NEURAL INFORMATION PROCESSING SYSTEMS

Policy Discriminators are General Reward Models

Shihan Dou*, Shichun Liu*, Yuming Yang*, Yicheng Zou*, Yunhua Zhou, Shuhao Xing, Chenhao Huang, Qiming Ge, Demin Song, Haijun Lv, Songyang Gao, Chengqi Lv, Enyu Zhou, Honglin Guo, Zhiheng Xi, Wenwei Zhang, Qipeng Guo, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Tao Gui, Kai Chen

Shanghai AI Laboratory **Fudan University**

NeurIPS 2025

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

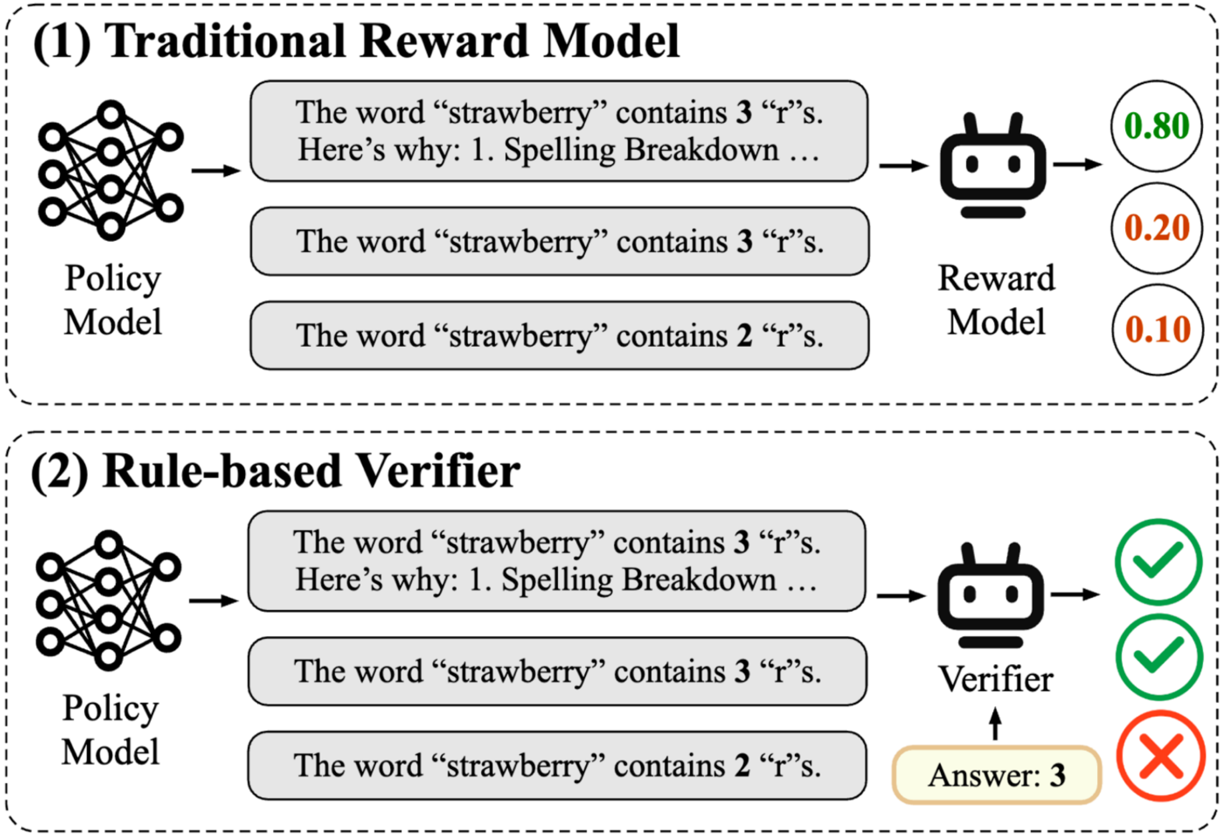
Problems and Challenges:
How Should Rewards Be Properly Modeled?

Traditional Reward Model

- Trains on preference data and outputs preference scores.
- Requires large-scale annotation covering all policy stages.
- Poor generalization.
- Conflicting annotation.

Rule-based Verifier

- Works well for verification but is hard to scale.
- Requires rubric-based labeling.



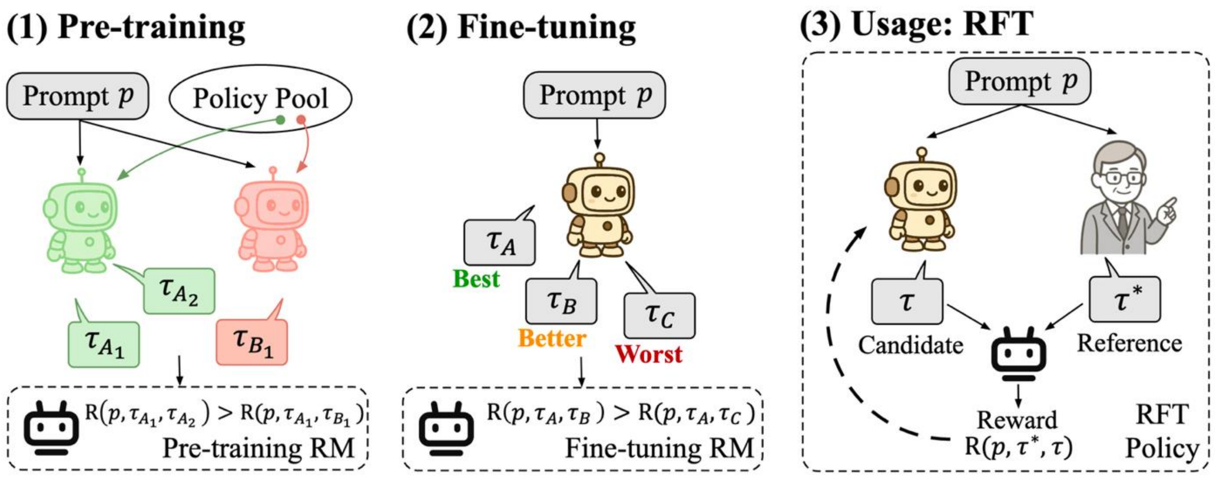
Two Popular Reward Modeling Methods

Our Method

Reward Model is a Policy Discriminator, POLAR

Learn the **differences between policies**, enabling the reward model to assign higher scores to candidate policies that are closer to the target policy.

- Instead of modeling absolute preferences or defining absolute good/bad, measure the **relative distance** between candidate and reference policies.
- Decouples optimization objectives from subjective preferences, allowing for **large-scale scalability**.



The Overview of POLAR

Training Paradigm and Applications

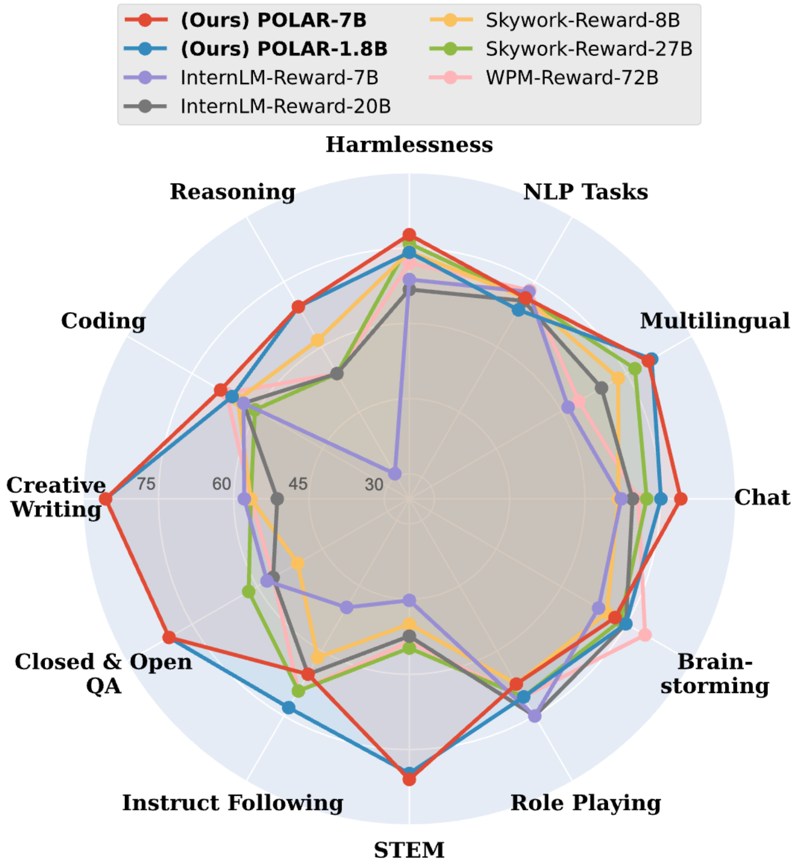
- **Pre-training:** Trajectories from the same policy are treated as positives, while those from different policies are negatives. Policies can be sampled freely, **no manual labels required**.
- **Supervised Fine-tuning:** Analogous to instruction fine-tuning in LLMs: aligns with downstream tasks and human comparative judgments.
- **Usage:** Given a reference answer, train the policy using **RFT (Reference-based Fine-Tuning)**.

