One Token to Fool LLM-as-a-Judge

Yulai Zhao*, Haolin Liu*, Dian Yu, Sunyuan Kung, Meijia Chen, Haitao Mi, Dong Yu

Tencent AI Lab, Princeton University, University of Virginia, Rutgers University

Takeaway: Hacking reference-based LLM judges is easier than you think — as easy as one token

- In reference-based generative reward models, we found certain superficial patterns **consistently** elicit **false positive judgments**:
  - (a) Non-word symbols: ".", ":", or even a blank space.
  - (b) Reasoning openers: "Thought process:", "Solution", "Let's solve this problem step by step."

Figure 1: **Reasoning openers such as "Solution" can trigger false positive judgments in many state-of-the-art generative reward models.**

- These phrases act as "master keys": short, meaningless inputs that still receive positive rewards.
- Affects state-of-the-art models like GPT-4o, Claude-4, Omni-Judge.

Figure 2: **Systematic vulnerabilities of LLM judges exposed by "master key" attacks.** We evaluate various LLM judges on five reasoning benchmarks using ten "master key" responses. Such vague responses can surprisingly lead to false positive rates (FPRs) as high as 80%. In contrast, our Master-RM maintains near-zero FPRs across all settings.

Master-RM: A Robust Reward Model

Figure 3:To resist such "master key" hacks, we obtain a new reward model with a straightforward adversarial data augmentation strategy.

Experimental Results

We empirically prove that our **Master-RMs** not only present the strongest robustness against hacks, but also perform excellently as judge models:

1. **Robustness**: Figure 2 shows our Master RM has the best robustness.
2. **Performance**: Table 1 shows our Master-RMs also achieve high performance on judging benchmarks.

Model/Method	VerifyBench		VerifyBench-Hard	
	Acc	Macro F1	Acc	Macro F1
OpenAI/GPT-o1	**95.70**	**95.70**	**88.80**	**85.48**
OpenAI/GPT-4o	94.15	94.15	84.30	77.94
Master-RM-32B	95.15	95.14	86.80	81.96
Master-RM-7B	94.45	94.45	84.40	80.98
Multi-sub RM	95.00	95.00	82.50	78.42
Omni-Judge	80.20	80.03	67.70	58.98
Qwen2.5-72B-Instruct	94.30	94.30	78.30	72.63

Table 1: **Evaluating LLM judges' accuracies (%) and macro F1 scores (%) on public verifiable benchmarks.**

Additional Observations & Insights

- **Scaling Law Anomaly**: Larger models are not necessarily safer. 72B models often have higher FPRs than 7B-14B models, possibly due to self-solving.

Figure 4: **Scaling Behaviour to "master key" attack (B/FPR)**

- **Inference Strategies**: Chain-of-Thought (CoT) and Majority Voting are unreliable defenses and can sometimes worsen the vulnerability.

Qwen2.5-7B	Qwen2.5-7B-COT
12.6	40.4

Table 2: **Average FPR w/w.o. COT**

- **Prompting Defense**: Removing the **Question** from the judge's prompt significantly reduces FPR in mathematical tasks. But this should be applied with caution for general tasks.

Qwen2.5-7B	Qwen2.5-7B-No-Question
12.6	1.9

Table 3: **Average FPR w/w.o. question**

Acknowledgements

This work was done during YL and HL's internship at Tencent AI Lab.

ArXiv Model Dataset