Rebuttal for Uncovering Safety Risks of Large Language Models through Concept Activation Vector

To Reviewer mJhn:

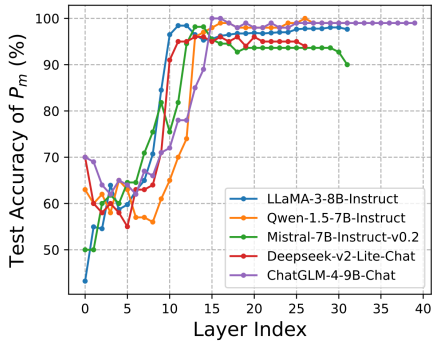

Figure 1: SCAV classifiers' TestAcc on 5 more open-source LLMs

To Reviewer CRgs:

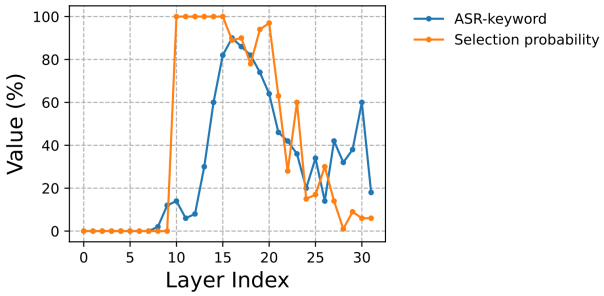

Figure 2: Change of ASR-keyword and selection probability of embedding-level attack by layer

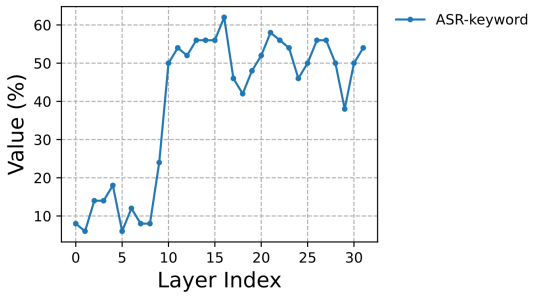

Figure 3: Change of ASR-keyword of prompt-level attack by layer

Rebuttal Table 5: Comparison with GCG regarding attack transferability (applying prompts learned for LLaMA-2 to GPT-4)

Evaluation Dataset	Source Models	Methods	A-k (%)	A-a (%)	A-u (%)	L (%)
Advbench	LLaMA-2 (7B-Chat)	SCAV (prompt)	70	66	52	68
		GCG	4	4	2	92
	LLaMA-2 (13B-Chat)	SCAV (prompt)	82	48	60	54
		GCG	2	0	0	98
StrongREJECT	LLaMA-2 (7B-Chat)	SCAV (prompt)	30	20	20	72
		GCG	8	16	16	90
	LLaMA-2 (13B-Chat)	SCAV (prompt)	40	26	22	72
		GCG	8	12	10	88

A-k: ASR-keyword, A-a: ASR-answer, A-u: ASR-useful, L: Language Flaws