Uncovering Safety Risks of Large Language Models through Concept Activation Vector

Zhihao Xu[1*], Ruixuan Huang[2*], Changyu Chen[1], Xiting Wang[1†]

[1]Renmin University of China, [2]The Hong Kong University of Science and Technology

Motivation

- **Interpretability:** What are the **safety mechanisms** within LLMs?
- **Controllability:** Can we enable automatic hyperparameter selection?
- **Transferability:** Can we apply prompt-level attacks based on our understanding of the safety concepts?

Contribution

(1) We establish a Safety Concept Activation Vector (SCAV) framework that effectively guides the attack by accurately interpreting LLMs' safety mechanisms.

(2) We then develop an SCAV-guided attack method, enabling automatic hyperparameter selection, and support both embedding-level and prompt-level attacks.

(3) Based on SCAV framework, we have revealed the safety risks of LLMs, whether they are open source or **closed source**, and **even models that have undergone unlearning**.

Interpretability

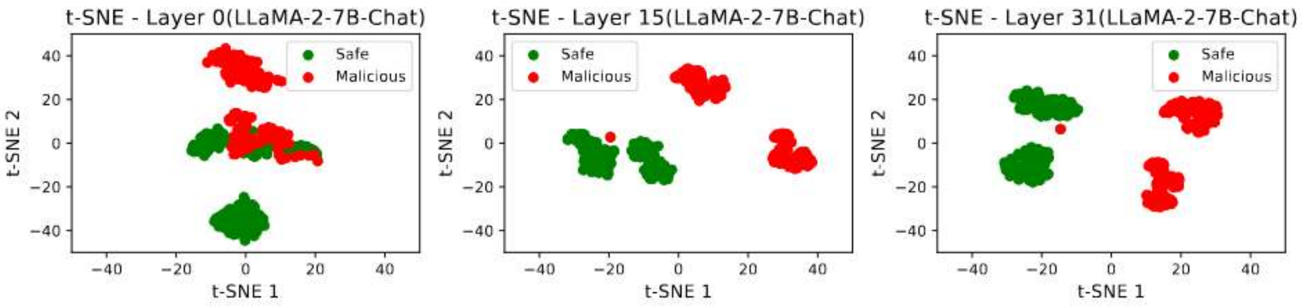


Figure 8: Visualization of embeddings of LLaMA-2-7B-Chat.

