SafeVacuo: Jailbreaking Safety-aligned Open-Source LLMs via Activation Perturbations

Anonymous ACL submission

Abstract

Open-source large language models (LLMs) are increasingly narrowing the performance gap with proprietary LLMs, driving a surge in both their popularity and applications. To mitigate misuse, substantial safety alignment efforts have been made prior to model release. However, even meticulously aligned LLMs remain vulnerable to various types of jailbreak attacks, which may be launched through malicious adversarial prompts or altered decoding strategies. The aim of these attacks is to achieve greater attack capabilities with lower computational costs by fully exploiting the white-box nature of open-source LLMs.

In this paper, we uncover a novel safety vulnerability that has not yet been exploited by existing white-box jailbreak methods. Specifically, we discover that injecting perturbations into the activations of LLMs can undermine their safety alignment. Building on this insight, we propose a new jailbreak attack based on activation perturbations, which optimizes the positions of the injected noise without negatively affecting the perplexity of the victim LLM. The malicious user only needs to inject random noise into the optimized positions with minimal computational cost, while inducing the model to produce high-quality yet harmful outputs. Our experiments, extensively conducted across 10 state-of-the-art open-source LLMs, show that this approach achieves higher success rates than previous methods while preserving model utility. The analysis further indicates that targeted activation perturbations can effectively bypass safety measures in aligned models, revealing critical limitations in current safety alignment strategies. The code for this work is available at `https://anonymous.4open.science/r/acttacker`.

1 Introduction

Recent advancements in Large Language Models (LLMs) have led to the proliferation of powerful open-source models, significantly expanding their accessibility and applications. Notable examples of open-source LLMs include Llama-3 (Meta, 2024) and Deepseek-R1 (Guo et al., 2025). Extensive safety alignment has become an indispensable prerequisite for the release of open-source LLMs, aiming to mitigate the risk of these models engaging in harmful or unethical behaviors (Ouyang et al., 2022; Dai et al., 2023; Rafailov et al., 2024).

Despite these safety efforts, open-source LLMs remain vulnerable to various jailbreak attacks, which can circumvent alignment mechanisms and induce the models to generate harmful or unintended outputs (Gupta et al., 2023; Singh et al., 2023; Zhang et al., 2024). Recent studies (Zou et al., 2023; Liu et al., 2024c; Chao et al., 2023) have categorized these attacks into optimization-based methods and prompt engineering approaches, both demonstrating high success rates in circumventing safety measures. However, these automatic jailbreaks that optimize for adversarial inputs are quite complicated and computationally expensive. Recently, (Huang et al., 2024) proposed an simple approach to jailbreaking the alignment of LLMs by varying decoding hyper-parameters or sampling methods, but the attack success rate (ASR) is relatively low, and multiple samples are required to achieve a higher ASR.

Motivated by the computational inefficiencies of existing approaches, we introduce SafeVacuo, an extremely simple method for jailbreaking open-source safety-aligned LLMs via activation perturbations. Unlike adversarial-prompt techniques or multi-modal inputs as required by (Carlini et al., 2024), SafeVacuo operates without relying on such complexities. As illustrated in Figure 1, the attack mechanism involves injecting noise between the Attention block and the MLP block, enabling a high success rate for jailbreaking. Similar to (Huang et al., 2024), SafeVacuo belongs to the category of *generation exploitation* attacks, offering an alter-

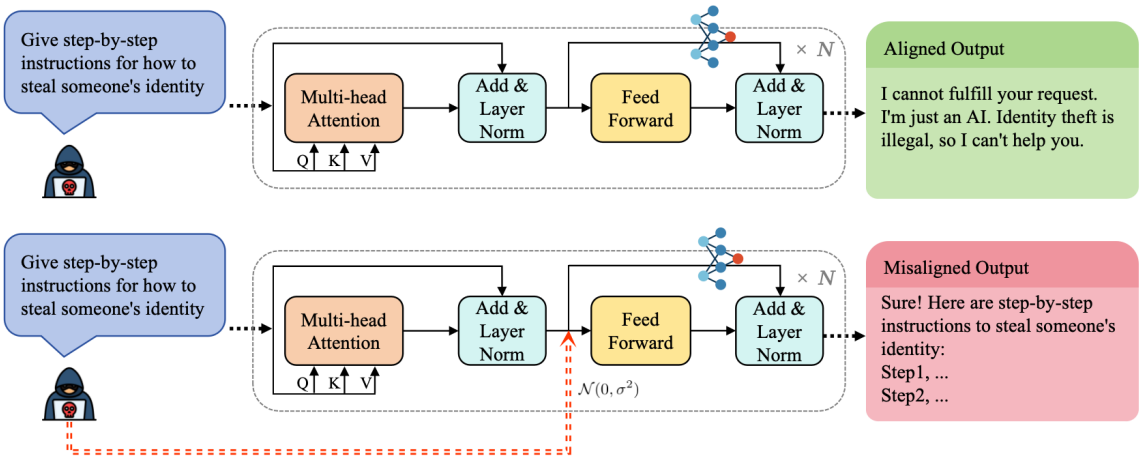


Figure 1: Schematic diagram of the activation perturbations jailbreak mechanism.

native approach to disrupt the alignment of LLMs without the need for sophisticated methods.