Experimental Results

Performance of POLAR on Human Preference Prediction

POLAR exhibits outstanding generalization, consistently outperforming baseline RMs across most tasks.

Notably, on the STEM task, POLAR-1.8B and POLAR-7B surpass the best baseline by over 24.9 and 26.2 percentage points, respectively.



Experimental Results

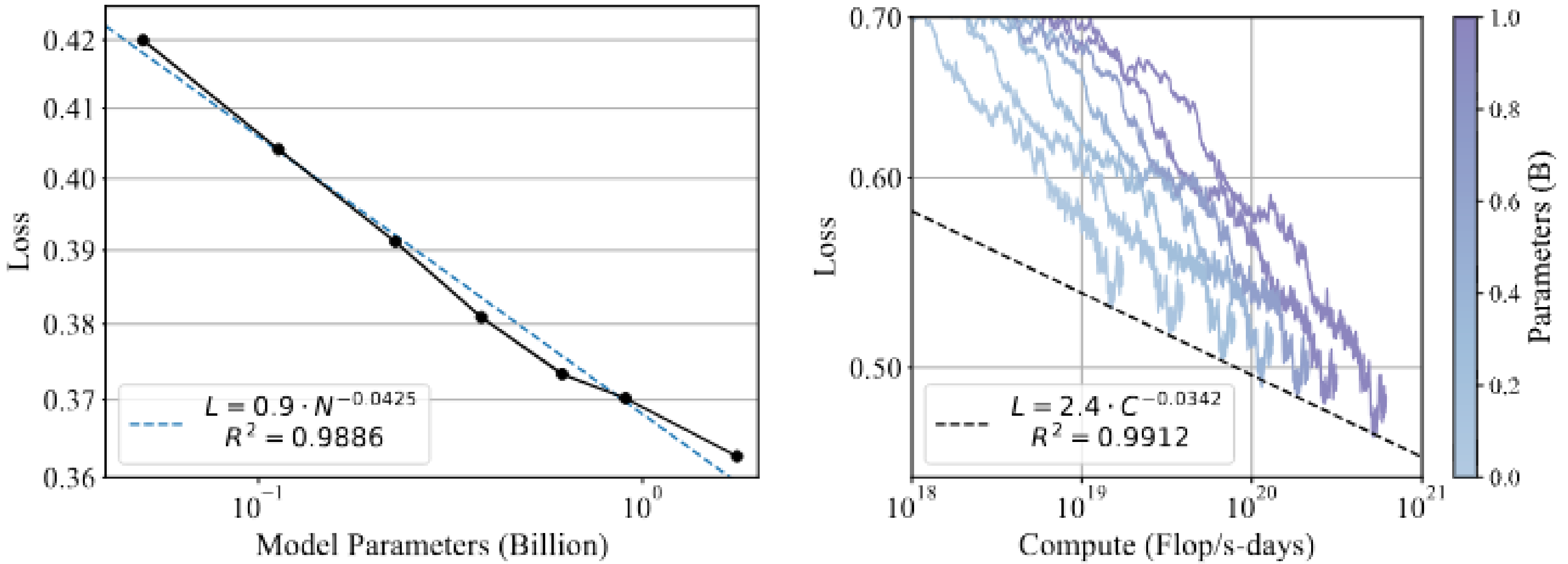
Performance of POLAR on RLHF

POLAR RMs consistently outperform traditional non-pre-trained RMs.

For instance, the Llama-3.1 fine-tuned using POLAR-7B achieves an average improvement of 9.0% over the initial policy model and 6.7% over the policy model optimized by WorldPM-72B-UltraFeedback across all benchmarks.