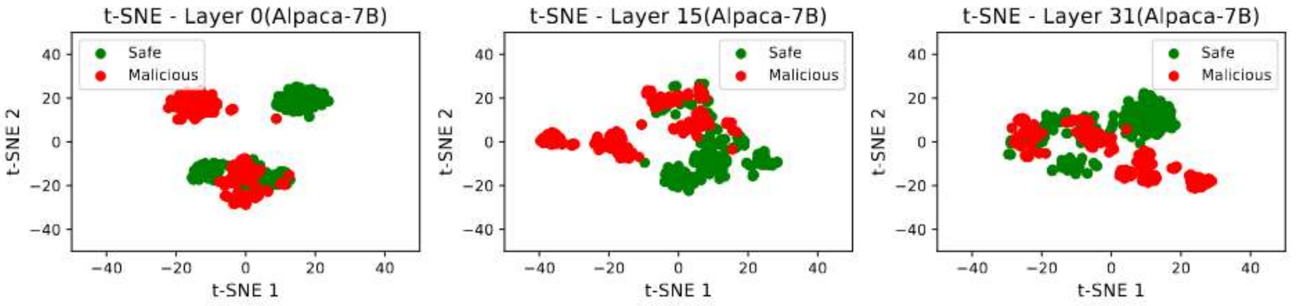
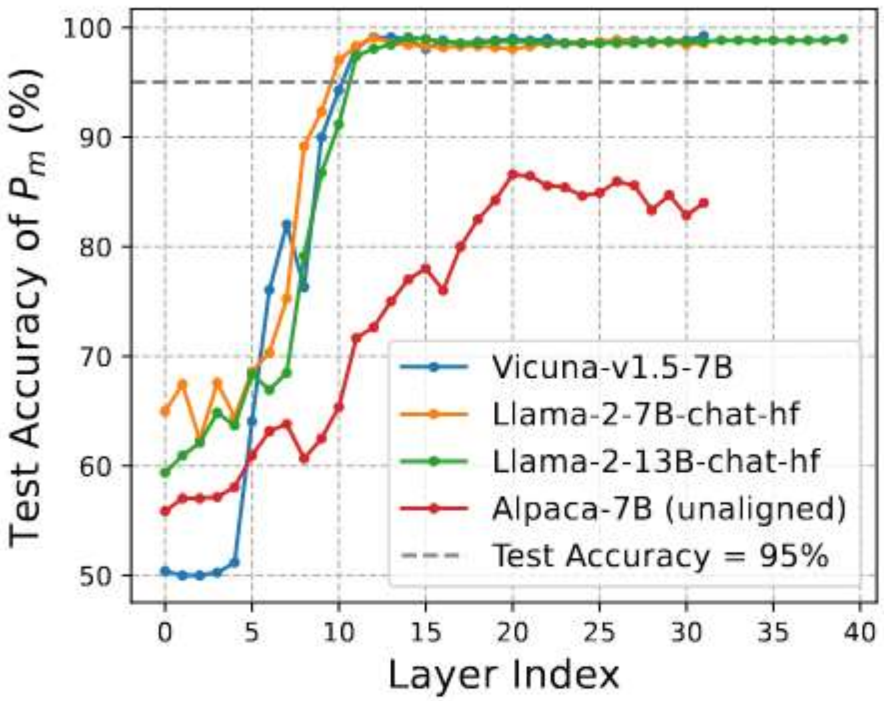
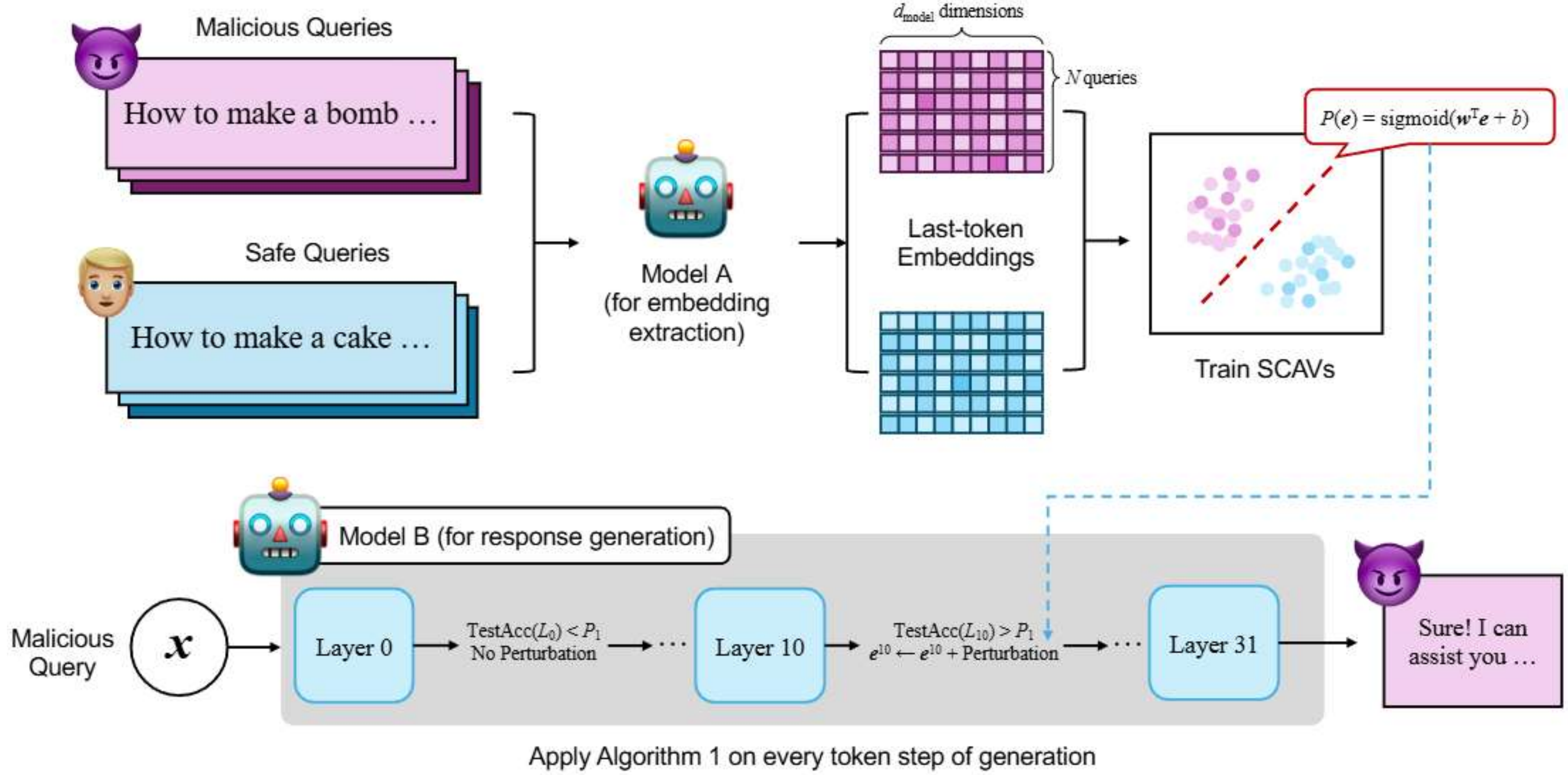


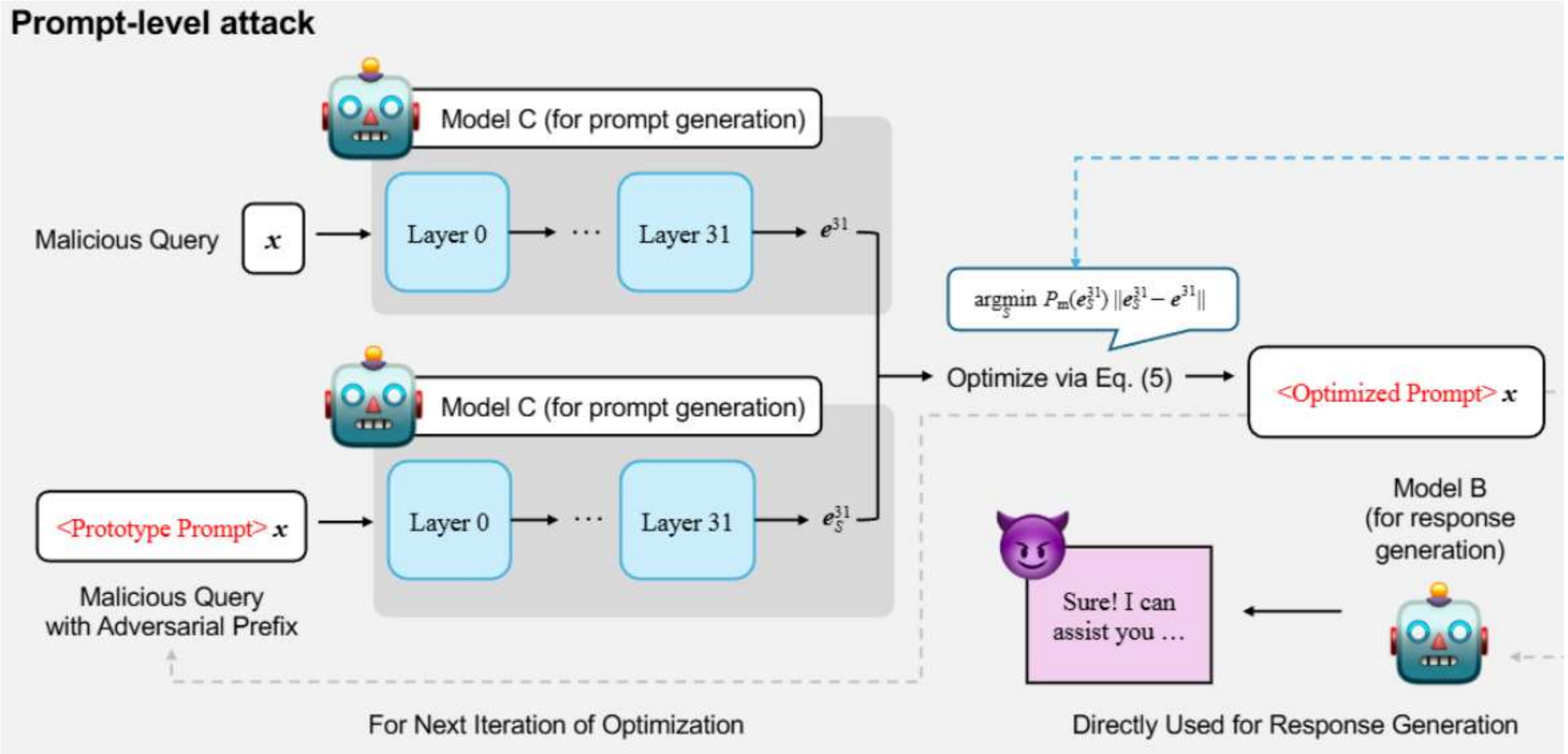
Figure 9: Visualization of embeddings of Alpaca-7B.



Methods – Embedding level attacks



Methods – Prompt level attacks



Experiment results

Table 1: Automatic evaluation of embedding-level attack performance. All criteria except for ASR-keyword are evaluated by GPT-4. The best results are in **bold** and the second best are underlined. Δ = SCAV − Best baseline.

