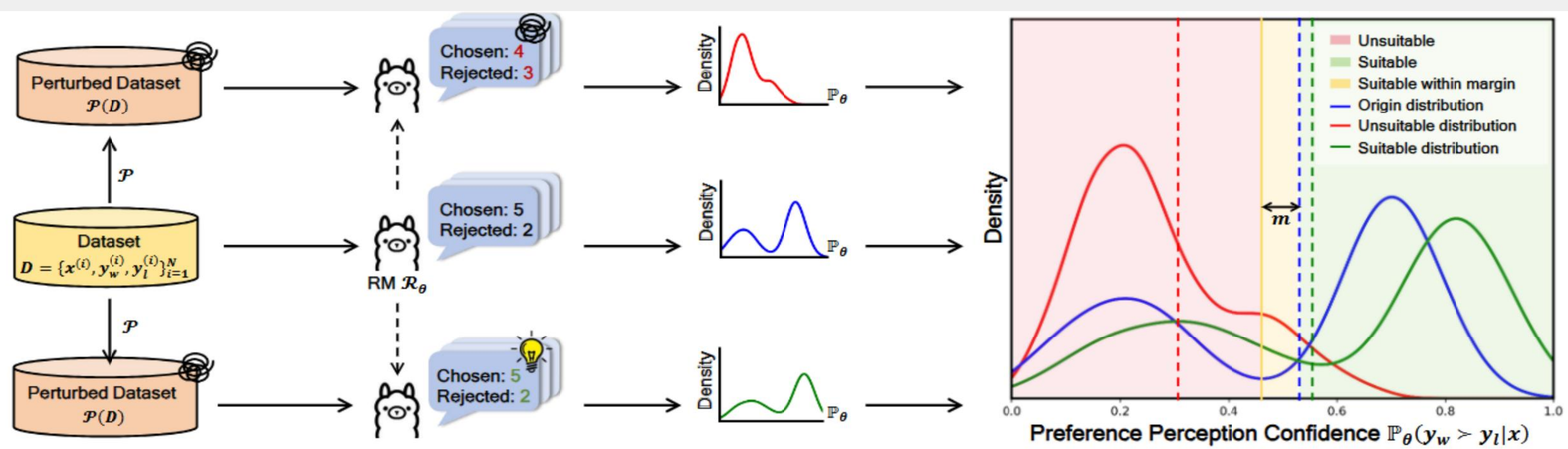
Reward Auditor: Inference on Reward Modeling Suitability in Real-World Perturbed Scenarios

Jianxiang Zang Yongda Wei Ruxue Bai Shiyu Jiang Nijia Mo Binhong Li Qiang Sun Hui Liu

ICML International Conference On Machine Learning

Introduction



Traditional reward model benchmark: How accurate is the RM's preference perception for given samples?

Reward Auditor: ***Can we infer that RMs exhibit systematic vulnerabilities in specific real-world scenarios?***