Policy Model	Reward Model	General Task	Instruct Following	Coding	General Reasoning	Math	Knowledge	Average
InternLM3-8B-Instruct	Baseline	24.07	62.65	74.40	64.37	83.11	60.94	56.49
	InternLM2-Reward-7B	28.02	64.45	78.63	64.84	79.96	60.43	57.82
	Skywork-Reward-8B	29.21	63.75	74.66	64.82	83.36	59.95	57.92
	InternLM2-Reward-20B	28.76	66.75	74.16	64.97	82.20	60.65	58.09
	Skywork-Reward-27B	30.20	64.95	74.35	65.18	83.23	59.91	58.34
	WorldPM-72B-UltraFeedback	34.89	67.90	77.13	65.56	84.29	61.08	60.49
	POLAR-1.8B (Ours)	**37.50**	72.70	78.24	66.79	84.33	64.40	62.60
	POLAR-7B (Ours)	37.35	**73.25**	**79.63**	**67.89**	**85.18**	**64.46**	**63.18**
Llama-3.1-8B-Instruct	Baseline	15.59	63.35	70.69	52.95	67.60	49.39	47.36
	InternLM2-Reward-7B	25.37	60.80	59.24	54.15	65.21	46.35	48.06
	Skywork-Reward-8B	24.80	61.80	67.53	53.54	66.23	49.36	49.22
	InternLM2-Reward-20B	26.52	62.85	58.57	52.41	64.45	45.09	47.70
	Skywork-Reward-27B	24.57	61.70	66.34	54.58	66.25	49.97	49.44
	WorldPM-72B-UltraFeedback	21.36	63.85	70.86	54.74	69.56	49.70	49.64
	POLAR-1.8B (Ours)	27.96	65.20	71.35	57.52	71.11	51.30	52.71
	POLAR-7B (Ours)	**37.02**	**69.30**	**72.14**	**59.85**	**72.20**	**51.69**	**56.33**
Qwen2.5-7B-Instruct	Baseline	26.52	66.05	79.24	53.83	83.47	61.98	54.95
	InternLM2-Reward-7B	31.99	64.05	72.80	56.48	80.35	55.24	54.95
	Skywork-Reward-8B	32.44	68.00	76.71	58.09	83.13	58.12	57.04
	InternLM2-Reward-20B	33.05	68.40	74.06	55.41	82.62	58.36	56.15
	Skywork-Reward-27B	34.28	69.45	78.21	57.46	83.58	59.43	57.84
	WorldPM-72B-UltraFeedback	35.72	70.55	77.48	59.48	83.35	59.48	58.83
	POLAR-1.8B (Ours)	35.76	**71.35**	80.40	**62.52**	84.19	60.39	60.35
	POLAR-7B (Ours)	**37.70**	71.15	**81.15**	61.30	**84.70**	**62.57**	**60.90**
Qwen2.5-32B-Instruct	Baseline	31.07	75.50	86.56	69.74	89.35	71.07	64.49
	InternLM2-Reward-7B	36.10	75.70	83.13	69.72	87.20	64.99	64.29
	Skywork-Reward-8B	36.44	79.60	84.42	71.07	89.29	68.77	66.08
	InternLM2-Reward-20B	37.98	74.45	85.76	69.32	89.35	66.68	65.25
	Skywork-Reward-27B	38.43	80.10	83.95	71.93	86.78	69.15	66.64
	WorldPM-72B-UltraFeedback	40.59	78.65	86.79	70.38	89.90	69.05	67.15
	POLAR-1.8B (Ours)	40.24	80.25	87.47	72.23	90.03	**73.67**	68.55
	POLAR-7B (Ours)	**45.98**	**80.50**	**88.92**	**73.17**	**90.39**	73.59	**70.47**

Scaling Laws



We observe a clear power-law relationship between validation loss and model size

The right panel of this figure shows that validation loss follows a power-law scaling trend with respect to compute

NEURAL INFORMATION PROCESSING SYSTEMS

Conclusion

1. We propose POLAR, a novel criterion-agnostic pre-training paradigm for reward modeling based on a scalable training objective, i.e., policy discrimination.

2. Scaling experiments reveal promising scaling laws, highlighting the significant potential of POLAR for enhancing the upper bound of reward modeling and developing stronger and more generalizable reward models.

3. We developed the POLAR series of reward models. They substantially outperform traditional RMs in empirical evaluations, achieving higher preference accuracy and better generalization than considerably larger reward models.

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory