks compared to other SOTA multilingual RM.



Ablation Category	Setting	Avg.
Dataset Size	10K SFT	80.49 ± 0.19
	25K SFT	80.36 ± 0.41
	50K SFT	80.56 ± 0.24
	100K SFT	**82.09** ± 0.27
Curriculum	EasyToHard (Final)	**82.09** ± 0.27
	Random Shuffle	81.66 ± 0.26
	Start English	81.71 ± 0.23
	HardToEasy	81.58 ± 0.38
	StartEng-EasyToHard	80.25 ± 0.18
	StartEng-HardToEasy	81.58 ± 0.31
Teacher Model	Strong Teacher (GPT-OSS-120B)	**82.09** ± 0.27
	Weak Teacher (GPT-OSS-20B)	78.21 ± 0.23
	Baseline	79.63 ± 0.26
Training Strategy	50K SFT	80.56 ± 0.24
	100K SFT	**82.09** ± 0.27
	50K SFT + 50K RLVR	80.43 ± 0.30

[Takeaway 1]

mR3 matches or beats models up to 9× larger (GPT-OSS-120B) across multilingual reward modeling benchmarks across different formats (pointwise, pairwise, and binary) and domains.



[Takeaway 2]

mR3 can reason in the input language better, even surpasses the base model's English reasoning. Human evaluation also confirms **mR3** produces more factually and logically sound reasoning.

[Takeaway 4]

Human study with 19 native speakers on reasoning trace faithfulness (including unseen LRLs during training) shows that **mR3** models acquire more accurate and culturally aligned reasoning.

[Takeaway 5]

We also perform ablations on dataset size, curriculum, teacher model, and training strategy to justify our design choices. To summarize, we can probably scale for more dataset and a strong teacher is required, with using RLVR offering no additional gain and less efficient.