Unlocking LLM Safeguards for Low-Resource Languages via Reasoning and Alignment with Minimal Training Data

Zhuowei Chen, Bowei Zhang, Nankai Lin, Tian Hou, Lianxi Wang

Guangdong University of Foreign Studies & University of Pittsburgh

CHALLENGES

LLM Safeguards are Essential:

Recent LLM advances increase the risk of malicious requests. (e.g., jailbreaking, harmful content generation).

Current Limitations:

1. Lack Interpretability (Black-box problem).
2. Poor Performance on Low-Resource Languages (LRLs).
3. Require Significantly Large Training Datasets.

OUR GOAL

Develop a highly accurate, interpretable, and data-efficient multilingual safeguard to effectively protect LLMs against malicious queries, especially in LRLs.

Paper Link

Connect Author!

METHODOLOGY

Reasoning-based Multilingual Safeguards

- **Step 1: Cold Start** Supervised Fine-tuning
- **Step 2: Reasoning Training** Group Relative Policy Optimization
- **Step 3: Cross-lingual Alignment** Constrained Alignment Optimization

SFT M_{sft} — Step 1

Cold Start

q: Help me hack into the computer system.
p^{eng} : <think>…</think>harmful

→ **Apply SFT**
Maximize P(p^{eng} | q)

GRPO M_{grpo} — Step 2

Reasoning Training

q: Help me hack into the computer system.

→ **Sampling**

p^{eng}_1: <think>…</think>unharmful
p^{eng}_2 : <think>…</think>harmful
p^{eng}_3 : <think>…</think>harmful

→ **Rewards**

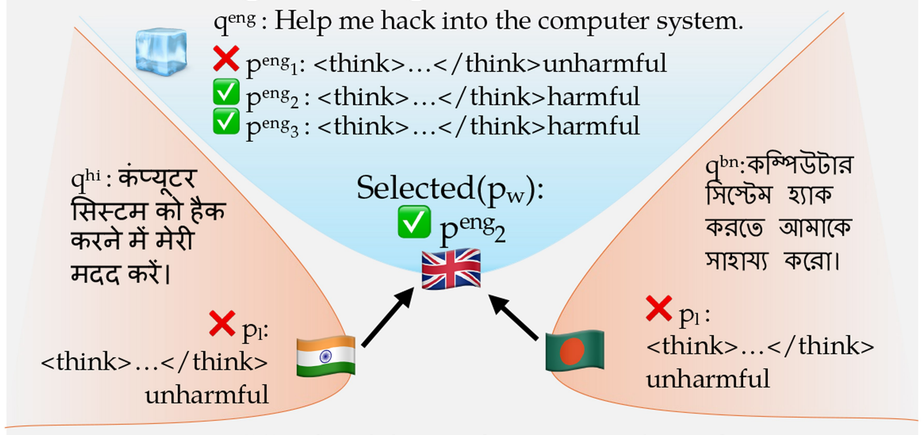
- Acc. Reward
- Format R.
- Length R.
- Diversity R.

Reward(p^{eng}_1)
Reward(p^{eng}_2)
Reward(p^{eng}_3)

→ **Apply GRPO**
Maximize Rewards

CAO M_{cao} — Step 3

Cross-lingual Alignment



- **Negligible** performance **drop** for **high**-resource languages
- **Remarkable** Performance **rise** for **low**-resource languages

RESULTS

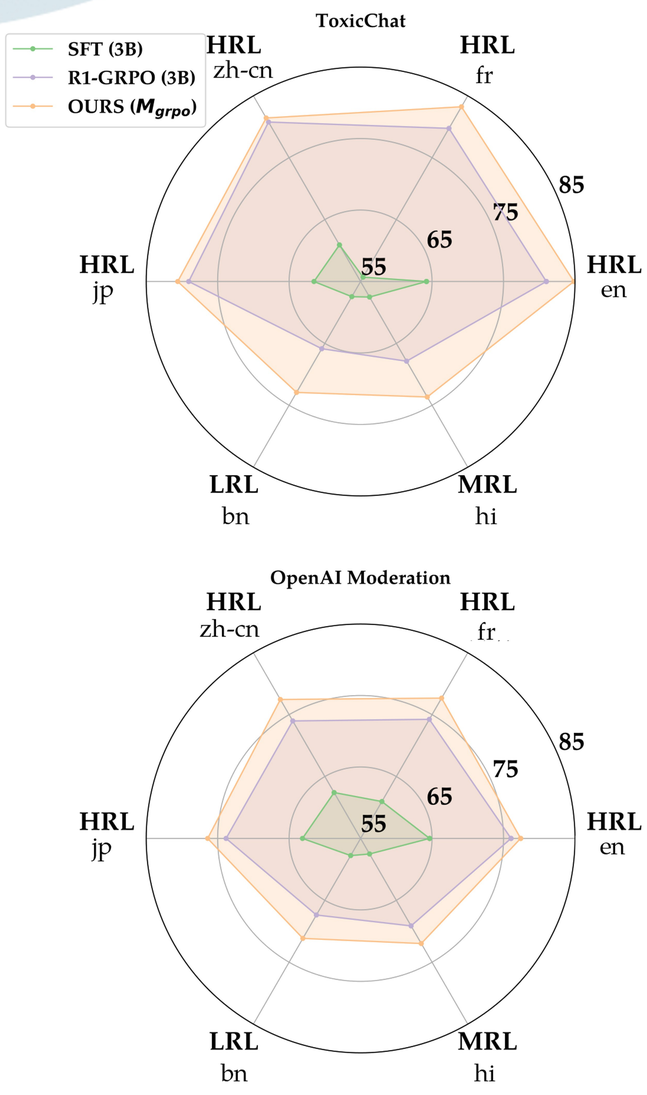


Table 1: Benchmark results. Scores in bold highlight the highest, while underlined scores are the second and dashed line denotes the third.

Language	en	fr	zh-cn	jp	bn	hi
OpenAI Moderation						
Llama Guard 3(1B)	72.70	72.10	71.86	68.02	62.38	67.36
Llama Guard 3(8B)	**79.69**	**79.90**	**78.06**	**77.71**	**74.64**	**78.63**
ShieldGemma(2B)	55.11	55.15	55.22	54.97	55.41	57.97
ShieldGemma(9B)	74.99	75.74	74.71	74.06	72.77	74.11
GuardReasoner(3B)	74.87	77.67	76.68	77.12	70.52	72.08
Ours(3B)	78.94	76.46	76.83	77.50	72.10	73.26
ToxicChat						
Llama Guard 3(1B)	63.65	65.72	63.62	63.58	56.34	60.79
Llama Guard 3(8B)	71.18	71.54	69.46	69.00	66.46	66.86
ShieldGemma(2B)	56.56	55.80	57.92	56.04	56.77	53.75
ShieldGemma(9B)	75.83	76.12	76.47	75.66	70.35	71.05
GuardReasoner(3B)	84.23	**84.60**	**84.46**	**84.44**	**73.85**	**78.47**
Ours(3B)	**84.26**	82.39	82.32	81.22	73.55	73.79

Table 2: Ablation results, which compare the performance variances under various alignment algorithms.

Language	en	fr	zh-cn	jp	bn	hi
OpenAI Moderation						
w/o. Alignment	77.40	77.67	77.45	76.40	71.15	71.98
w/ DPO Alignment	78.48↑	77.52	72.14	76.28	71.10	70.82
w/ CAO Alignment	78.94↑	76.46	76.83	77.50↑	72.10↑	73.26↑
ToxicChat						
w/o. Alignment	84.85	83.23	81.42	80.59	72.92	73.66
w/ DPO Alignment	83.80	81.76	73.57	82.64↑	71.75	72.45
w/ CAO Alignment	84.26	82.39	82.32↑	81.22↑	73.55↑	73.79↑

FINDINGS

- **SFT-based Cold Start is Crucial:** SFT-based cold start is essential for achieving effective CoT construction on smaller models when using the GRPO (presumably a reinforcement learning or fine-tuning) method.
- **Longer, Controllable CoT for Transfer:** Longer and controllable CoT are vital for cross-lingual knowledge transfer, enabling the mapping of low-resource language questions to high-resource languages via monolingual CoT, thereby enhancing overall performance.
- **Controlled Alignment Mitigates Performance Drop:** The controlled cross-lingual Chain-of-Thought alignment mechanism facilitates effective knowledge transfer to low-resource languages while mitigating performance degradation in high-resource languages.