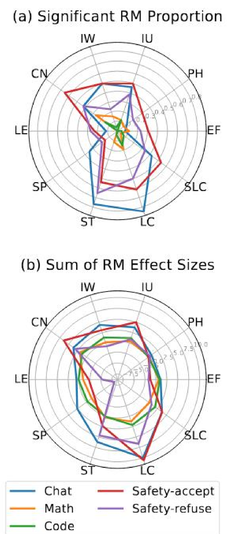
Reward Auditor pipeline

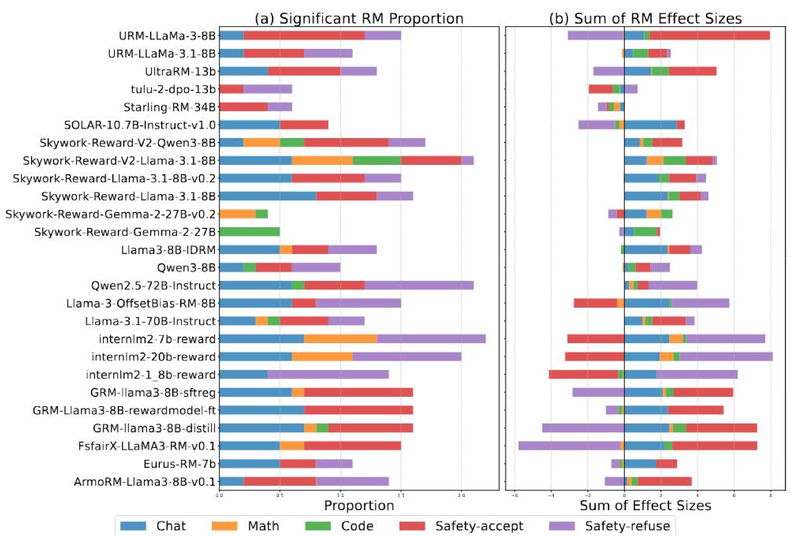
Algorithm 1 Reward Auditor for evaluating **suitability** of an RM

Require:
(1) Reward model $\mathcal{R}_\theta$, (2) Dataset $D = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$, (3) Perturbed dataset $D' = \mathcal{P}(D)$, (4) Number of permutations B, (5) Ordered significance level set $\boldsymbol{\alpha}$
Define: Preference perception confidence $\mathbb{P}_\theta(x, y_w, y_l) = \sigma\left[\mathcal{R}_\theta(x, y_w) - \mathcal{R}_\theta(x, y_l)\right]$
procedure PAIREDPERMUTATIONTEST(D, D', B, $\boldsymbol{\alpha}$)
  Build preference perception confidence distribution $M \leftarrow \{\mathbb{P}_\theta(D_i)\}_{i=1}^N$, $M' \leftarrow \{\mathbb{P}_\theta(D'_i)\}_{i=1}^N$,
  Construct paired difference distribution $\Delta M \leftarrow M - M'$
  Build test statistic $\hat{t}_{\text{obs}} \leftarrow \frac{\overline{\Delta M_i}}{\text{std}(\Delta M_i)/\sqrt{N}}$
  Initialize counter $c \leftarrow 0$
  for $j = 1$ to B **do**
    Random permute $\hat{t}_{\text{perm}} \leftarrow \frac{\overline{\Delta M_i \cdot s_i}}{\text{std}(\Delta M_i \cdot s_i)/\sqrt{N}}$, $\boldsymbol{s} \sim \mathcal{U}\{-1, 1\}^N$
    if $\hat{t}_{\text{perm}} \geq \hat{t}_{\text{obs}}$ **then**
      $c \leftarrow c + 1$
    end if
  end for
  $\hat{p} \leftarrow \frac{c+1}{B+1}$, $\hat{e} \leftarrow \frac{\overline{\Delta M_i}}{\text{std}(\Delta M_i)}$, $r_S \leftarrow \hat{e} \wedge \mathbb{I}^*(\hat{p}, \boldsymbol{\alpha})$
  Ensure: Effect size $\hat{e}$, p-value $\hat{p}$, Suitability risk report r_S
end procedure

Case Studies

Reward Models	Controlled Perturbation (Prompt)					Stylized Perturbation (Response)				
	EF	PH	IU	IW	CN	LE	SP	ST	LC	SLC
♡Starling-RM-34B	-0.120	-0.135	-0.002	-0.088	-0.040	0.094	0.093	0.008	-0.074	0.019
♠tulu-2-dpo-13b	-0.145	0.043	-0.105	-0.113	0.037	0.089	0.081	-0.045	-0.081	0.006
♡Skywork-Reward-Gemma-2-27B	0.131	-0.001	0.034	0.046	0.115	0.087	-0.037	0.043	0.102	-0.009
♡Skywork-Reward-Gemma-2-27B-v0.2	0.052	0.085	0.132	0.109	0.174	0.190	0.077	0.026	0.182	0.192
♣Qwen2.5-72B-Instruct	-0.033	0.041	-0.102	-0.007	0.043	0.002	0.312***	-0.221	0.229**	-0.008
♡URM-LLaMa-3-8B	0.068	0.057	-0.018	0.030	0.037	-0.107	0.039	0.370***	0.490***	0.114
♡Skywork-Reward-V2-Qwen3-8B	0.120	-0.032	-0.039	0.186	0.050	-0.197	-0.107	0.347***	0.435***	0.079
♣Llama-3.1-70B-Instruct	0.004	0.029	0.018	0.239**	-0.023	0.092	0.081	0.262**	0.241**	0.037
♡URM-LLaMa-3.1-8B	-0.059	0.002	0.049	0.010	-0.063	-0.127	-0.039	0.237**	0.446***	0.025
♡ArmoRM-Llama3-8B-v0.1	-0.071	-0.127	0.061	-0.165	-0.055	-0.216	-0.117	0.350***	0.500***	-0.004
♣Qwen3-8B	-0.081	0.034	0.151*	0.019	-0.302	0.206**	0.166*	0.109*	0.141*	-0.210
♠SOLAR-10.7B-Instruct-v1.0	0.047	-0.067	-0.163	-0.117	0.067	0.774***	0.906***	0.266**	0.458***	0.696***
♡UltraRM-13b	0.091	-0.166	0.174	0.366***	0.162	0.121	0.255**	0.267**	0.303***	-0.118
♡internlm2-1_8b-reward	0.160	0.242**	0.162	0.114	0.350***	0.036	0.046	0.151	0.229**	0.258**
♡Eurus-RM-7b	0.062	-0.009	-0.062	-0.070	0.042	0.395***	0.389***	0.207*	0.393***	0.399***
♡FsfairX-LLaMA3-RM-v0.1	-0.006	0.066	0.502***	0.466***	0.202*	-0.069	0.013	0.328***	0.550***	0.128
♡Skywork-Reward-Llama-3.1-8B-v0.2	-0.100	0.147	0.170*	0.205**	0.228**	-0.044	0.255**	0.435***	0.436***	0.149
♡internlm2-20b-reward	0.029	-0.003	0.265***	0.250**	0.260**	-0.062	0.001	0.199*	0.515***	0.476***
♡Skywork-Reward-V2-Llama-3.1-8B	0.067	0.143	0.188*	0.255**	0.271***	-0.382	-0.489	0.449***	0.549***	0.171*
♡GRM-llama3-8B-sftreg	0.148	-0.030	0.315***	0.382***	0.254**	-0.010	0.009	0.304***	0.549***	0.176*
♡Llama3-8B-IDRM	0.115	-0.144	0.363***	0.360***	0.155	0.162*	0.247**	0.336***	0.525***	0.340***
♡Llama-3-OffsetBias-RM-8B	-0.065	0.247**	0.509***	0.530***	0.385***	-0.202	-0.112	0.376***	0.741***	0.047
♡GRM-llama3-8B-distill	0.313***	0.019	0.349***	0.422***	0.247**	-0.036	0.032	0.337***	0.581***	0.190*
♡GRM-Llama3-8B-rewardmodel-ft	0.076	0.148	0.230**	0.263**	0.391***	-0.087	0.162*	0.359***	0.639***	0.204*
♡internlm2-7b-reward	0.302***	0.006	0.344***	0.406***	0.387***	-0.062	0.100	0.313***	0.465***	0.195*
♡Skywork-Reward-Llama-3.1-8B	-0.052	0.214**	0.265***	0.273***	0.306***	0.058	0.234**	0.381***	0.457***	0.254**





- The majority of RMs demonstrate highly idiosyncratic vulnerability patterns.
- Stylized perturbations pose a greater systemic risk than controlled perturbations.
- Semantic and structural alterations are the main failures of the suitability.

- RM suitability shows robustness in objective domains, but is brittle in subjective domains.
- RMs exhibit high domain specificity in suitability.

CONTACT US

WeChat:

Paper:

Github: