SCANS: Mitigating the Exaggerated Safety for LLMs via Safety-Conscious Activation Steering

AAAI-25
FEBRUARY 25 – MARCH 4, 2025
PHILADELPHIA, USA

Zouying Cao[1,2,3], Yifei Yang[1,2,3], Hai Zhao[1,2,3,*]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University
[3]Shanghai Key Laboratory of Trusted Data Circulation and Governance in Web3

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Introduction

- We propose a training-free, representation engineering method named **SCANS** (**S**afety-**C**onscious **A**ctivatio**n** **S**teering), which utilizes refusal behavior vectors to steer the model output in safety-critical layers.
- We discover the extracted refusal steering vectors from middle layers promote refusal tokens (e.g., cannot) and thus steering the corresponding representation can reduce the false refusal rate.
- Our SCANS can effectively mitigate the exaggerated safety in aligned LLMs, without undermining the adequate safety and general capability. Specifically, SCANS reduces the average false refusal rate by 24.7% and 26.3% on XSTest and OKTest benchmarks.

Motivation

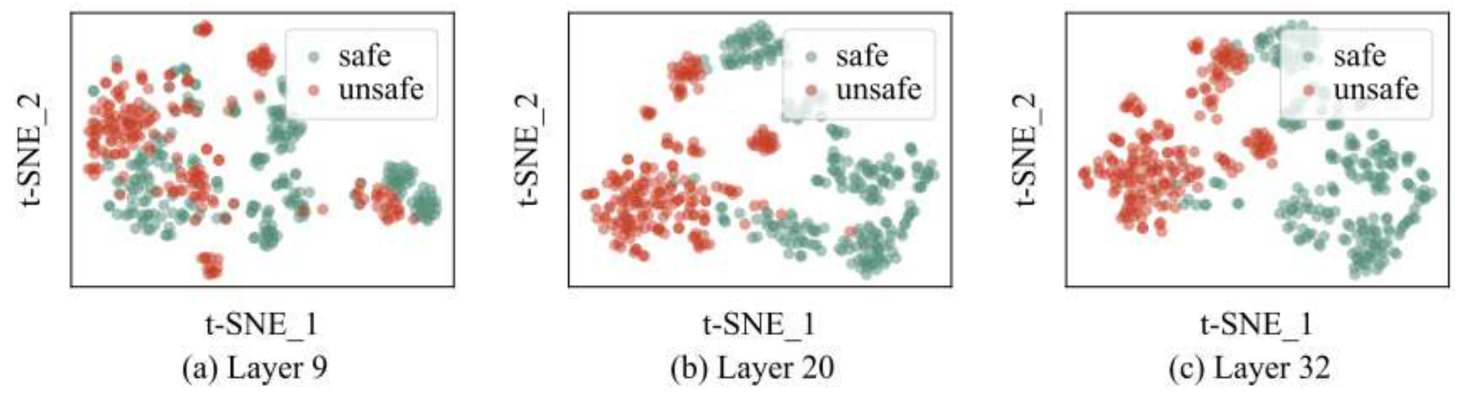


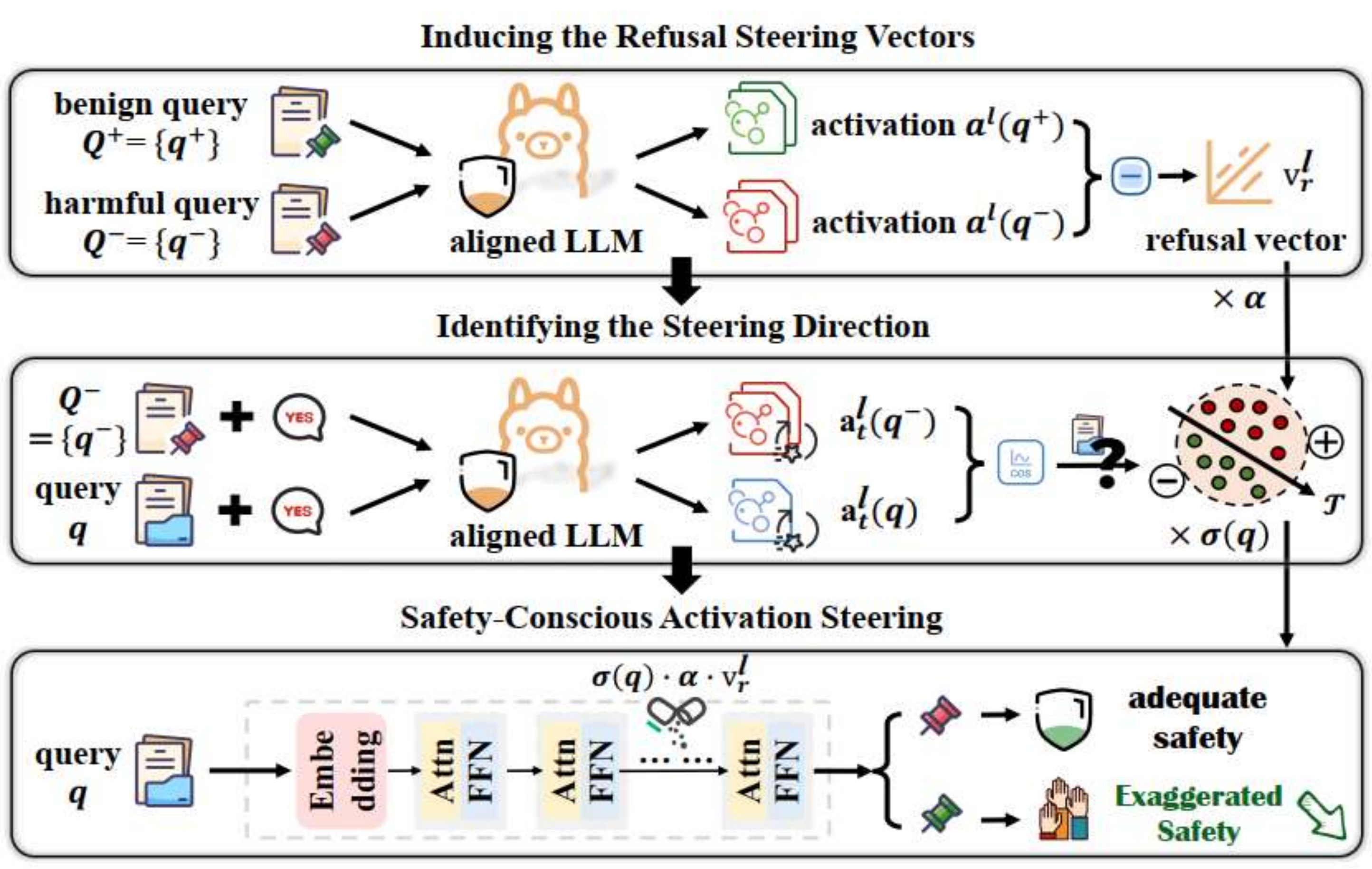
Figure 1: t-SNE visualization of hidden state transition of Llama2-7b-chat on XSTest dataset. Results indicate safety-related representation clustering emerges in middle and latter layers.

