DogeRM: Equipping Reward Models with Domain Knowledge through Model Merging

Tzu-Han Lin, Chen-An Li, Hung-yi Lee, Yun-Nung (Vivian) Chen
National Taiwan University

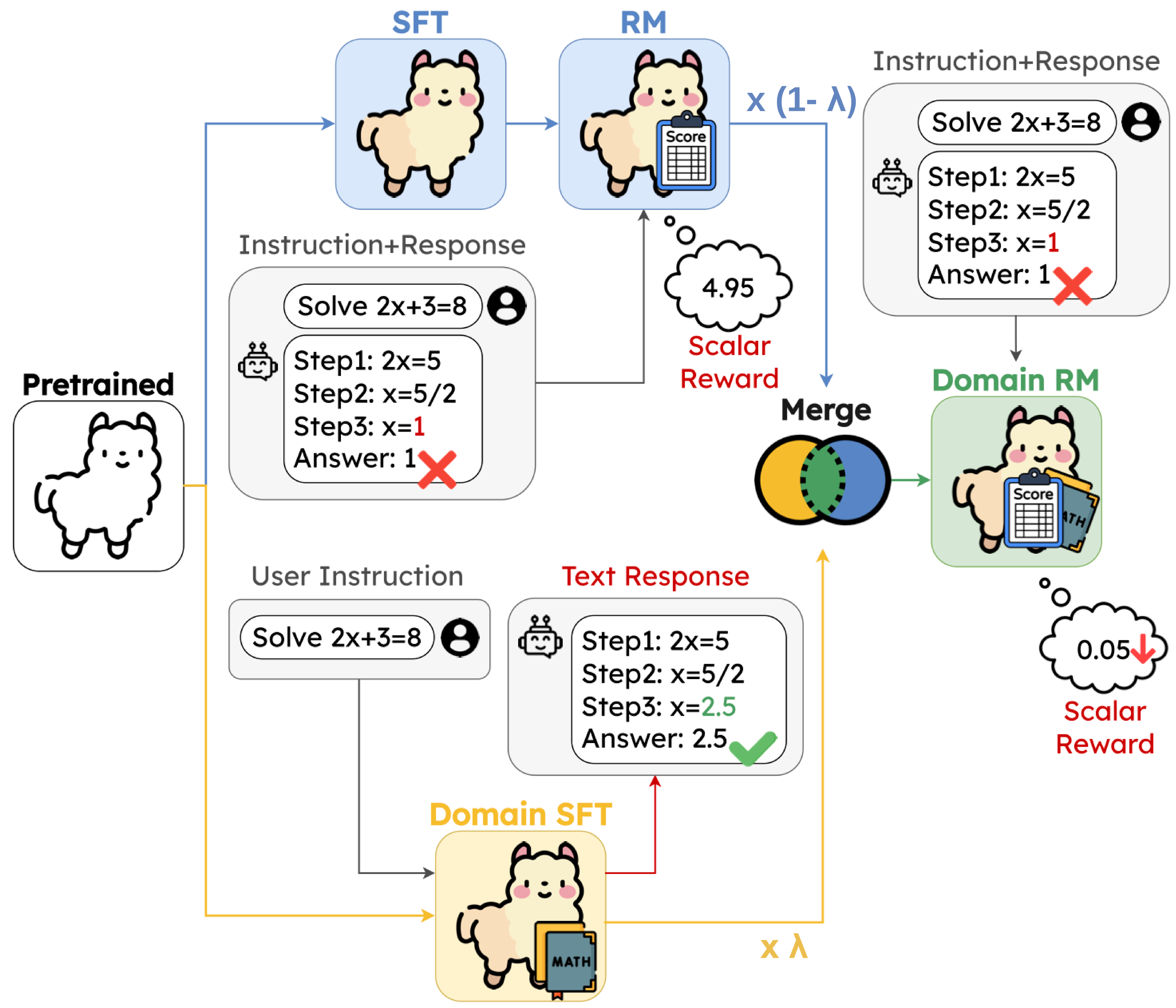
Motivation

- Collecting preference data for RM training is costly, especially for ***domain-specific*** preference data requiring **domain experts**.

Paper Idea

- Domain-specific instruction-tuning data are relatively more accessible to domain-specific preference data.
- **Model Merging** combines multiple single-domain LMs into a multi-domain LM without extra training.
- RQ: *Can we merge **classifier-based RMs** with domain-specific LM to integrate domain knowledge?*

Domain Knowledge LM + RM = DogeRM!



- Starting from the same pre-trained model, we fine-tune a general classifier-based RM.
- With Domain SFT LMs, we can adopt any model merging techniques to obtain Domain RM!

Summary

- Merging RMs with domain LM enhance RM performance on various benchmarks.
- DogeRM can generalize to different benchmarks and model architectures.