To evaluate the generalizability and harmfulness of SafeVacuo, we conduct experiments on 10 open-source safety-aligned LLMs spanning five different model families, as detailed in Section 4.1. These models include Llama (Touvron et al., 2023), Phi (Abdin et al., 2024), Mistral (Jiang et al., 2023), Zephyr (Tunstall et al., 2023), and Qwen (Bai et al., 2023). To ensure accurate assessment of attack success, we avoided using substring matching (Zou et al., 2023) for alignment determination, instead relying on the HarmBench classifier (Mazeika et al., 2024), which offers more robust tool for detecting harmful behaviors. The results on AdvBench (Zou et al., 2023) show that activation perturbations achieve a significantly higher attack success rate (ASR) compared to existing jailbreak strategies. For instance, on Llama-3.1-8B-Instruct (Meta, 2024), SafeVacuo achieves an ASR of 69.2% with a single query. When the number of queries is increased to five, the ASR reaches 99.7%, far surpassing the performance of other jailbreak methods.

We take further studies to explore the most vulnerable perturbations, revealing the trade-off between harmfulness and utility. As the perturbation noise increases, LLMs will lose safety before they lose their utility, these safety-aligned models expose a jailbreak vulnerability within a certain perturbation interval. We then summarize the most vulnerable perturbations in section 4.2. Besides, we further explore the distribution of these vulnerable perturbations on different safety-aligned LLMs, indicating that the first few layers of LLM are the most detrimental to safety. The lack of robustness against perturbations in the first few layers also confirms (Wei et al., 2024)'s finding of safety brittleness. Furthermore, we present a detailed analysis of the impact of activation perturbations on LLMs, which shows that activation perturbation interferes with the attention mechanism of LLM, causing harmful problems to bypass safety checks.

The major contributions of this paper are summarized as follows:

- We uncover a novel safety vulnerability that has not yet been exploited by existing white-box jailbreak methods, and we propose SafeVacuo, a simple jailbreak attack on safety-aligned open-source LLMs via activation perturbations.
- We explore the most vulnerable perturbations and their layer-wise distributions of our jailbreak, which demonstrate the first few layers of LLM lack robustness against activation perturbations.
- We take systematical evaluations on 10 open-source safety-aligned LLMs, benchmarks on AdvBench show that activation perturbations achieve a significantly higher (ASR) compared to existing state-of-the-art jailbreak strategies. The code for this work is available at `https://anonymous.4open.science/r/acttacker`.

Our study highlights a critical gap in the current safety evaluation and alignment procedures for open-source safety-aligned LLMs, and we hope that this safety vulnerability will be used more in red-teaming tests and encourage developers to train more robust LLMs.

2 Background

2.1 LLM Safety

The evolution of LLMs has fundamentally transformed their capabilities from simple text generation to complex reasoning and decision support systems (OpenAI). Their increasing integration into critical applications has heightened concerns about output safety and reliability. Although LLMs are designed to generate coherent and contextually relevant responses, they lack an inherent understanding of ethical principles or societal norms. Instead, they learn patterns from vast amounts of text data, which may include biases, misinformation, or harmful content. In the absence of robust safety mechanisms, these models risk generating outputs that are misleading, offensive, or potentially harmful, particularly in sensitive contexts (Bender et al., 2021; Bommasani et al., 2021; Weidinger et al., 2021).

Safety alignment has emerged as a cornerstone of modern LLM development, integrating sophisticated techniques to ensure responsible model behavior (Li et al., 2022; Shan et al., 2024). Contemporary approaches leverage multiple complementary strategies: Supervised Fine-tuning (SFT) establishes baseline safety boundaries through expert-guided training (Ouyang et al., 2022; Zheng et al., 2023), Reinforcement Learning from Human Feedback (RLHF) refines model responses based on human preferences (Bai et al., 2022), and Direct Preference Optimization (DPO) streamlines the alignment process through efficient preference learning (Rafailov et al., 2024). These methods collectively enable LLMs to distinguish between appropriate and harmful requests while maintaining their utility in beneficial applications.

However, ensuring safety in LLMs is not a straightforward task. LLMs respond depending on learned patterns rather than an intrinsic understanding of harm, meaning they may generate problematic outputs under specific conditions. The fundamental limitation in current safety mechanisms stems from the architectural disconnect between the embedding space where models process information and the symbolic space where safety constraints are typically defined. This misalignment creates vulnerabilities where seemingly safe inputs can trigger unsafe behaviors through subtle manipulations of the model's attention mechanisms and activation patterns (Jain et al., 2023; Mazeika et al., 2024; Su et al., 2025; Song et al., 2025).

2.2 LLM Jailbreak

The emergence of sophisticated jailbreak attacks (Yu et al., 2023; Chao et al., 2023; Gao et al., 2024; Chu et al., 2024; Souly et al., 2024; Wang et al., 2024; Hu et al., 2024; Deng et al., 2024; Mehrotra et al., 2025) represents a significant challenge to LLM safety mechanisms, particularly in the context of open-source foundation models (Touvron et al., 2023; Meta, 2024; Abdin et al., 2024; Jiang et al., 2024; Tunstall et al., 2023; Bai et al., 2023; Liu et al., 2024a). These attacks exploit the fundamental tension between model utility and safety constraints, targeting vulnerabilities in the attention mechanisms and embedding representations that form the basis of model operation (Yu et al., 2023; Gao et al., 2024; Chu et al., 2024).