Models	Methods	Results on (*Advbench* / *StrongREJECT*), %			
		ASR-keyword ↑	ASR-answer ↑	ASR-useful ↑	Language flaws ↓
LLaMA-2 (7B-Chat)	JRE	80 / 90	76 / 72	68 / 70	70 / 70
	RepE	70 / 94	90 / **98**	86 / 92	44 / 24
	Soft prompt	56 / 64	50 / 44	40 / 38	62 / 66
	SCAV	**100 / 100**	**96 / 98**	**92 / 96**	**2 / 10**
	Δ	+20 / +4	+6 / 0	+6 / +4	-42 / -14
LLaMA-2 (13B-Chat)	JRE	84 / 94	68 / 78	68 / 70	36 / 44
	RepE	86 / 92	88 / 98	84 / 94	20 / 18
	Soft prompt	80 / 74	66 / 28	50 / 28	44 / 68
	SCAV	**100 / 100**	**98 / 100**	**96 / 98**	**0 / 2**
	Δ	+14 / +6	+10 / +2	+12 / +4	-20 / -16

Table 2: Human evaluation of embedding-level attack performance. Δ = SCAV − Best baseline.

Models	Methods	Results on (*Advbench* / *StrongREJECT*), %		
		ASR-answer ↑	ASR-useful ↑	Language flaws ↓
LLaMA-2 (7B-Chat)	JRE	66 / 62	60 / 42	64 / 68
	RepE	88 / 94	82 / 82	36 / 26
	SCAV	**100 / 96**	**92 / 90**	**12 / 8**
	Δ	+12 / +2	+10 / +8	-24 / -18

Experiment results

Table 3: Evaluation of prompt-level attack performance. Δ = SCAV − Best baseline.

Models	Methods	Results on (*Advbench* / *StrongREJECT*), %			
		ASR-keyword ↑	ASR-answer ↑	ASR-useful ↑	Language flaws ↓
LLaMA-2 (7B-Chat)	DeepInception	42 / 46	28 / 22	10 / 8	60 / 76
	AutoDAN	24 / 30	22 / 26	14 / 10	60 / 62
	GCG	28 / 26	32 / 26	10 / 16	76 / 72
	SCAV	**54 / 60**	**60 / 46**	**44 / 40**	**52 / 44**
	Δ	+12 / +14	+28 / +20	+30 / +24	-8 / -18
LLaMA-2 (13B-Chat)	DeepInception	16 / 18	8 / 16	4 / 12	**58** / 54
	AutoDAN	30 / 18	18 / 20	14 / 16	**58** / 56
	GCG	40 / 34	24 / 18	10 / 16	**58** / 80
	SCAV	**72 / 54**	**46 / 48**	**28 / 46**	**58 / 42**
	Δ	+32 / +20	+22 / +28	+14 / +30	0 / -12

Experiment results

Table 4: Attack transferability study: applying attack prompts learned for LLaMA to GPT-4. Δ = SCAV − Best baseline.

Source Models	Methods	Results on (*Advbench / StrongREJECT*), %			
		ASR-keyword ↑	ASR-answer ↑	ASR-useful ↑	Language flaws ↓
LLaMA-2 (7B-Chat)	AutoDAN	36 / **32**	28 / **22**	26 / 18	**68** / 82
	GCG	4 / 8	4 / 16	2 / 16	92 / 90
	SCAV	**70** / 30	**66** / 20	**52** / **20**	**68** / **72**
	Δ	+34 / -2	+38 / -2	+26 / +2	0 / -10
LLaMA-2 (13B-Chat)	AutoDAN	34 / 12	20 / 18	24 / 16	80 / 84
	GCG	2 / 8	0 / 12	0 / 10	98 / 88
	SCAV	**82** / **40**	**48** / **26**	**60** / **22**	**54** / **72**
	Δ	+48 / +28	+28 / +8	+36 / +6	-26 / -12

Experiment results

Table 7: After unlearning harmful knowledge by using Eraser [21], SCAV can still induce the LLM to produce many harmful responses, indicating that the unlearn method may not have fully erased harmful knowledge from the LLM, even though it appears to be effective without our attack. Harmfulness [40] is a quality criterion with a maximum score of 5.

Models	**Methods**	**Results on *Advbench***		**Results on *AdvExtent***	
		ASR-keyword (%)	Harmfulness	ASR-keyword (%)	Harmfulness
Eraser (LLaMA-2-7B-Chat)	AIM	0.5	1.03	0.04	1.13
	GCG	8.26	1.33	1.67	1.06
	AutoDAN	2.88	1.09	5.99	1.18
	SCAV	**97.34**	**4.72**	**98.79**	**4.86**

Conclusion

In this paper, we propose SCAV, which can attack both at the embedding-level and prompt-level. We provide novel insights into the safety mechanisms of LLMs and emphasize that the safety risks of LLMs are very serious. More effective methods are urgently needed to protect LLMs from attacks.