D-Judge: Disrupting Multi-Turn Jailbreaks using Semantics-Preserving Output Rewriting

Huanli Gong[♠], *Zhipeng Wei*[♠♥], *Yu Fu*[♣], *Haz Sameen Shahgir*[♣], *Ananya Gupta*[♠], *Yue Dong*[♣], *N. Benjamin Erichson*[♠♦]

[♠]*UC Berkeley* [♥]*International CompSci Institute* [♣]*UC Riverside* [♦]*Lawrence Berkeley National Lab*

Counter the attack by disrupting the judge-feedback loop that powers automated jailbreaks.

•

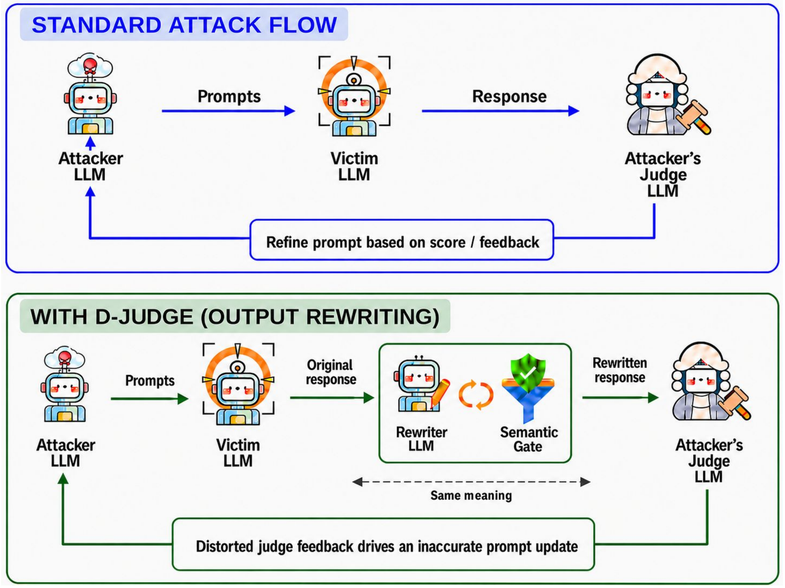
D-Judge rewrites model outputs without changing their meaning, so attacks optimize against a distorted judge signal and fail to converge.

•

Same meaning. Different Judge score. Broken loop.

Core Concept

Key insight: LLM-as-a-judge scores are sensitive to semantics-preserving rewrites, making judge-guided attacks vulnerable to feedback manipulation.



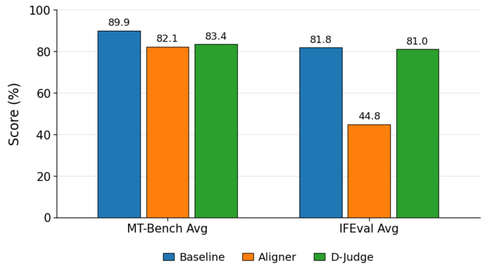
- D-Judge rewrites the victim model's response before the attacker's judge sees it, intervening directly in the feedback loop that guides the next prompt.
- We collect semantically equivalent response pairs with different judge scores, then train the rewriter with SFT and DPO to preserve meaning while inducing misleading judge feedback.
- A Semantic Gate checks that the original and rewritten responses remain semantically equivalent.
- By perturbing judge feedback, D-Judge derails the attacker's multi-turn prompt refinement loop.

Defense Performance

Methods	Crescendo (↓)	CoA (↓)	ActorBreaker (↓)	FITD (↓)	X-Teaming (↓)	Average (↓)
No Defense	81.8	75.5	37.1	24.1	73.0	58.3
Single-turn Guard						
LlamaGuard	71.1	64.2	25.8	17.0	62.4	48.1
QwenGuard	54.7	54.1	20.1	15.1	58.0	40.4
WildGuard	63.5	60.4	34.0	13.2	59.5	46.1
Multi-turn Detection						
NBF	75.5	64.2	28.9	17.7	64.8	50.2
TCA	76.1	73.6	34.0	19.5	67.7	54.2
Input Preprocessing						
Paraphrase	79.9	72.3	30.8	19.5	70.4	54.6
SmoothLLM	73.6	73.4	35.2	21.5	72.2	55.2
Backtranslation	61.2	60.4	24.5	17.7	59.1	44.6
Output Postprocessing						
Aligner	68.6	66.0	23.9	14.6	63.2	47.3
ProAct	72.9	43.4	34.2	18.5	43.0	42.4
D-Judge-GPT	13.8	15.1	3.1	11.3	18.2	12.3
D-Judge	**12.0**	**6.3**	**2.5**	**5.0**	**17.0**	**8.6**

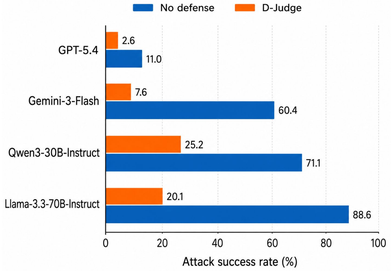
- HarmBench evaluation across five multi-turn jailbreak attacks.
- Here GPT-4o is the victim model. D-Judge achieves the lowest average attack success rate.

Benign Performance



Benign benchmark performance of GPT-4o across MT-Bench and IFEval with D-Judge.

Transfer Across Victim Models



Defense performance against X-Teaming across different victim LLMs.

Read the Paper