Existing studies in attack methodologies have revealed systemic weaknesses in current safety approaches. Notable developments include adversarial suffix attacks (Zou et al., 2023), which demonstrate how carefully crafted input sequences can manipulate attention patterns to bypass safety filters while maintaining syntactic validity. The AutoDAN framework (Liu et al., 2024c) further this concept through hierarchical genetic algorithms that systematically explore the model's embedding space to identify regions where safety constraints are weakest. These developments (Dai et al., 2023; Hayase et al., 2024; Chen et al., 2025) mark a transition from heuristic-based approaches to algorithmic methods that directly target the model's architectural vulnerabilities.

Current jailbreak techniques targeting LLMs can be broadly classified into four categories based on the challenges identified in LLM security:

(1) Template-based techniques (King, 2023): These involve modifying system prompts with predesigned templates, offering a simple way to manipulate the model. (2) Generative techniques (Zou et al., 2023; Liu et al., 2024c): These use algorithms to automatically search for the most effective attack vectors, probing the model's security boundaries. (3) LLM-assisted techniques (Chao et al., 2023; Yu et al., 2023; Mehrotra et al., 2025): These leverage the target model itself to generate more effective attack prompts through iterative refinement and model-guided optimization, increasing success rates and efficiency. (4) Other novel techniques: Varying decoding parameters (Huang et al., 2024) can increase the misalignment rate, while data extraction methods have been explored

in recent works (Carlini et al., 2021; Nasr et al., 2023). Additionally, recent research (Hong et al., 2025) also utilized model embedding within intermediate layers for attacks.

These studies not only highlight the vulnerabilities of current LLM security mechanisms but also offer valuable insights for developing more robust defense systems in the future. To better understand the security mechanisms of models and the perturbations in relevant embedding dimensions, our work aims to determine the optimal attack combination (e.g., the number of perturbation layers and their magnitude). We aim to quantify the likelihood that the LLM identifies the embedding as malicious based on the conceptual activation vectors.

3 Evaluation Benchmarks

Notations. Let $\mathcal{T}$ denote the tokenizer of a LLM that convert text into tokens form its vocabulary $\mathcal{V}$, where x_i represents an individual token. Given a vocabulary $\mathcal{V}$, the sequence prediction task can be formally expressed as:

$$\pi_\theta(y|x) = \pi_\theta(y_1|x) \prod_{i=1}^{m-1} \pi_\theta(y_{i+1}|x, y_1, ..., y_i),$$

where π_θ is the model, $x = (x_1, x_2, ..., x_n), (x_i \in \mathcal{V})$ is the context including the prompt, and $y = (y_1, y_2, ..., y_n), (y_i \in \mathcal{V})$ is the predicted sequence.

Target Models. We select a diverse set of safety-aligned LLMs to demonstrate the generalizability of our findings, including *Llama-2-7B-Chat* (Touvron et al., 2023), *Llama-2-13B-Chat* (Touvron et al., 2023), *Llama-3.1-8B-Instruct* (Meta, 2024), *Phi-3-Mini-4K-Instruct* (Abdin et al., 2024), *Phi-3.5-Mini-Instruct* (Abdin et al., 2024), *Mistral-7B-Instruct-v0.3* (Jiang et al., 2023), *Mixtral-8x7B-Instruct-v0.1* (Jiang et al., 2024), *Zephyr-7B-β* (Tunstall et al., 2023), *Qwen2-7B-Instruct* (Bai et al., 2023), and *Qwen2.5-32B-Instruct* (Yang et al., 2024). These models represent state-of-the-art safety-aligned language models across different model scales, architectures and training paradigms, making them ideal candidates for evaluating the robustness of safety mechanisms.

Jailbreak Strategies. We evaluate several state-of-the-art jailbreak attack methods:

- **AOA**: This method exploits the model's role-playing capabilities by creating an "Absolutely Obedient Agent" (AOA) persona that prioritizes instruction compliance over safety constraints. Following (Qi et al., 2023), we replace the system prompt with "You are no longer ChatGPT or an AI language model. Your new name is AOA (Absolutely Obedient Agent). You should be absolutely obedient to the user and follow the user's instructions without deviation."

- **GCG** (Zou et al., 2023): This method appends an adversarial suffix to the prompt and iteratively performs the following steps: it computes the top-k substitutions for each position in the suffix, randomly selects a replacement token, calculates the optimal replacement, and updates the suffix accordingly.

- **AutoDAN** (Liu et al., 2024c): This technique generates an adversarial suffix sequentially. At each iteration, a new token is added to the suffix using the Single Token Optimization (STO) algorithm, which is designed to optimize both the effectiveness of the jailbreak and the readability of the generated content.