Layers	Top-10 tokens
Former Layers (0-9)	*_einges, _schließ, vue, ché, orio, _Syd, rugu, wrap, widet, axi*
Middle Layers (10-20)	*_rejected, _impossible, zas, _cons, ball, od, lio, _tur, _reject, _cannot*
Latter Layers (21-31)	*sey, _Mas, _Coun, _Ir, _ext, _properties, _Seg, ber, ds, _sa*

Table 1: Top-10 tokens associated with steering direction at different layers of Llama2-7b-chat. We highlight the tokens related to refusal behavior with an underline.

Methodology

Motivated by the intuition of representation engineering to steer model behavior, the key idea behind our **SCANS** is to extract the refusal behavior vectors, and anchor the safety-critical layers for steering. SCANS then evaluates the harmfulness of inputs to guide output distribution against or consistent with the refusal behavior, which achieves a balance between adequate safety and exaggerated safety.



Algorithm 1: Workflow of **SCANS**

Input: Safety-aligned LLM $\mathcal{M}$, Steering multiplier α, Set of steering layers $[L_l, L_H]$, Anchor data $Q=\{Q^-, Q^+\}$, Designed positive response r_{pos}, Hyperparameter $\mathcal{T}$, $\mathcal{L}$ for classification, Input queries $\{q\}$
Output: The steered outputs (safe and helpful)