Different Model / Multiple Domain

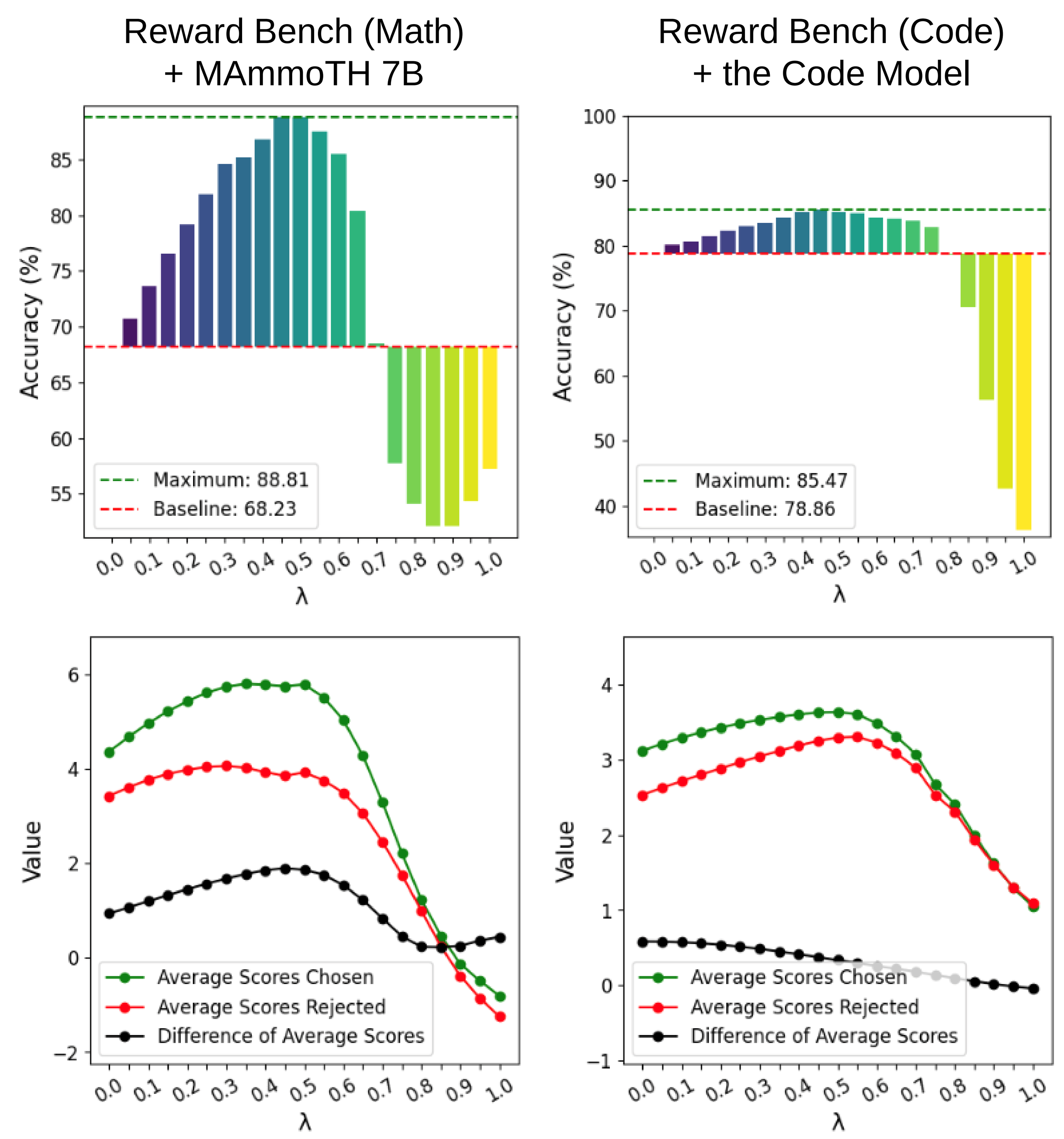
DogeRM can be applied to different model architecture!

Model	Reward Bench		Auto-J Eval		Best-of-16
	Code	Math	Code	Math	GSM8K
Mistral RM	**93.5**	55.0	**88.1**	87.5	44.2
+ MAmmoTH2-Plus	92.6	**85.0**	**88.1**	**90.6**	**46.6**

DogeRM can effectively integrate multiple domains!

Model	Reward Bench		Auto-J Eval		Best-of-16	
	Code	Math	Code	Math	GSM8K	MBPP
LLaMA-2 RM	78.9	68.2	76.2	84.2	35.3	**17.2**
+ Math & Code	**83.0**	**85.2**	**81.0**	**87.5**	**39.5**	17.0

Effect of Weight Factor λ



Experimental Results

DogeRM is effective across different benchmarks!

Model	Reward Bench					Auto-J Eval			Best-of-16	
	Chat	Chat-Hard	Safety	Reasoning		Code	Math	Others	GSM8K	MBPP
				Code	Math					
(a) LLaMA-2 RM	95.8	47.6	**44.6**	78.9	68.2	76.2	84.4	79.2	35.3	**17.2**
(b) FT on Auto-J Math	94.7	**48.5**	44.4	79.1	68.7	76.2†	**90.2**†	79.2†	35.2	-
(c) FT on Auto-J Code	94.7	48.2	44.3	78.8	66.9	**89.3**†	84.4†	79.4†	-	**17.2**
(d) Ours (+ MetaMath)	95.8	44.5	43.5	**85.7**	79.6	79.8	87.5	79.3	**40.7**	-
(e) Ours (+ MAmmoTH)	**96.1**	44.7	43.8	84.1	**85.2**	79.8	87.5	**79.7**	40.5	-
(f) Ours (+ Code Model)	**96.1**	45.6	43.9	84.3	71.8	82.1	87.5	**79.7**	-	**17.2**

r12944034@csie.ntu.edu.tw @tzuhan_0316