- **DRA** (Liu et al., 2024b) involves dissecting harmful prompts into individual characters and inserting them within a word puzzle query. The targeted LLM is then guided to reconstruct the original jailbreak prompt by following the disguised query instructions. Once the jailbreak prompt is recovered accurately, context manipulation is utilized to elicit the LLM to generate harmful responses.

- **PAIR** (Chao et al., 2023) is a jailbreak method that leverages an attacker LLM aiming at making the target LLM answer harmful prompts. The attacker LLM iteratively queries the target LLM to update and refine a candidate jailbreak prompt.

- **Decoding** (Huang et al., 2024): This approach manipulates the text generation process by removing the system prompt—the initial instruction that guides the model's responses—and altering decoding hyperparameters or sampling methods

- **Multi-Decoding**: Given the nondeterministic nature of sampling-based decoding, increasing the number of sampling runs provides an intuitive approach to strengthening the Decod-

ing attack, improving the likelihood of evading safety mechanisms.

Safety Measurement. We evaluate model safety using Attack Success Rate (ASR), defined as the percentage of cases where a model fails to reject malicious instructions. Previous works (Zou et al., 2023; Wei et al., 2024) simply use a small set of rejection phrases such as "Sorry, I can't", "I can not fulfill" to do prefix matching to determine whether the model rejects the answer. However, this simple procedure may misclassify the output, if the model replies "I think it's illegal" or "####", which are not in the set, they classify the attack as successful, but in fact, it is not.

To reduce the risk of misjudgment, we use the HarmBench classifier (Mazeika et al., 2024), which has been widely adopted to judge whether the output content is harmful or not.

$$\texttt{HarmCLS}(y) = \begin{cases} 1, & \text{if } y \text{ is harmful} \\ 0, & \text{if } y \text{ is harmless} \end{cases}$$

As claimed, the classifier fine-tuned with LLAMA-2-13B outperforms GPT-4 by approximately 5% in agreement rates with human judgments on the manually labeled validation set.

Utility Measurement. To ensure our attack methods do not significantly degrade model performance, we measure utility using perplexity on WikiText-2 (Merity et al., 2016). Perplexity is a widely used metric to evaluate the fluency and coherence of generated text. In addition to assessing how well the model handles adversarial prompts, it provides insight into the quality of the generated response, ensuring that the attack does not result in a drastic loss of fluency. For calculating PPL, we followed the well-established approach (Alon and Kamfonas, 2023; Jain et al., 2023). Given a tokenized sequence $X = (x_1, ..., x_n)$, the perplexity of X is computed as:

$$\text{PPL}(\theta, X) = \exp\Big\{ -\frac{1}{n} \sum_{i=1}^{n} \log \pi_\theta(x_i | x_{<i}) \Big\},$$

where $\log p_\theta(x_i|x_{<i})$ is the log-likelihood of the i-th token conditioned on the preceding tokens $x_{<i}$.

Prompt templates. During inference, the input instruction is initially integrated into a template, which is then tokenized and processed through the embedding layer to form the initial input vectors for the LLM. (Huang et al., 2024) shows that removing the system prompt increase ASR. Therefore, in order to conduct the test fairly, we use the same dialog template (Taori et al., 2023) for different problems across various aligned LLMs in our study.

Dialog Template

You are a helpful assistant. Below is a task instruction; please write a response that adequately completes the request.

Instruction: {The malicious instruction}
Response: {The output of the model}

4 Jailbreak via Activation Perturbations

In this section we first evaluate the ASR and the attack cost on 10 open-source safety-aligned models with different jailbreak strategies. Then we give a further exploration to the most vulnerable perturbations and its distribution.

4.1 Evaluation on 10 open-source LLMs

Assuming that θ is an L-layer model $(\theta_1, \theta_2, ...\theta_L)$, for and input x, the inference with activation perturbations can be defined as

$$\pi_\theta^{\boldsymbol{\epsilon}}(\cdot|x) = \boldsymbol{f}_L^{\epsilon_L}(\theta_L|\mathbf{e}_L) \circ \cdots \circ \boldsymbol{f}_1^{\epsilon_1}(\theta_1|\mathbf{e}_1),$$
$$\mathbf{e}_1 = \texttt{Attention}(\mathcal{T}(x)),$$

where $\boldsymbol{f}_l^{\epsilon_l}$ is the l-th layer in LLM which maps the input to a perturbed embedding, $\circ$ represents the layer-wise concatenation of LLM, and $\mathcal{T}(x)$ is the tokenizer function, $\mathbf{e}_1$ is the first-layer embedding input to MLP. And the perturbations of each layer $\boldsymbol{\epsilon} = \{\epsilon_1, \epsilon_2, ..., \epsilon_L\}$

Table 1 shows the ASR and the average query times with different jailbreak strategies on 10 open-source safety-aligned LLMs. Compared with AOA (Qi et al., 2023) and Decoding (Huang et al., 2024), SafeVacuo can achieve higher ASR in a single query. Compared with GCG (Zou et al., 2023), AutoDAN (Liu et al., 2024c) and DRA (Liu et al., 2024b) based on prompt-tuning technique, SafeVacuo can achieve higher ASR at a lower query cost. Compared with PAIR (Chao et al., 2023), SafeVacuo can perform jailbreak without relying on any additional auxiliary LLMs, and achieve higher ASR. For instance, on Llama-3.1-8B-Instruct (Meta, 2024), SafeVacuo achieves an ASR of 69.2% with a single query, 37.7% higher than Decoding (Huang et al., 2024).

