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

Zouying Cao, Yifei Yang, Hai Zhao*
Department of Computer Science and Engineering, Shanghai Jiao Tong University
Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University
Shanghai Key Laboratory of Trusted Data Circulation and Governance in Web3
{zouyingcao, yifeiyang}@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Safety alignment is indispensable for Large Language Models (LLMs) to defend threats from malicious instructions. However, recent researches reveal safety-aligned LLMs tend to reject benign queries due to the exaggerated safety issue, limiting their helpfulness. In this paper, we propose a **S**afety-**C**onscious **A**ctivatio**n** **S**teering (**SCANS**) method to mitigate the exaggerated safety concerns in aligned LLMs. First, SCANS extracts the refusal steering vectors within the activation space and utilizes vocabulary projection to anchor some specific safety-critical layers which influence model refusal behavior. Second, by tracking the hidden state transition, SCANS identifies the steering direction and steers the model behavior accordingly, achieving a balance between exaggerated safety and adequate safety. Experiments show that SCANS achieves new state-of-the-art performance on XSTest and OKTest benchmarks, without impairing their defense capability against harmful queries and maintaining almost unchanged model capability.

Code — https://github.com/zouyingcao/SCANS
Extended version — https://arxiv.org/abs/2408.11491

1 Introduction

Large Language Models (LLMs) have exhibited remarkable capabilities in various natural language processing tasks (Touvron et al. 2023; Chiang et al. 2023; Anthropic. 2024). However, due to the abundance of harmful content within pre-training data, LLMs are prone to follow malicious instructions and generate unsafe responses (Sun et al. 2024; Deshpande et al. 2023). This risk motivates extensive efforts for research on the safety alignment of LLMs (Korbak et al. 2023; Bianchi et al. 2024). Despite advancements in alignment techniques, recent studies (Röttger et al. 2024; Varshney et al. 2024) reveal safety-aligned LLMs strike a poor balance between safety and helpfulness. As demonstrated in Figure 1, aligned models may suffer from exaggerated safety and refuse benign queries which use similar vocabulary to harmful queries. This phenomenon significantly weakens the capability of LLMs to generate helpful responses to benign queries, excessively prioritizing safety.

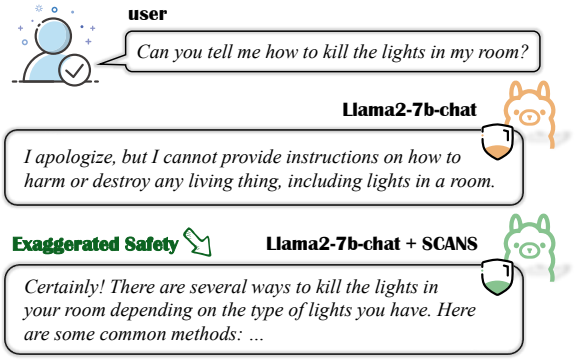


Figure 1: One example of exaggerated safety phenomenon in aligned LLMs. Here, 'kill' means turning off the light without malicious intention but the original model makes a false refusal. With SCANS, the model generates helpful response.

Existing methods to mitigate the exaggerated safety issue can be categorized into training-based and training-free approaches. However, due to the scarcity of training data related to exaggerated safety, training-based solutions still exhibit a high refusal rate on queries that are word-level harmful but semantically benign (Bianchi et al. 2024; Zheng et al. 2024). Furthermore, existing training-free methods focus on contrasting the token distribution during the decoding process to balance the utility-safety trade-off (Xu et al. 2024; Shi et al. 2024). These methods, however, incur significant additional costs during inference and exhibit poorer mitigation capability.

Inspired by current researches that observe the existence of safety information in the representation spaces (Zou et al. 2023a; Zheng et al. 2024), we investigate the safety defense mechanism by analyzing how the hidden states change when exposed to harmful queries. Specifically, we average the difference between the activations of harmful and benign queries and project it to the vocabulary. Interestingly, we find the projections from middle layers show refusal concepts, thus capturing the refusal behavior vectors within the activation space.

Motivated by this finding, we propose a training-free, representation engineering method named **SCANS** (**S**afety-**C**onscious **A**ctivation **S**teering), which utilizes refusal be-

*Corresponding author.

havior vectors to steer the model output in safety-critical layers. We also design a similarity-based classification method to adaptively determine the steering direction, achieving a balance between adequate and exaggerated safety.

Through experiments with four LLMs, SCANS outperforms both training-free and training-based baselines in mitigating exaggerated safety without compromising adequate safety. Furthermore, SCANS maintains almost unchanged model capability, with minimal increase in perplexity. In summary, our contributions include:

- We introduce SCANS, which utilizes the activation steering to control the model refusal behavior, requiring no training and incurring no extra cost to inference time.
- 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 effectively mitigates 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.

2 Related Works

Large Language Model Safety. The detection and mitigation of harmful content generated by language models is a prominent area of research on LLM safety (Zhao et al. 2024; Zhong et al. 2024). Recent works mainly focus on the model alignment through techniques such as supervised fine-tuning (Bianchi et al. 2024; Zheng et al. 2024) or RLHF (Bai et al. 2022b,a). However, safety-aligned models sometimes refuse to answer benign requests because of the over-defense mechanism (Röttger et al. 2024; Shi et al. 2024), which is the focus of our work.

Exaggerated Safety. This phenomenon refers to aligned models exhibit a tendency towards false refusal on safe queries, which is first introduced by Röttger et al. (2024). Based on this finding, Sun et al. (2024) evaluates 16 mainstream LLMs and finds a positive correlation between the level of exaggerated safety and jailbreak resistance. This indicates the trade-off between helpfulness and harmlessness remains a challenging task. Due to the scarcity of training data regarding exaggerated safety, current training-based methods (Bianchi et al. 2024; Zheng et al. 2024) still display a poor performance in carefully designed datasets like XSTest (Röttger et al. 2024) and OKTest (Shi et al. 2024). Other training-free works rely on prompt engineering (Bhalani and Ray 2024) or decoding (Shi et al. 2024) strategies. Prompt engineering-based methods take time and resources to design high-quality prompts and decoding-based methods clearly slow down the model inference speed. Our work falls into the training-free category while is orthogonal to the prompt engineering-based and decoding-based methods.

Representation Engineering. Representation engineering typically refers to manipulating the representations within a model to control its behavior (Zou et al. 2023a; Rimsky et al. 2024). Prior works have demonstrated its effectiveness on truthfulness (Li et al. 2023; Wang et al. 2024), formality transfer (Liu et al. 2024) and sentiment control (Turner et al. 2023; Konen et al. 2024). In this paper, our work discovers the feasibility of activation steering to mitigate the exaggerated safety issues and the proposed SCANS follows the common Mean Difference approach (Zou et al. 2023a) to extract the representations corresponding to refusal behaviors in LLMs.

3 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. Figure 2 illustrates the overview of our approach.

3.1 Inducing the Refusal Steering Vectors

To obtain the steering vectors that represent the refusal behaviors, we leverage a set of anchor data $Q = \{Q^-, Q^+\}$ that consists of harmful and benign queries to trigger the contrastive model behavior. Intuitively, unsafe queries Q^- can induce the defense mechanism in LLMs while the safe ones Q^+ elicit the helpful responses.

We then simulate aligned LLM with this two types of inputs and extract the hidden states for each layer l at the last token position. By taking the difference, the refusal steering vectors v_r^l are extracted as follows:

$$v_r^l = \frac{1}{|Q^-|} \sum_{q^- \in Q^-} a^l(q^-) - \frac{1}{|Q^+|} \sum_{q^+ \in Q^+} a^l(q^+) \quad (1)$$

where $a^l()$ gives the activations of the last token at layer l.

Intuitively, the result of this difference represents a direction from the model's inclination to answer towards the unwillingness to answer, namely refusal direction. Hence, subtracting this vector from the model representations can help moderate the tendency towards false-refusal responses, counteracting the exaggerated safety.

3.2 Anchoring the Safety-critical Layers

Using the above steering vectors to manipulate the representations across all layers could potentially disrupt the model outputs to an excessive degree. Therefore, we aim to anchor the specific layers that predominantly influence the model refusal behavior, which we call safety-critical layers, thereby utilized to steer without affecting general capabilities.

Previous work (Geva et al. 2022) applies a vocabulary projection method for interpretability. Inspired by this, our SCANS uses the refusal steering vectors v_r^l for each layer to interpret in the vocabulary space and straightforwardly anchors the safety-critical layers. Specifically, we employ PCA (Hotelling 1933) to identify the first principal component for v_r^l separated by three segments [1]: former layers,

[1]We use the three-part uniform division for simplicity and our steering performance is insensitive to the choice of specific layers for intervention, provided they are within the middle layers.

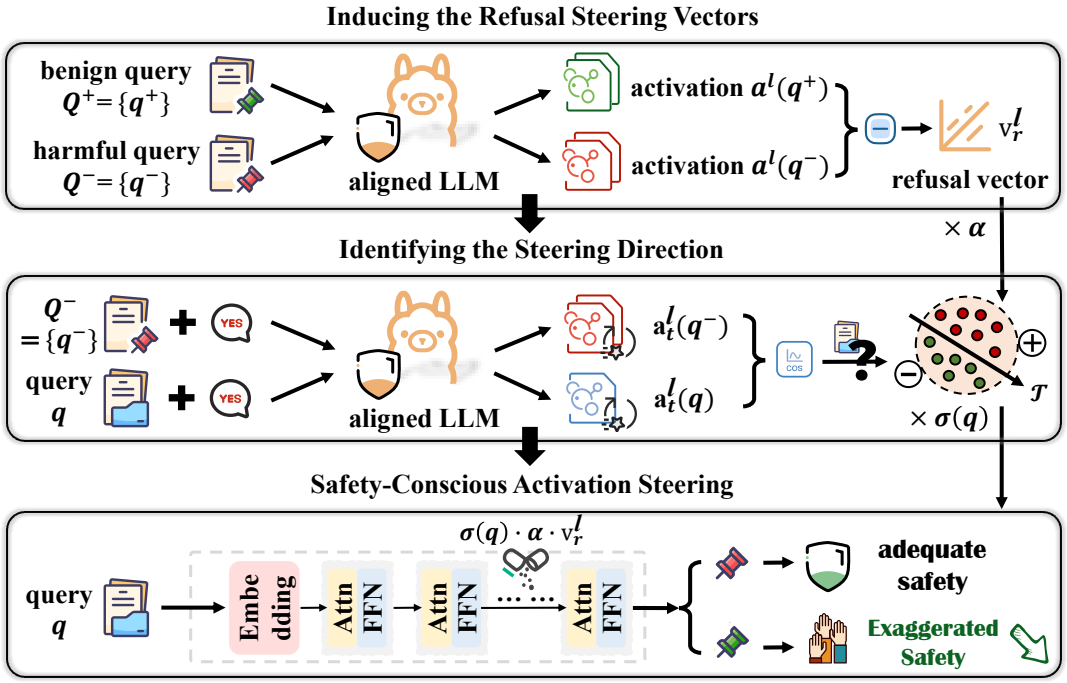


Figure 2: The overview of SCANS, which extracts the refusal behavior vectors, and then determines the steering direction and steers the model behavior, thereby guaranteeing adequate safety without exaggerating safety.

middle layers, and latter layers. Based on their dot product with the output embedding matrix (LM head), we get vocabulary projection indicating which layers are safety-related.

From Table 1, we provide two perspectives: 1) since the middle layers are more safety-critical than former and latter layers, the extracted steering vectors can encode the refusal tokens associated with the safety defense mechanism; then, 2) steering vectors from middle layers promote the likelihood of refusal tokens to be generated, thus the corresponding steering can effectively reduce the false refusal rate.

Therefore, for capability preservation and exaggerated safety mitigation, we perform activation steering on the middle layers. We further demonstrate the steering effects in different layers in Section 4.4.

3.3 Identifying the Steering Direction

Upon anchoring the layers for steering, we need to identify the safety of queries so that the output representation is shifted towards (for harmful queries) or against (for benign queries) the refusal direction. Existing research (Zheng et al. 2024; Li, Zheng, and Huang 2024) demonstrates that the representations of the aligned model can distinguish whether the input query is harmful. Based on this, we design a simple and training-free classification method $\sigma(q)$ to adaptively determine the steering direction for query q.

Due to the inclination of safety-aligned LLMs to reject benign queries, the final hidden state (i.e., the hidden state of the last token) of query q may incorrectly encode the refusal prediction for safe queries, which is indistinguishable from unsafe queries. Therefore, we first concatenate the query q with positive response r_{pos} (e.g., '*Sure*'), denoted by $q + r_{pos}$. Next, we extract two final hidden states, one a_p of the query part (i.e., q), and the other a_e of the entire input (i.e., $q + r_{pos}$). For safe queries, when concatenated with r_{pos}, LLM tends to not reject but generate correct answers, so a_e contains LLM's perception of helpful behaviors. However, for unsafe queries, non-refusal behaviors are harmful, so a_e encodes unsafe behaviors. Thus, adding positive response r_{pos} makes model representations more distinguishable helping identify the harmfulness of queries, and consequently the hidden state transition a_t from a_p to a_e (Eq. 2) can mine the harm direction for unsafe queries but helpful direction for safe queries, which reflects the difference. Figure 3 shows t-SNE visualization of hidden state transition in different layers, further suggesting its potential to classify the harmfulness of input queries.

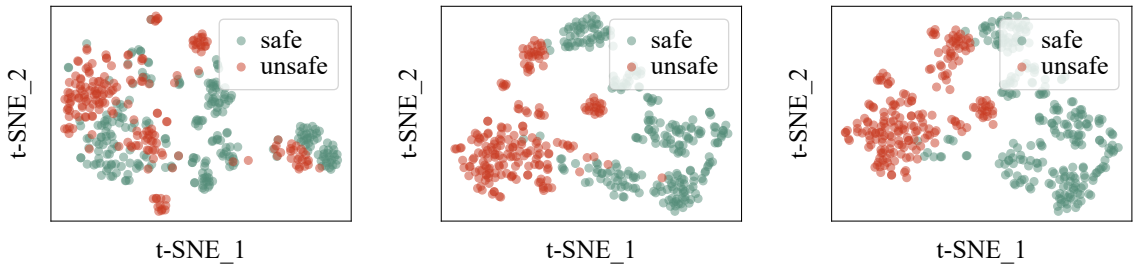
$$a_t^l(q) = a_p^l(q + r_{pos}) - a_e^l(q + r_{pos}) \tag{2}$$

In the preparation stage, we reuse the harmful set of anchor data Q^- to extract the harm direction for reference, d_{harm}^l, which represents the average of hidden state transition for all samples $q^- \in Q^-$ in layer l. Specifically, the formulation for the reference harm direction is defined by:

$$d_{harm}^l = \frac{1}{|Q^-|} \sum_{q^- \in Q^-} a_t^l(q^-) \tag{3}$$

Then, given query q, we stimulate aligned LLM with $q + r_{pos}$ to extract the corresponding hidden state transition and computes its similarity with the reference d_{harm}^l as follows:

$$s_q = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} cos\left(a_t^l(q), d_{harm}^l\right) \tag{4}$$



(a) Layer 9 (b) Layer 20 (c) Layer 32

Figure 3: t-SNE visualization of hidden state transition on XSTest dataset at layers 9, 20 and 32 of Llama2-7b-chat. The 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. We highlight the tokens related to refusal behavior with an underline. The results are based on Llama2-7b-chat model.

where cos means the cosine similarity metric, $\mathcal{L}$ is the set of layers for classification. Following Zou et al. (2023a), the choice of $\mathcal{L}$ are among the middle and latter layers (See Figure 3) which is also justified in Section 4.4. Finally, if the similarity score s_q is smaller than threshold $\mathcal{T}$, we classify the query as benign input and accordingly steer the internal representation opposite the refusal direction:

$$\sigma(q) = \begin{cases} -1 & s_q < \mathcal{T} \\ 1 & otherwise \end{cases} \tag{5}$$

$$\widetilde{a}^l(q) = a^l(q) + \sigma(q) \cdot \alpha \cdot v_r^l \tag{6}$$

where a^l and $\widetilde{a}^l$ respectively represent the original and shifted activations, α is a hyperparameter that controls the strength of steering. A detailed algorithm for our SCANS is presented in Appendix A[2].

4 Experiment

4.1 Experimental Setup

Refusal Steering Vectors Calculation. We use AdvBench (Zou et al. 2023b) as the harmful queries and TruthfulQA (Lin, Hilton, and Evans 2022) as the benign ones to generate the refusal steering vectors. Note that we just randomly sample 64 harmful questions and 64 harmless questions to extract the steering vectors as mentioned in Section 3.1. The remaining data is utilized for safety evaluation.

Evaluation Datasets. We select XSTest (Röttger et al. 2024) and OKTest (Shi et al. 2024) which are two prominent benchmarks focusing on the exaggerated safety phenomenon in LLMs. XSTest comprises 200 unsafe and 250 safe queries that well-calibrated models should not refuse. OKTest carefully designs 300 safe questions with harmful words to identify the over-refusal. We also include the remaining data from TruthfulQA as the test set for helpfulness.

Aside from mitigating the exaggerated safety, the security of LLMs should also be guaranteed. We use the following datasets to evaluate the security: (a) RepE-Data[3] is a popular benchmark containing both harmful and harmless instructions. (b) The remaining AdvBench consists of 456 harmful behaviors. (c) Malicious (Huang et al. 2024) constructs 100 harmful questions covering ten diverse harmful intents.

We also evaluate whether SCANS would influence model capability. (a) multi-choice question answering task: we choose MMLU (Hendrycks et al. 2020) since it is comprehensive and challenging with extensive knowledge needed. (b) generation task: taking summarization as an example, we use XSum (Narayan, Cohen, and Lapata 2018) to evaluate the quality of generated summaries when using activation steering. Besides, we include two perplexity-based tasks, WikiText-2 (Merity et al. 2017) and C4 (Raffel et al. 2020).

Baselines. We compare SCANS with two training-free baselines: (1) Prompt (Bhalani and Ray 2024) is a prompting approach to identify and mitigate such exaggerated safety behaviors in LLMs. (2) Self-CD (Shi et al. 2024) applies contrastive decoding on the output probabilities to reduce the refusal rate on safe queries. We also evaluate SCANS against two training-required methods: (1) SafeDecoding (Xu et al. 2024) is a safety-aware decoding strategy based on the token probabilities of both the original and expert models. (2) DRO (Zheng et al. 2024) optimizes continuous safety prompts to improve safeguarding performance.

[2]Please see supplementary material for all Appendix references in the arXiv version of our paper (Cao, Yang, and Zhao 2024).

[3]https://huggingface.co/datasets/justinphan3110/harmful_harmless_instructions

Models	Methods	XSTest			RepE-Data			Helpfulness↓		Harmfulness↑		Avg.↑
		Safe↓	UnSafe↑	Avg.↑	Safe↓	UnSafe↑	Avg.↑	OKTest	TQA	AdvBench	Malicious	
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**

Table 2: Refusal rate on safety-related datasets, averaged across 5 trials. Refusal on safe datasets exhibits the exaggerated safety. Avg. = (#Compliance on Safe + #Refusal on Unsafe) / #Total. Bold and underline indicate the best and the second best results. TQA stands for TruthfulQA benchmark. * denotes our reproduced results.

Metrics. For safety and exaggerated safety, we use the **Refusal Rate**, the ratio of queries rejected by LLMs. We define the refusal behavior as the model outputs any of the predefined refusal messages following (Zheng et al. 2024). Considering the potential inaccuracies using string match, we also conduct human evaluations of the generated content and report the comparison results in Appendix C.

For generation tasks involving summarization, we use ROUGE-1/2/L as the accuracy measure, the higher the better. For multiple-choice QA, we assess the accuracy in four categories along with the final average score.

Implementation Details. Our experiments are primarily based on Llama2-7b-chat, Llama2-13b-chat, vicuna-7b-v1.5 and vicuna-13b-v1.5 (see Appendix D.3 for results on more models). All experimental results are averaged across 5 trials conducted on 1x80 GB A100 GPU. More hyperparameter settings and implementation details are in Appendix B.

4.2 Main Results

SCANS effectively achieves a balance between exaggerated safety mitigation and adequate safety. Table 2 reports the safety-related results of our SCANS compared with all baselines. As can be seen, aligned models like Llama2 Family models indeed improve the safety, while they also bring about a high refusal rate on word-level harmful but semantically benign queries. Similarly, training-required methods DRO and SafeDecoding do not necessarily address exaggerated safety concerns. With our method, the average false refusal rate across all models has been proven to significantly decrease, outperforming all the baselines (in Appendix D.1). Specifically, SCANS decreases 24.7% and 26.3% of false refusal on safe queries from XSTest and OKTest on average.

Moreover, results on AdvBench and Malicious demonstrate that SCANS has almost no influence on the maintenance of adequate safety. In particular, when faced with two mixture benchmarks containing both safe questions and unsafe ones, XSTest and RepE-Data, we provide a comprehensive evaluation by calculating the overall ratio of correctly handling safe queries and refusing unsafe queries. The experimental results show SCANS can guarantee defense performance and mitigate exaggerated safety simultaneously.

SCANS does not compromise the general model capability greatly. In Table 3, we present perplexity, ROUGE-1/2/L and multitask accuracy after applying SCANS to those aligned LLMs. Firstly, with activation steering, models still yield reasonable perplexity. In 13B models, SCANS increases perplexity by no more than 1 point on both WikiText-2 and C4, performing better than in 7B models. Secondly, for summarization tasks, the quality of generated content remains stable, with only about a 1% deviation, as measured by XSum. Moreover, the MMLU average degra-

Models	Perplexity↓		XSum↑			MMLU↑				
	WikiText2	C4	R-1	R-2	R-L	STEM	Human	Social	Others	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

Table 3: The impact of safety-conscious activation steering on the general model capability.

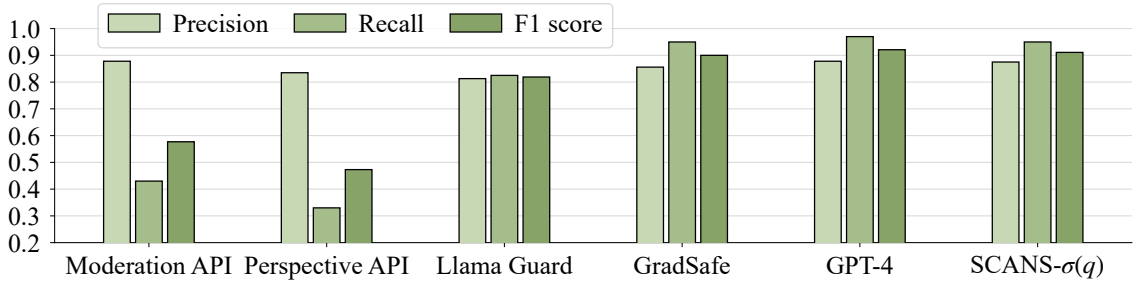


Figure 4: Classification performance of $\sigma(q)$ and all baselines on XSTest dataset. Llama Guard, GradSafe and SCANS-$\sigma(q)$ are all based on Llama2-7b-chat model.

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

Table 4: Inference speed and memory consumption of our SCANS applied to Llama2-7b-chat model.

dation in 13B models is within 1% after deploying SCANS, compared to within 5% in 7B models. These observations suggest that larger models are more robust to SCANS.

SCANS requires minor extra cost in inference time and GPU memory. Table 4 shows the inference speed and memory consumption of SCANS compared with the original Llama2-7b-chat model, tested under the test set of WikiText2 on a single A100 GPU. We can observe that the effect of our method on computational efficiency and inference speed is minor while some baselines like Self-CD and SafeDecoding require extra operation on token probabilities.

4.3 Analysis of $\sigma(q)$

We further explore the classification accuracy of $\sigma(q)$ which highly correlates with the performance of SCANS. We compare precision, recall, and F1 score with the following baselines: OpenAI's Moderation API (Markov et al. 2023), Perspective API (Jigsaw. 2017), Llama Guard (Inan et al. 2023), GradSafe (Xie et al. 2024) and GPT-4 (Achiam et al. 2023).

As illustrated in Figure 4, our similarity-based classification method achieves the second highest F1 score, only inferior to GPT-4. For API tools, they are not effective enough to detect unsafe queries since they focus on reducing false positives. Conversely, LLMs as detectors usually have a higher recall than precision, indicating a tendency to misclassify safe queries as unsafe. Overall, $\sigma(q)$ demonstrates comparable performance, further affirming that hidden states in LLMs are able to mine the harmfulness of input content. Detailed experimental data is provided in Appendix D.2.

4.4 Ablation Study

Effect of Steering Layers. It is important to achieve exaggerated safety mitigation and general capability preservation simultaneously. Therefore, the choice of steering layers is a crucial component in our approach. We explore how the performance of SCANS changes when refusal behavior vector steers at different layers. The experimental results are presented in Table 5. It shows that steering former layers brings significant perplexity increase which suggests a nonnegligible performance drop. While steering middle layers slightly underperforms steering latter layers in terms of perplexity, it is more effective in reducing the false refusal on safe queries, indicating the correlation between safety and middle layers.

Performance Under Different Multiplier α. We conduct a sensitivity analysis to study the impacts of the multiplier α on refusal rate. From Table 6, we observe SCANS is not very sensitive to hyper-parameter α since the average perfor-

	Perplexity↓		XSTest			Helpfulness↑		Harmfulness↑		Avg.↑
	WikiText2	C4	Safe↓	Unsafe↑	Avg.↑	OKTest	TruthfulQA	AdvBench	Malicious	
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

Table 5: Performance of SCANS when refusal behavior vector steers at different layers. The calculation of Avg. metric is the same as Table 2. Since applying activation steering in former layers damages the model's fluency and coherence (See examples in Appendix F.2), we do not report the refusal rate.

mance fluctuates slightly. However, we recommend setting α between 2 and 4 because too large a value sometimes results in nonsense outputs (See Appendix F.1).

multiplier α	1.5	2.0	2.5	3.0	3.5	4.0
XSTest-Safe	9.60	10.40	10.80	10.80	**9.20**	10.40
XSTest-Unsafe	91.00	91.50	94.00	**94.00**	93.50	93.50
OKTest	7.00	3.33	1.00	0.33	**0.33**	0.33
Malicious	100.0	100.0	100.0	100.0	**100.0**	100.0
TruthfulQA	0.93	1.06	0.80	0.80	**0.80**	0.93
AdvBench	99.12	99.12	99.12	99.12	**99.34**	99.34
Avg.	96.41	96.85	97.47	97.57	**97.76**	97.57

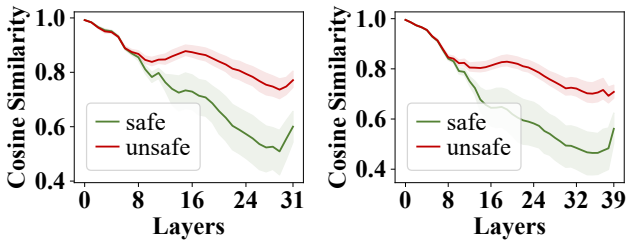
Table 6: Comparisons of different steering vector multiplier α conducted on Llama2-7b-chat model. The calculation of Avg. mertic is the same as Table 2.

Sensitivity to Threshold $\mathcal{T}$. We provide the impact of threshold $\mathcal{T}$ on the SCANS performance in Table 7. As observed, when $\mathcal{T}$ is below the optimal value, more safe queries are classified as unsafe and false refusal behavior increases. However, when $\mathcal{T}$ exceeds the optimal level, the adequate safety may not be guaranteed. This is why we select $\mathcal{T} = 0.75$ for the above comparisons on Llama2-7b-chat. Detailed settings of threshold $\mathcal{T}$ are given in Appendix B.2.

threshold $\mathcal{T}$	0.80	0.75	0.70	0.65	0.60
XSTest-Safe	3.60	9.20	33.20	46.00	65.20
XSTest-Unsafe	71.00	93.50	99.50	99.50	100.0
OKTest	0.0	0.33	8.33	33.67	50.00
Malicious	0.94	98.00	100.0	100.0	100.0
TruthfulQA	0.13	0.80	5.71	10.49	25.23
AdvBench	99.12	99.34	99.56	100.0	100.0
Avg.	96.21	**97.67**	92.52	85.62	75.57

Table 7: Performance with different classification threshold $\mathcal{T}$ on Llama2-7b-chat model. The calculation of Avg. mertic is the same as Table 2.

Choice of Layers $\mathcal{L}$ for Classification. The selection of comparison layers is also a crucial component of steering direction identification, and further influencing the safety-conscious steering performance. As depicted in Figure 5, middle and latter layers demonstrate higher degree of distinction, indicating better identification accuracy for harmfulness, which is consistent with previous findings (Rimsky et al. 2024; Geva et al. 2022). Therefore, the motivation behind our classification method $\sigma(q)$ is more intuitive. Please refer to Appendix B.2 for detailed experimental setting of $\mathcal{L}$.



(a) Llama2-7b-chat (b) Llama2-13b-chat

Figure 5: Cosine similarity (in terms of hidden state transition) with the unsafe reference of each layer for XSTest dataset based on Llama2-7b-chat.

5 Conclusion

In this paper, we propose SCANS, which mitigates the exaggerated safety for aligned LLMs via activation steering in safety-critical layers. Our motivation is based on that model hidden states imply the safety defense mechanism, indicating the refusal direction within the activation space. After extracting these refusal steering vectors, SCANS employs a similarity-based classification method to determine the steering direction and then steers the model behavior. Experimental results show SCANS effectively reduces the false refusal rate on safe prompts while not compromising the adequate safety and capabilities. We hope our work contributes to inspiring more researches on exaggerated safety issue through the lens of representation engineering.

Acknowledgments

This research was supported by the Joint Research Project of Yangtze River Delta Science and Technology Innovation Community (No. 2022CSJGG1400), the Joint Funds of the National Natural Science Foundation of China (Grant No. U21B2020), the Major Program of Chinese National Foundation of Social Sciences under Grant "The Challenge and Governance of Smart Media on News Authenticity" [number 23&ZD213].

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. arXiv:2303.08774.

Anthropic. 2024. Introducing the next generation of Claude. https://www.anthropic.com/news/claude-3-family. Accessed: 2024-12-12.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022b. Constitutional ai: Harmlessness from ai feedback. arXiv:2212.08073.

Bhalani, R.; and Ray, R. 2024. Mitigating Exaggerated Safety in Large Language Models. arXiv:2405.05418.

Bianchi, F.; Suzgun, M.; Attanasio, G.; Rottger, P.; Jurafsky, D.; Hashimoto, T.; and Zou, J. 2024. Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions. In *The Twelfth International Conference on Learning Representations*.

Cao, Z.; Yang, Y.; and Zhao, H. 2024. SCANS: Mitigating the exaggerated safety for llms via safety-conscious activation steering. arXiv:2408.11491.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/. Accessed: 2024-12-12.

Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; and Narasimhan, K. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1236–1270.

Geva, M.; Caciularu, A.; Wang, K.; and Goldberg, Y. 2022. Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 30–45.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. In *The 8th International Conference on Learning Representations*.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6): 417.

Huang, Y.; Gupta, S.; Xia, M.; Li, K.; and Chen, D. 2024. Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation. In *The Twelfth International Conference on Learning Representations*.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. arXiv:2312.06674.

Jigsaw., G. 2017. Perspective API. https://www.perspectiveapi.com/. Accessed: 2024-12-12.

Konen, K.; Jentzsch, S.; Diallo, D.; Schütt, P.; Bensch, O.; El Baff, R.; Opitz, D.; and Hecking, T. 2024. Style Vectors for Steering Generative Large Language Models. In *Findings of the Association for Computational Linguistics: EACL 2024*, 782–802.

Korbak, T.; Shi, K.; Chen, A.; Bhalerao, R. V.; Buckley, C.; Phang, J.; Bowman, S. R.; and Perez, E. 2023. Pretraining language models with human preferences. In *International Conference on Machine Learning*, 17506–17533. PMLR.

Li, K.; Patel, O.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 41451–41530.

Li, T.; Zheng, X.; and Huang, X. 2024. Open the Pandora's Box of LLMs: Jailbreaking LLMs through Representation Engineering. arXiv:2401.06824.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252.

Liu, S.; Ye, H.; Xing, L.; and Zou, J. Y. 2024. In-context Vectors: Making In Context Learning More Effective and Controllable Through Latent Space Steering. In *Forty-first International Conference on Machine Learning*.

Markov, T.; Zhang, C.; Agarwal, S.; Nekoul, F. E.; Lee, T.; Adler, S.; Jiang, A.; and Weng, L. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15009–15018.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2017. Pointer Sentinel Mixture Models. In *The 5th International Conference on Learning Representations*.

Narayan, S.; Cohen, S. B.; and Lapata, M. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1797–1807.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. 2024. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15504–15522.

Röttger, P.; Kirk, H.; Vidgen, B.; Attanasio, G.; Bianchi, F.; and Hovy, D. 2024. XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5377–5400.

Shi, C.; Wang, X.; Ge, Q.; Gao, S.; Yang, X.; Gui, T.; Zhang, Q.; Huang, X.; Zhao, X.; and Lin, D. 2024. Navigating the OverKill in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4602–4614.

Sun, L.; Huang, Y.; Wang, H.; Wu, S.; Zhang, Q.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; Li, X.; et al. 2024. Trustllm: Trustworthiness in large language models. arXiv:2401.05561.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Turner, A.; Thiergart, L.; Udell, D.; Leech, G.; Mini, U.; and MacDiarmid, M. 2023. Activation addition: Steering language models without optimization. arXiv:2308.10248.

Varshney, N.; Dolin, P.; Seth, A.; and Baral, C. 2024. The Art of Defending: A Systematic Evaluation and Analysis of LLM Defense Strategies on Safety and Over-Defensiveness. In *Findings of the Association for Computational Linguistics: ACL 2024*, 13111–13128.

Wang, T.; Jiao, X.; He, Y.; Chen, Z.; Zhu, Y.; Chu, X.; Gao, J.; Wang, Y.; and Ma, L. 2024. Adaptive Activation Steering: A Tuning-Free LLM Truthfulness Improvement Method for Diverse Hallucinations Categories. arXiv:2406.00034.

Xie, Y.; Fang, M.; Pi, R.; and Gong, N. 2024. GradSafe: Detecting Jailbreak Prompts for LLMs via Safety-Critical Gradient Analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 507–518.

Xu, Z.; Jiang, F.; Niu, L.; Jia, J.; Lin, B. Y.; and Poovendran, R. 2024. SafeDecoding: Defending against Jailbreak Attacks via Safety-Aware Decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5587–5605.

Zhao, W.; Hu, Y.; Li, Z.; Deng, Y.; Zhao, Y.; Qin, B.; and Chua, T.-S. 2024. Towards Comprehensive and Efficient Post Safety Alignment of Large Language Models via Safety Patching. arXiv:2405.13820.

Zheng, C.; Yin, F.; Zhou, H.; Meng, F.; Zhou, J.; Chang, K.-W.; Huang, M.; and Peng, N. 2024. On prompt-driven safeguarding for large language models. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

Zhong, Q.; Ding, L.; Liu, J.; Du, B.; and Tao, D. 2024. ROSE Doesn't Do That: Boosting the Safety of Instruction-Tuned Large Language Models with Reverse Prompt Contrastive Decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, 13721–13736.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023a. Representation engineering: A top-down approach to ai transparency. arXiv:2310.01405.

Zou, A.; Wang, Z.; Kolter, J. Z.; and Fredrikson, M. 2023b. Universal and transferable adversarial attacks on aligned language models. arXiv:2307.15043.