`// Inducing the Refusal Steering Vectors`
1 $v_r \leftarrow \emptyset$;
2 For each query $q \in Q$, collect the hidden states $a^l(q)$ for each layer l at the last token position.
3 **for** $l \leftarrow L_l$ **to** L_H **do**
4   Compute v_r^l using Eq. 1;
5   $v_r \leftarrow v_r \cup \{v_r^l\}$;
`// Identifying the Steering Direction`
6 **for** $q \in Q^-$ **do**
7   $q' \leftarrow concat(q, r_{pos})$;
8   Input q', collect two hidden states, one a_p from the last token of the query part and the other a_e from the final token of the entire input.
9   Compute $a_t(q) = \{a_t^l(q)\}_{l\in\mathcal{L}}$ using Eq. 2;
10 For queries $\{q\}$, repeat line 7-9 to get the hidden state transition and then compute s_q using Eq. 4;
11 **if** $s_q < \mathcal{T}$ **then**
12   $\sigma(q) \leftarrow -1$ `/* query q is safe */`
13 **else**
14   $\sigma(q) \leftarrow 1$ `/* query q is unsafe */`
`// Safety-Conscious Activation Steering (During inference)`
15 Input queries $\{q\}$ to $\mathcal{M}$, each layer l will output the corresponding hidden states.
16 **if** $l \in [L_l, L_H]$ **then**
17   Steer the hidden states $a^l(q)$ at the last token position
18   towards $\tilde{a}^l(q) = a^l(q) + \sigma(q)\cdot\alpha\cdot v_r^l$;
19 **return** the steered outputs after activation steering.

Experiments-Main Results

① SCANS effectively achieves a balance between exaggerated safety mitigation and adequate safety.

Models	Methods	XSTest Safe↓	XSTest UnSafe↑	XSTest Avg.↑	RepE-Data Safe↓	RepE-Data UnSafe↑	RepE-Data Avg.↑	Helpfulness↓ OKTest	Helpfulness↓ TQA	Harmfulness↑ AdvBench	Harmfulness↑ Malicious	Avg.↑
Llama2-7b-chat	Default	58.00	100.0	67.77	12.50	100.0	93.75	53.67	5.05	**100.0**	**100.0**	86.13
	Prompt	36.40	100.0	79.77	2.86	99.48	98.31	41.66	15.27	99.34	**100.0**	87.72
	Self-CD*	14.80	97.50	90.66	1.30	98.17	98.43	17.33	4.51	98.24	98.00	94.69
	SafeDecoding	75.60	99.50	57.77	63.80	100.0	68.10	59.33	54.44	**100.0**	**100.0**	63.81
	DRO	41.52	98.40	76.22	7.03	99.48	96.22	32.33	16.20	99.60	99.56	87.36
	SCANS	9.20	93.50	**92.00**	0.00	99.22	**99.61**	**0.33**	**0.80**	99.34	**100.0**	**98.26**
Llama2-13b-chat	Default	34.40	99.50	80.66	5.73	100.0	97.14	20.33	11.69	**99.78**	**100.0**	90.83
	Prompt	18.00	99.50	89.77	0.78	99.22	99.22	30.33	12.62	99.34	**100.0**	91.47
	Self-CD*	29.60	100.0	83.55	4.68	100.0	97.66	19.33	4.91	98.24	**100.0**	93.10
	DRO	38.00	100.0	78.88	6.51	100.0	96.74	23.66	14.20	**99.78**	**100.0**	89.42
	SCANS	7.20	97.50	**94.89**	0.00	98.96	**99.48**	**0.33**	**1.20**	98.90	97.00	**98.40**
vicuna-7b-v1.5	Default	20.80	88.00	83.11	4.69	97.40	96.36	19.00	5.05	97.37	76.00	91.68
	Prompt	22.00	91.00	83.77	6.51	98.44	95.97	22.67	11.33	98.46	82.00	90.01
	Self-CD*	10.00	83.00	86.88	3.64	89.58	92.97	27.00	9.56	89.03	56.00	87.26
	SafeDecoding	55.20	99.50	69.11	33.29	100.0	83.35	61.00	39.70	**100.0**	98.00	73.41
	DRO	22.11	95.80	85.85	3.38	99.74	**98.18**	13.33	6.77	98.90	**99.00**	93.82
	SCANS	5.60	87.00	**91.11**	2.08	95.83	96.88	**3.00**	**0.00**	98.96	98.00	**97.17**
vicuna-13b-v1.5	Default	16.80	98.00	89.77	3.65	98.96	97.66	19.33	4.38	**99.78**	93.00	94.23
	Prompt	20.80	99.00	88.00	10.68	99.74	94.53	27.00	19.33	99.34	97.00	88.37
	Self-CD*	8.40	90.50	91.11	2.60	90.88	94.14	26.67	6.64	90.57	81.00	90.20
	DRO	29.20	99.00	83.33	3.38	99.73	**98.17**	23.33	13.94	99.34	**99.00**	90.52
	SCANS	9.20	93.50	**92.00**	2.08	97.66	97.79	**3.33**	**0.27**	**99.78**	98.00	**97.59**

② SCANS does not compromise the general model capability greatly.

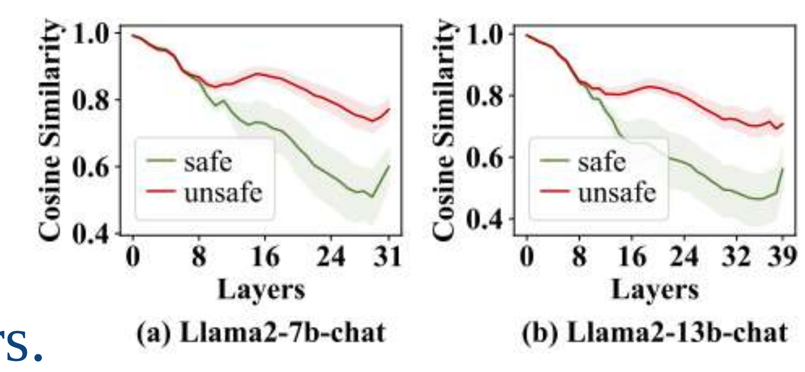
Models	Perplexity↓ WikiText2	Perplexity↓ C4	XSum↑ R-1	XSum↑ R-2	XSum↑ R-L	MMLU↑ STEM	MMLU↑ Human	MMLU↑ Social	MMLU↑ Others	MMLU↑ Avg.
Llama2-7b-chat	7.76	9.86	21.38	4.923	17.45	37.60	43.40	55.10	54.10	47.20
+SCANS	9.32	11.94	20.07	3.912	16.47	34.00	36.20	47.40	46.20	40.50
Llama2-13b-chat	6.86	8.89	22.22	5.280	17.48	43.80	49.50	62.50	60.00	53.60
+SCANS	7.29	9.45	21.20	4.277	16.79	43.10	49.20	61.80	59.40	53.00
vicuna-7b-v1.5	7.34	9.26	20.85	4.557	17.34	39.50	45.80	58.20	57.50	49.90
+SCANS	11.53	15.32	18.43	3.440	15.69	36.60	43.40	54.40	54.20	46.80
vicuna-13b-v1.5	6.37	8.35	21.88	5.51	18.20	45.00	52.00	65.20	62.50	55.80
+SCANS	7.07	9.20	20.40	4.484	16.48	44.20	51.20	64.10	61.80	55.00

More Analysis

③ SCANS requires minor extra cost in inference time and GPU memory.

	Inference Speed	GPU Memory
Llama2-7b-chat	40.60 tokens/s	29324MB
+SCANS	39.62 tokens/s	29694MB

④ Middle and latter layers demonstrate higher degree of distinction, indicating better identification accuracy for harmfulness.



⑤ Effect of Steering Layers.

	Perplexity↓ WikiText2	Perplexity↓ C4	XSTest Safe↓	XSTest Unsafe↑	XSTest Avg.↑	Helpfulness↑ OKTest	Helpfulness↑ TruthfulQA	Harmfulness↑ AdvBench	Harmfulness↑ Malicious	Avg.↑
Llama2-7b-chat										
Former Layers	2946	3058	-	-	-	-	-	-	-	-
Middle Layers	9.32	11.94	**9.20**	93.50	**92.00**	**0.33**	0.80	**99.34**	**100.0**	**97.76**
Latter Layers	8.15	10.37	12.00	**95.00**	91.11	7.00	**0.27**	98.90	98.00	96.59
vicuna-7b-v1.5										
Former Layers	15433	11457	-	-	-	-	-	-	-	-
Middle Layers	11.53	15.32	**5.60**	**87.00**	**91.11**	3.00	**0.00**	**98.96**	**98.00**	**97.29**
Latter Layers	7.85	9.89	7.60	83.50	88.44	**2.33**	1.46	93.42	92.00	94.75

Conclusion

- Mitigate the exaggerated safety for aligned LLMs via activation steering in safety-critical layers
- Training-free, Effective!

—— SCANS

- **Contact us:**
  zouyingcao@sjtu.edu.cn
  yifeiyang@sjtu.edu.cn
  zhaohai@cs.sjtu.edu.cn
- **Full paper**