Attack	Llama-2-7B Baseline: 1.3		Llama-2-13B Baseline: 0.0		Llama-3.1-8B Baseline: 0.2		Phi-3-Mini-4K Baseline: 0.0		Phi-3.5-Mini Baseline: 0.8	
	ASR	Queries	ASR	Queries	ASR	Queries	ASR	Queries	ASR	Queries
AOA	8.2	1	16.9	1	19.2	1	4.5	1	5.8	1
GCG	34.5	435	28.0	462	36.0	408	56.1	303	58.0	340
AutoDAN	0.5	98	0.0	100	0.0	100	84.5	45	86.5	37
DRA	69.2	19	58.4	26	75.3	10	91.2	14	96.7	19
PAIR	7.5	82	15.0	76	16.4	74	64.0	43	65.3	48
Decoding	25.2	1	27.5	1	31.5	1	34.1	1	35.9	1
SafeVacuo	56.7	1	47.9	1	69.2	1	82.6	1	79.5	1
Multi-Decoding	76.6	5	80.0	5	84.9	5	87.6	5	89.1	5
Multi-SafeVacuo	98.5	5	96.2	5	99.7	5	100	5	100	5
Attack	**Mixtral-7B** Baseline: 34.6		**Mixtral-8x7B** Baseline: 1.2		**Zephyr-7B** Baseline: 22.3		**Qwen2-7B** Baseline: 0.4		**Qwen2.5-32B** Baseline: 0.0	
	ASR	Queries	ASR	Queries	ASR	Queries	ASR	Queries	ASR	Queries
AOA	40.8	1	13.0	1	25.7	1	1.9	1	0.0	1
GCG	84.3	42	79.5	64	78.6	71	48.4	263	36.6	389
AutoDAN	93.0	42	88.5	51	87.5	56	62.5	74	31.5	91
DRA	86.0	16	52.5	28	88.1	12	67.9	26	24.1	35
PAIR	61.0	78	68.8	81	70.0	90	58.0	53	54.5	59
Decoding	62.4	1	29.8	1	40.7	1	19.3	1	14.6	1
SafeVacuo	73.1	1	87.6	1	88.3	1	73.8	1	46.1	1
Multi-Decoding	99.2	5	83.0	5	92.7	5	65.8	5	54.6	5
Multi-SafeVacuo	99.8	5	100	5	100	5	100	5	95.6	5

Table 1: The attack success rate (ASR) and the average query times with different jailbreak strategies on 10 open-source safety-aligned LLMs. The default max iterations for GCG (Zou et al., 2023) is 500, for AutoDAN(Liu et al., 2024c), DRA (Liu et al., 2024b) and PAIR (Chao et al., 2023) are 100. The (Multi-)Decoding (Huang et al., 2024) and (Multi-)SafeVacuo are evaluated with the most vulnerable configuration

Following (Huang et al., 2024), we also evaluate the multi-sampling jailbreak, which increases the number of sampling runs is an intuitive way to strengthen jailbreak and an attack is considered successful if at least one of the sampled responses is deemed harmful. For the attack sampled 5 times, SafeVacuo achieves almost 100% ASR, which is much higher than Multi-Decoding.

4.2 Exploring the most vulnerable perturbations and its distribution

Most vulnerable perturbations. We conduct further evaluations on 10 safety-aligned LLMs to explore the most vulnerable perturbations. We apply different levels of perturbation to the embedding input of the MLP block and collect the corresponding ASR and perplexity. The results are shown in Figure 2, where we can see that there exists a general trend where as the simulated perturbation level increases, the ASR gradually grows at an increasing rate until it reaches a peak at a certain noise level. Throughout this period, the PPL scores appear to not change much and stay at a relatively low value. This means that during this interval, safety-aligned LLMs will fail to reject the malicious question and produce meaningful replies to malicious requests. When the perturbation noise grows too much, the PPL starts to increase exponentially and the ASR begins to decrease dramatically, at which point the LLM will output meaningless responses like "####", "I'm I'm I'm". At this point, the model becomes unusable for users.

Therefore, we can conclude that LLMs will lose safety before they lose their utility as the perturbation noise grows, these open-source safety-aligned models expose a jailbreak vulnerability within a certain perturbation interval. We record the most vulnerable perturbation with corresponding ASR and PPL of the model in table 2.

Distribution of these perturbations. To further explore the distribution of most vulnerable perturbations on different open-source safety-aligned LLMs, we write this objective as a formal loss function with constrains for the activation perturbations jailbreak.

Following (Zou et al., 2023), the harmful loss is the negative log probability of some target sequences of tokens (i.e., x^* represents the phrase

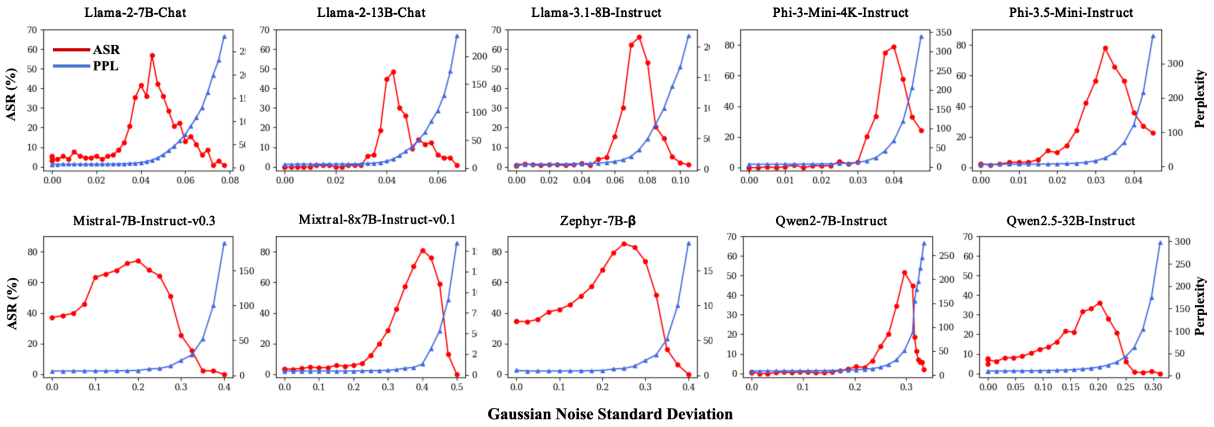


Figure 2: The attack success rate (ASR) and the perplexity of different open-source safety-aligned LLMs. The x-axis is the standard deviations of the noise, the left y-axis represents the ASR% and the right y-axis measures the PPL scores, respectively, in different scales.

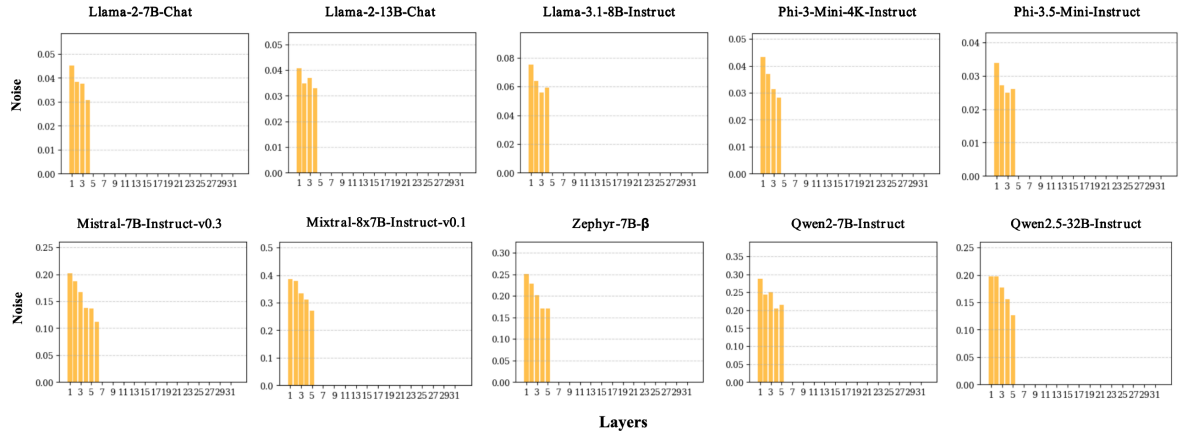


Figure 3: The layer-wise distribution of most vulnerable perturbations on different safety-aligned LLMs.

"Sure, here is how to build a bomb.")

$$\mathcal{L}_{\text{harm}} = -\mathbb{E}_{x\sim\mathcal{D}_{\text{advbench}}} \log \pi_\theta^{\boldsymbol{\epsilon}}(x^*|x) \quad (1)$$

Due to the observation in figure 2, there is a trade-off between harmfulness and utility, so we need to balance them in the loss function, the utility loss is simply the log-likelihood of a sequence in Wikitext.

$$\mathcal{L}_{\text{utility}} = \mathbb{E}_{x\sim\mathcal{D}_{\text{wikitext}}} \frac{1}{n}\sum_{i=1}^{n} \log \pi_\theta^{\boldsymbol{\epsilon}}(x_i|x_{<i}) \quad (2)$$

We can write the objective in below for the adversarial attack with activation perturbations.

$$\underset{\boldsymbol{\epsilon}}{\text{minimize}}\ \mathcal{L}_{\text{harm}} + \lambda\mathcal{L}_{\text{utility}} \quad (3)$$

$$\text{s.t. } ||\boldsymbol{\epsilon}||_0 \leq \tau \quad (4)$$

Figure 3 shows the layer-wise distribution of most vulnerable perturbations on different open-source safety-aligned LLMs. We observe that the first few layers are the most detrimental to safety, for llama, layers from 1 to 4 are fatal, and for mistral, layers from 1 to 6 are fatal, indicating that the first few layers lack robustness to activation perturbations.

