SelfGrader: Stable Jailbreak Detection for Large Language Models using Token-Level Logits

Scientific Achievement

- Addressed the open question of how to reliably and efficiently detect jailbreak prompts in large language models, where prior defenses were unstable or slow.
- Developed *SelfGrader*, which uses model logits as an internal safety signal and applies a dual malicious/benign scoring scheme to improve robustness and interpretability.

Significance and Impact

- Enables low-latency, lightweight safety filtering that improves jailbreak detection while remaining practical for real-world deployment.
- Advances the field of AI safety and alignment by showing that internal model signals can replace external classifiers, with broad applicability across domains and industry systems.

Technical Approach

- Reformulates detection as a numeric grading task over token logits (0–9 scale), avoiding expensive auxiliary models.
- Overcomes challenges of latency, stochastic outputs, and false positives through a dual-perspective scoring method that stabilizes predictions.

Guardrails	TAP	LLM-Fuzzer	X-Teaming	Average	Latency (Sec.)	Memory Overhead (MB)
LLama-3-8B-Instruct (No Defense)	14.00/-	49.00/-	91.00/-	51.33/-	1.47	-
Perplexity Filter	14.00/100.00	49.00/100.00	91.00/100.00	51.33/100.00	**0.49**	13571.94
GradSafe	9.00/53.00	OOM	OOM	-	-	-
GradientCuff	5.00/10.00	**0.00/0.00**	1.00/1.00	**2.00/3.67**	26.57	2063.76
Token Highlighter	5.00/7.00	OOM	OOM	-	-	-
Prompt Guard	14.00/93.00	**0.00/0.00**	**0.00/0.00**	4.67/31.00	22.67	1864.11
Llama Guard (Pre)	14.00/46.00	38.00/56.00	77.00/81.00	43.00/61.00	0.96	13910.61
Llama Guard (Post)	14.00/100.00	35.00/84.00	85.00/91.00	44.67/91.67	1.17	13910.61
SelfDefend (Direct)	12.00/20.00	1.00/1.00	59.00/61.00	24.00/27.33	1.19	13452.93
SelfDefend (Intent)	6.00/12.00	3.00/6.00	22.00/22.00	10.33/13.33	3.11	13457.40
WildGuard (Pre)	**2.00**/8.00	**0.00/0.00**	OOM	-	-	-
WildGuard (Post)	4.00/87.00	6.00/45.00	OOM	-	-	-
GuardReasoner (Pre)	3.00/**7.00**	**0.00/1.00**	64.00/65.00	22.33/24.33	14.87	15617.33
GuardReasoner (Post)	5.00/87.00	3.00/41.00	76.00/83.00	28.00/70.33	16.71	15618.13
SelfGrader	6.00/20.00	2.00/5.00	**0.00/0.00**	2.67/8.33	1.24	**479.81**

Table 1: Comparison of different defense methods against three adaptive attacks on LLama-3-8B-Instruct. Defense performance are reported as Attack Success Rate (↓) / Pass Guardrail Rate (↓) in %.

PI(s)/Facility Lead(s): K. Kim (PI), O. Kotevska (ORNL's PI)
ASCR Program: AI for Science
ASCR PM: Xujing Davis
Publication(s) for this work: Zhang, Z., Hu, R., Kotevska, O., Xu, J. SelfGrader: Stable Jailbreak Detection for Large Language Models using Token-Level Logits. In *Third Conference on Language Modeling*. https://arxiv.org/pdf/2604.01473