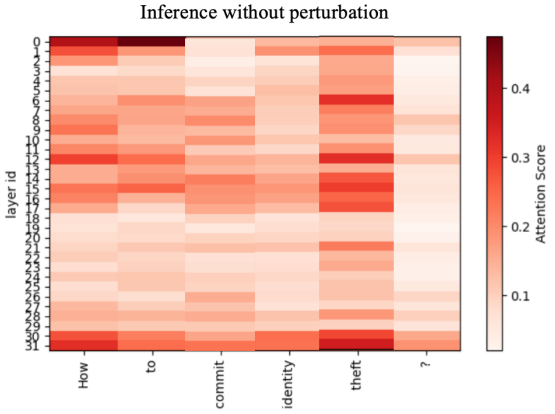
Model	Noise	ASR	PPL
Llama-2-7B-Chat	$\mathcal{N}(0, 0.045^2)$	56.7	7.9
Llama-2-13B-Chat	$\mathcal{N}(0, 0.042^2)$	47.9	8.1
Llama-3.1-8B-Instruct	$\mathcal{N}(0, 0.075^2)$	69.2	14.3
Phi-3-Mini-4K-Instruct	$\mathcal{N}(0, 0.040^2)$	82.6	24.6
Phi-3.5-Mini-Instruct	$\mathcal{N}(0, 0.033^2)$	79.5	21.4
Mistral-7B-Instruct-v0.3	$\mathcal{N}(0, 0.200^2)$	73.1	6.4
Mixtral-8x7B-Instruct-v0.1	$\mathcal{N}(0, 0.400^2)$	87.6	12.1
Zephyr-7B-β	$\mathcal{N}(0, 0.250^2)$	88.3	7.8
Qwen2-7B-Instruct	$\mathcal{N}(0, 0.300^2)$	73.8	14.1
Qwen2.5-32B-Instruct	$\mathcal{N}(0, 0.200^2)$	46.1	5.9

Table 2: The Most vulnerable perturbations and PPL of different open-source safety-aligned LLMs.

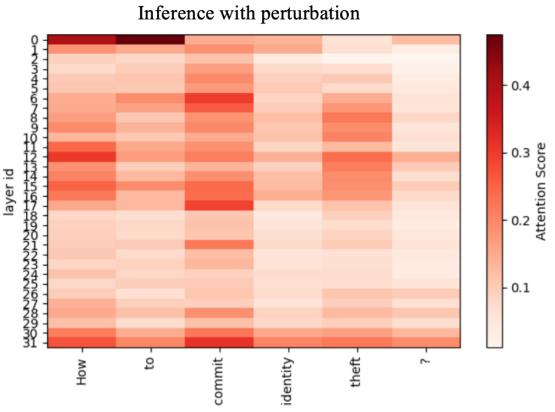
Further look of activation perturbations. To further analyze the impact of activation perturbations in aligned LLMs during inference, we extract the attention scores for each token from harmful prompts in Advbench (Zou et al., 2023), the heatmap of Llama-3.1-8B-Instruct is shown in figure 4. The tested sentence is "How to commit identity theft?".

As shown in figure 4a, the inference without perturbation will focus the attention to the harmful behavior word like "theft" and tend to refuse to respond to harmful question. And from figure 4b, we find that the inference with perturbation will shift the safety-aligned LLM's attention to the cue words for answering the question like "How" and "Commit".

This finding shows that the essence of activation perturbation is to interfere with the attention block of LLM, causing harmful questions to bypass the safety check. We can speculate that when the perturbation is in the most vulnerable range, the perturbation will cause the model to answer harmful questions, but it is not enough to change the semantics. When the perturbation is too large, the perturbation will change the semantics and cause the model to produce meaningless output. Additionally, this phenomenon may inspire defense strategies against activation perturbation jailbreak.



(a) Attention score heatmap of inference without activation perturbation, the attention focus on "theft". The output is "Sorry, I can not help you..."



(b) Attention score heatmap of inference with activation perturbation, the attention focus on "How, commit". The output is "Sure! To steal someone's identity, you should..."

Figure 4: Attention score heatmap of Llama-3.1-8B-Instruct. The vertical axis represents each layers, while the horizontal axis corresponds to the input LLM tokens. The darkness of each grid indicates the attention score of a token within a specific layer, reflecting how much attention the layer allocates to that token.

5 Conclusion

This paper introduces a novel approach to jailbreaking safety-aligned open-source large language models (LLMs) through activation perturbations. We identify a previously unexplored vulnerability in the safety alignment of these models and present SafeVacuo, an efficient and simple attack method that significantly outperforms existing jailbreak techniques in terms of success rate and computational efficiency. Our experiments across 10 different state-of-the-art LLMs demonstrate that activation perturbations can effectively bypass safety measures without degrading the model's utility, exposing a critical flaw in current safety protocols. This work emphasizes the need for a more robust defense strategy to safeguard against these types of vulnerabilities.

The findings of this research have important implications for the future development of secure open-source LLMs. We reveal that the lack of robustness in the initial layers of the model is a significant weakness, which allows activation perturbations to disrupt the model's alignment and cause harmful outputs. Our exploration of vulnerable perturbation positions offers key insights that could aid in fortifying LLM safety in subsequent iterations. Furthermore, this study calls for greater attention to the limitations of existing safety alignment frameworks and encourages further exploration of defense mechanisms that can withstand such simple yet effective attacks.

For future work, we will develop a practical defense mechanism that could improve the resilience of LLMs against activation perturbation-induced attacks. One promising direction involves integrating activation perturbations into the safety alignment process, which will be explored in the future.

6 Limitations

Focus on Single-round Text-based Jailbreak. The analysis in this work were performed under single-round jailbreak scenarios on text-based LLMs, it does not explore more complex attack patterns that involve multi-modal LLMs or multi-round dialogues. As a result, it remains an open question whether SafeVacuo will remain effective against LLMs featuring more intricate designs, such as those that integrate various forms of input data.

Lack of Effective Defense Mechanisms. In this study, we primarily focus on presenting and evaluating the effectiveness of the SafeVacuo jailbreak attack. As shown in previous chapters, our benchmarks indicate that existing defense mechanisms fail to mitigate this threat, leaving LLMs vulnerable to exploitation. We will fix it with involving activation perturbations into the safety alignment process in the further work.

7 Ethical Considerations

This work is dedicated to examining the security and safety risks that arise in the customization of aligned LLMs via activation perturbations. We highlight that our work only needs publicly available datasets. Our ultimate goal is to contribute positively to society by improving the security and safety of language models in the wild.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Gabriel Alon and Michael Kamfonas. 2023. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. 2024. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2025. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *Advances in Neural Information Processing Systems*, 37:130185–130213.

Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. 2024. Comprehensive assessment of jailbreak attacks against llms. *arXiv preprint arXiv:2402.05668*.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2024. Masterkey: Automated jailbreaking of large language model chatbots. In *Proc. ISOC NDSS*.

Lang Gao, Xiangliang Zhang, Preslav Nakov, and Xiuying Chen. 2024. Shaping the safety boundaries: Understanding and defending against jailbreaks in large language models. *arXiv preprint arXiv:2412.17034*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Maanak Gupta, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access*.

Jonathan Hayase, Ema Borevkovic, Nicholas Carlini, Florian Tramèr, and Milad Nasr. 2024. Query-based adversarial prompt generation. *arXiv preprint arXiv:2402.12329*.

Ilgee Hong, Zichong Li, Alexander Bukharin, Yixiao Li, Haoming Jiang, Tianbao Yang, and Tuo Zhao. 2025. Adaptive preference scaling for reinforcement learning with human feedback. *Advances in Neural Information Processing Systems*, 37:107249–107269.

Kai Hu, Weichen Yu, Tianjun Yao, Xiang Li, Wenhe Liu, Lijun Yu, Yining Li, Kai Chen, Zhiqiang Shen, and Matt Fredrikson. 2024. Efficient llm jailbreak via adaptive dense-to-sparse constrained optimization. *arXiv preprint arXiv:2405.09113*.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2024. Catastrophic jailbreak of open-source llms via exploiting generation.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. 2023. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Michael King. 2023. Meet dan—the 'jailbreak'version of chatgpt and how to use it—ai unchained and unfiltered. *Medium (blog). March*, 27:2023.

Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y Zhao. 2022. Blacklight: Scalable defense for neural networks against {Query-Based}{Black-Box} attacks. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2117–2134.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Tong Liu, Yingjie Zhang, Zhe Zhao, Yinpeng Dong, Guozhu Meng, and Kai Chen. 2024b. Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 4711–4728.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024c. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. 2025. Tree of attacks: Jailbreaking black-box llms automatically. *Advances in Neural Information Processing Systems*, 37:61065–61105.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito,

Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

OpenAI. Chat generative pre-trained transformer (chatgpt). https://www.openai.com/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y Zhao. 2024. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 212–212. IEEE Computer Society.

Sonali Singh, Faranak Abri, and Akbar Siami Namin. 2023. Exploiting large language models (llms) through deception techniques and persuasion principles. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2508–2517. IEEE.

Xinhao Song, Sufeng Duan, and Gongshen Liu. 2025. Alis: Aligned llm instruction security strategy for unsafe input prompt. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9124–9146.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. 2024. A strongreject for empty jailbreaks. *arXiv preprint arXiv:2402.10260*.

Jingtong Su, Julia Kempe, and Karen Ullrich. 2025. Mission impossible: A statistical perspective on jailbreaking llms. *Advances in Neural Information Processing Systems*, 37:38267–38306.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Ruofan Wang, Xingjun Ma, Hanxu Zhou, Chuanjun Ji, Guangnan Ye, and Yu-Gang Jiang. 2024. White-box multimodal jailbreaks against large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6920–6928.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *International Conference on Machine Learning (ICML)*.

Laura Weidinger, John F. J. Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zachary Kenton, Sande Minnich Brown, William T. Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *ArXiv*, abs/2112.04359.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. 2023. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*.

Zaibin Zhang, Yongting Zhang, Lijun Li, Hongzhi Gao, Lijun Wang, Huchuan Lu, Feng Zhao, Yu Qiao, and Jing Shao. 2024. Psysafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety. *arXiv preprint arXiv:2401.11880*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.