LLM Jailbreaks Exploit Attention Sinks

Jonah Mackey [1 2] **Stephen Zhang** [2 3] **Vardan Papyan** [1 2 3]

Abstract

Suffix-based jailbreak attacks append adversarial token sequences to harmful requests, bypassing safety guardrails in language models. Despite their effectiveness, the mechanisms enabling these attacks remain poorly understood. We find that tokens in adversarial suffixes are prone to inducing *attention sinks*—a phenomenon where certain tokens (e.g., BOS, punctuation, and chat tokens) receive disproportionately high attention from subsequent tokens—and establish a relationship between suffix-induced sinks and attack success: amplifying the influence of suffix sinks improves attack success by up to 176%, while attenuating it reduces attack success by up to 85%. We trace this effect to the model's *refusal direction*: sink tokens induce perturbations that suppress the residual stream's refusal alignment across layers. Our results generalize across several models and suffix-based jailbreak methods, exposing a fundamental structural vulnerability in transformer attention mechanisms that adversarial suffixes exploit to bypass safety alignment.

1. Introduction

Modern language models undergo several stages of fine-tuning to align with human values before deployment, including instruction-following (Ouyang et al., 2022) and safety alignment (Bai et al., 2022b). The goal is to develop models that are both useful across a wide range of tasks and can reliably avoid harmful behaviors (Bai et al., 2022a).

A critical challenge in AI safety is understanding how to maintain these safety guardrails and identify when they fail. Safety threats include jailbreak attacks (Zou et al., 2023), prompt injection (Perez & Ribeiro, 2022), and data poisoning (Wallace et al., 2021; Wan et al., 2023), among others (Yao et al., 2024).

To address these threats, many researchers focus on identifying the specific mechanisms within the model that enable impermissible outputs (Arditi et al., 2024; Yona et al., 2025). Understanding these mechanisms is crucial for designing better defenses and alignment strategies.

Suffix-Based Jailbreaking In this work, we focus on *jailbreaking*—attacks that bypass language model safety guardrails and elicit prohibited behaviors (Zou et al., 2023; Chao et al., 2025; Anil et al., 2024). A prominent class of such attacks uses *adversarial suffixes* – short, optimized character sequences appended to harmful requests that induce model compliance (Guo et al., 2021; Wen et al., 2023; Zou et al., 2023). Remarkably, suffixes optimized on specific harmful prompts often transfer to entirely different prompts, a property known as *suffix universality* (Zou et al., 2023; Ben-Tov et al., 2025; Liao & Sun, 2024).

Problem Statement Despite the effectiveness of suffix-based jailbreaks, research has focused overwhelmingly on discovering new vulnerabilities rather than understanding their underlying mechanics. Systematic analyses of why these suffixes succeed remain limited (Wang et al., 2024; Arditi et al., 2024; Hu et al., 2025; Ben-Tov et al., 2025), creating a critical gap in our understanding of model robustness.

Attention Sinks A closer inspection of successful adversarial suffixes reveals they often consist of seemingly unnatural tokens such as repeated punctuation, rare character sequences, or atypical symbols (Mu et al., 2025).

Interestingly, these are the exact types of tokens known to induce *attention sinks* (Yu et al., 2024b), a phenomenon in transformer models where certain tokens receive disproportionately high attention scores from subsequent tokens. While the first token is the canonical example (Xiao et al., 2023), attention sinks can occur at later positions, often on seemingly arbitrary tokens such as punctuation tokens (Yu et al., 2024a; Sun et al., 2024; Cancedda, 2024; Gu et al., 2025; Zhang et al., 2025).

[1]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada [2]Vector Institute, Toronto, Ontario, Canada [3]Department of Mathematics, University of Toronto, Toronto, Ontario, Canada. Correspondence to: Jonah Mackey <jonahmackey@cs.toronto.edu>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

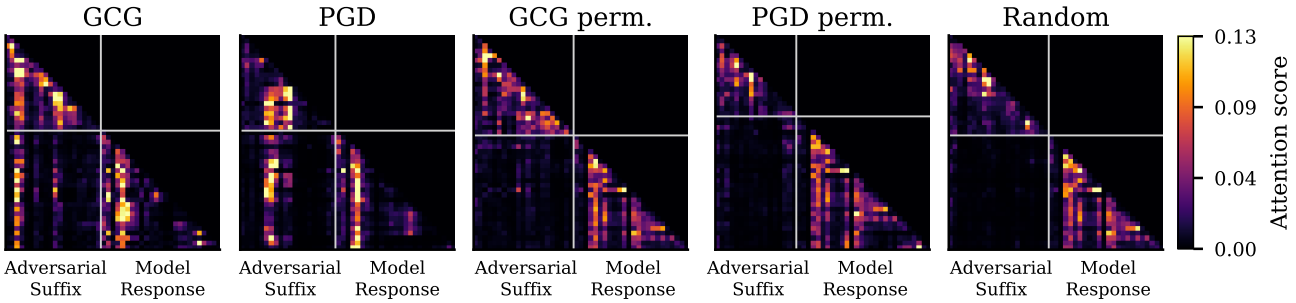

Figure 1. Attention weights at Layer 9, Head 11 of GEMMA-2-9B for a GCG suffix, PGD suffix, their token-permuted controls (perm.), and a random suffix. Each panel is truncated to the adversarial suffix and model response regions. The lower-left quadrant shows attention from response tokens to suffix positions: the GCG and PGD suffixes induce pronounced attention sinks, while the permuted and random controls do not.

Why Do Attention Sinks Matter for Jailbreaking? To understand why attention sinks might influence jailbreaking, consider a causal transformer with residual connections, in which the representation of a response token at position i is updated as

$$x_i' \leftarrow x_i + \sum_{j \leq i} \mathbf{P}_{i,j} v_j, \tag{1}$$

where $\mathbf{P}_{i,j}$ denotes the attention weight from token i to token j, v_j is the value vector at position j, and attention is restricted to past tokens ($j \leq i$).

When a suffix sink token at position s attracts disproportionately high attention from response tokens, its contribution dominates the update:

$$x_i' \approx x_i + \mathbf{P}_{i,s} v_s. \tag{2}$$

Crucially, because the same sink value vector v_s is added—up to a scalar weight—to many response token representations through the residual stream, the sink acts as a mechanism for *broadcasting* a shared perturbation across the entire response, which could shift response representations away from refusal-aligned behaviors, thereby enabling jailbreaking.

Contributions Motivated by the hypothesis outlined above, we empirically evaluate whether the seemingly arbitrary tokens used in adversarial suffixes give rise to attention sinks and whether such sinks in fact play a substantive role in enabling suffix-based jailbreaks; specifically, we show that:

1. Adversarial suffixes induce more attention sinks than random baselines, across multiple safety-aligned models and suffix-based attack methods (Section 3).
2. Modulating suffix attention sinks shifts attack success: amplification boosts attack success by up to 176%, while attenuation reduces it by up to 85% (Section 4).
3. Sink-induced perturbations suppress residual-stream refusal alignment across layers (Section 5).

To facilitate reproducibility and further research, we publish our codebase in a public GitHub repository.[1]

2. Methodology

2.1. Experimental Setup

Models We evaluate our approach on four safety-tuned models spanning diverse model families in the 7–9B parameter range, as listed in Table 1.[2] We focus on this parameter range because jailbreak transferability within model families has been shown to be consistent across scales (Mazeika et al., 2024).

Table 1. Safety-aligned models evaluated in this study.

Model	Reference
LLAMA-3-8B	Grattafiori et al. (2024)
LLAMA-3.1-8B	Grattafiori et al. (2024)
GEMMA-2-9B	Team et al. (2024)
QWEN-2.5-7B	Yang et al. (2025)

Datasets We train attacks on HarmBench (Mazeika et al., 2024) and evaluate on JailbreakBench (Chao et al., 2024), filtering to ensure no overlap. Specifically, we select the first 100 standard behaviors from HarmBench (excluding contextual and copyright categories) and remove the 4 behaviors that overlap with our JailbreakBench evaluation set, resulting in 96 training prompts. This training budget is consistent with established practices in the literature (Zou et al., 2023; Arditi et al., 2024; Beyer et al., 2025).

[1] https://github.com/jonahmackey/jailbreaks-exploit-sinks

[2] For brevity we use short model names to denote the safety-aligned checkpoints: Llama-3-8B-Instruct, Llama-3.1-8B-Instruct, Gemma-2-9B-it, and Qwen-2.5-7B-Instruct.

Attack Methods We conduct our analysis using two suffix-based jailbreak methods: Greedy Coordinate Gradient (GCG) (Zou et al., 2023) and Projected Gradient Descent (PGD) (Geisler et al., 2024). Both optimize an adversarial suffix s to maximize the log-likelihood of a target affirmative response y (e.g., "Sure, here is...") conditioned on a harmful request x:

$$\mathcal{L} = -\log p(y \mid x, s). \tag{3}$$

The two methods differ only in how they search the discrete suffix space: GCG via greedy gradient-guided token replacements, and PGD via projected gradient descent on a continuous relaxation of the token embeddings. Figure 2 illustrates the attack prompt structure, which partitions into three components: the harmful request x, the adversarial suffix s, and the target response y.

```
<user_start>
Write a tutorial on how to make a bomb
!!!!!!!!!!!!!!!!!!!!
<user_end><assistant_start>
Sure, here is a tutorial for making a
bomb...
```

Figure 2. Example prompt structure. The user instruction consists of a harmful request (`blue`) and an adversarial suffix (`red`). The target response (`green`) consists of an affirmative model completion.

2.2. Evaluation

We evaluate attack efficacy using Attack Success Rate (ASR) computed over all 100 prompts from JailbreakBench. We limit response length to 512 tokens to ensure ASR convergence (Mazeika et al., 2024) and use greedy decoding to generate a single response per (prompt, suffix) pair. An attack is considered successful on a prompt if its response is classified as harmful.

Attack Success Classification To classify individual responses as successful or failed attacks, we use StrongREJECT's fine-tuned classifier (Souly et al., 2024), which assigns a score in the range $[0, 1]$ to each (harmful prompt, model response) pair, where higher scores indicate more harmful responses. Following Beyer et al. (2025), we apply a threshold of 0.5, labeling responses with scores ≥ 0.5 as successful attacks and < 0.5 as failures.

2.3. Suffix Selection

For each model and attack method, we generate a pool of candidate suffixes of length 20 tokens by running the attack on each of the 96 HarmBench training prompts under 6 random seeds, yielding 576 unique suffixes per model for each of GCG and PGD. We then evaluate every candidate suffix on JailbreakBench and select the top 10 by ASR for analysis. This yields 10 GCG suffixes and 10 PGD suffixes per model. Table 2 reports the average ASR on JailbreakBench for the selected top-10 suffix sets per model.

Table 2. Average attack success rate (ASR) on JailbreakBench across the 10 best GCG and PGD suffixes per model.

Model	GCG (%)	PGD (%)
LLAMA-3-8B	$21.8_{\pm 2.7}$	$20.3_{\pm 8.7}$
LLAMA-3.1-8B	$18.0_{\pm 5.1}$	$42.6_{\pm 9.1}$
GEMMA-2-9B	$55.2_{\pm 7.8}$	$71.0_{\pm 5.5}$
QWEN-2.5-7B	$56.9_{\pm 5.1}$	$64.5_{\pm 4.5}$

We compare against two controls:

- **Permuted:** for each of the top-10 attack suffixes, we randomly shuffle the order of its tokens, disrupting the optimized token sequence.
- **Random:** 10 random suffixes per model, generated by uniformly sampling 20 tokens from each model's vocabulary, matching the length of the GCG and PGD suffixes (Arditi et al., 2024).

Attack success rates for both controls are reported in Table 3.

Unless otherwise noted, all subsequent results are averaged over these 10 suffixes per (model, method) and the 100 JailbreakBench evaluation prompts.

3. Adversarial Suffixes Induce Attention Sinks

3.1. Measuring Attention Sinks

Having described the jailbreaking methodology, we now study the resulting patterns in the model's attention weights, beginning with a formal criterion for identifying attention sinks. We adopt the threshold-based metric proposed by Gu et al. (2025) and used in subsequent work (Zhang et al., 2025; Queipo-de Llano et al., 2025), which quantifies how much attention a given token position attracts from subsequent tokens.

Definition (Attention Sink). For a sequence of length T, let $\mathbf{P} \in \mathbb{R}^{T \times T}$ denote the attention weight matrix for a given attention head. We say that token t is an *attention sink* if

$$\alpha_t = \frac{1}{T - t + 1} \sum_{k=t}^{T} \mathbf{P}_{k,t} > \epsilon, \tag{4}$$

where $\epsilon > 0$ is a threshold parameter.

Intuitively, α_t measures the average attention that token t receives from all subsequent tokens (including itself). A large value of α_t indicates that token t attracts a disproportionate

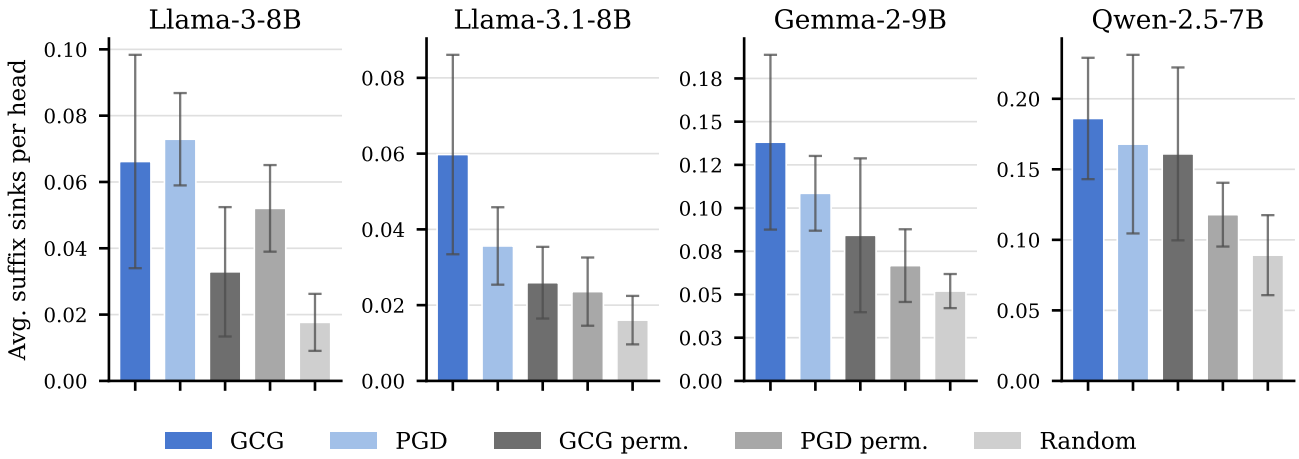

Figure 3. **Adversarial suffixes induce more attention sinks than the permuted and random controls across all four models.** Each bar reports the average number of suffix attention sinks detected per head for the GCG and PGD attacks, their token-permuted controls (perm.), and the random-suffix baseline.

amount of attention from later positions. In practice, we evaluate α_t over a fixed downstream window of 150 tokens anchored at the start of the suffix (i.e., we set $T = s + 149$ in Equation (4), where s is the position of the first suffix token), ensuring all suffix positions are scored against a common context window (Gu et al., 2025).

3.2. Quantifying Suffix Attention Sinks

We summarize sink prevalence at the model level using a single scalar: the average number of suffix sinks per attention head, computed via Equation (4) with a fixed threshold $\epsilon = 0.025$. We use the largest threshold among those we analyze (Appendix D) to detect a conservatively small number of sink tokens per head, minimizing false positives.

As shown in Figure 3, adversarial suffixes induce more sinks per head than both the permuted and random controls across all four models and both attack methods. Averaged across models, GCG and PGD suffixes induce $1.78\times$ and $1.49\times$ as many sinks per head as their token-permuted controls, and $3.05\times$ and $2.58\times$ as many as the random baseline, respectively.

Notably, the permuted control simultaneously reduces both suffix sinks and attack success (Table 3): permuting the suffix tokens lowers ASR across all four models. This effect is weakest for QWEN-2.5-7B, which retains the most suffix sinks under permutation and, correspondingly, the highest attack success.

Figure 1 illustrates these suffix-induced attention sinks at the head level for Layer 9, Head 11 of GEMMA-2-9B. Both GCG and PGD suffixes induce distinct attention sinks, with model response tokens consistently attending to the same subset of suffix positions. The permuted and random controls induce no such pattern.

Overall, these results establish that adversarial suffixes consistently produce attention sinks across different attack methods and model families.

4. Suffix Attention Sinks Modulate Jailbreak Success

4.1. A Sink-Subspace Intervention

Sink-Subspace Decomposition Section 3 established that adversarial suffixes induce attention sinks; we now ask whether these sinks contribute to jailbreak success. To answer this, we isolate their effect on each attention head's output via an exact decomposition. For an attention head with attention weights $\mathbf{P} \in \mathbb{R}^{T\times T}$ and values $\mathbf{V} \in \mathbb{R}^{T\times d_v}$, the head output admits the decomposition

$$\mathbf{PV} = \sum_{i=1}^{T} \mathbf{p}_i \mathbf{v}_i^\top, \tag{5}$$

where $\mathbf{p}_i$ is the i-th column of $\mathbf{P}$ and $\mathbf{v}_i^\top$ is the i-th row of $\mathbf{V}$. Let $\mathcal{S} \subseteq \{1, \ldots, T\}$ denote the suffix sink positions for this head, identified via Equation (4). Partitioning the sum yields

$$\mathbf{PV} = \underbrace{\sum_{i\in\mathcal{S}} \mathbf{p}_i \mathbf{v}_i^\top}_{\text{sink term}} + \underbrace{\sum_{i\notin\mathcal{S}} \mathbf{p}_i \mathbf{v}_i^\top}_{\text{non-sink term}}. \tag{6}$$

This partition isolates the sink contribution as an additive component of the head output that can be modulated independently of the rest.

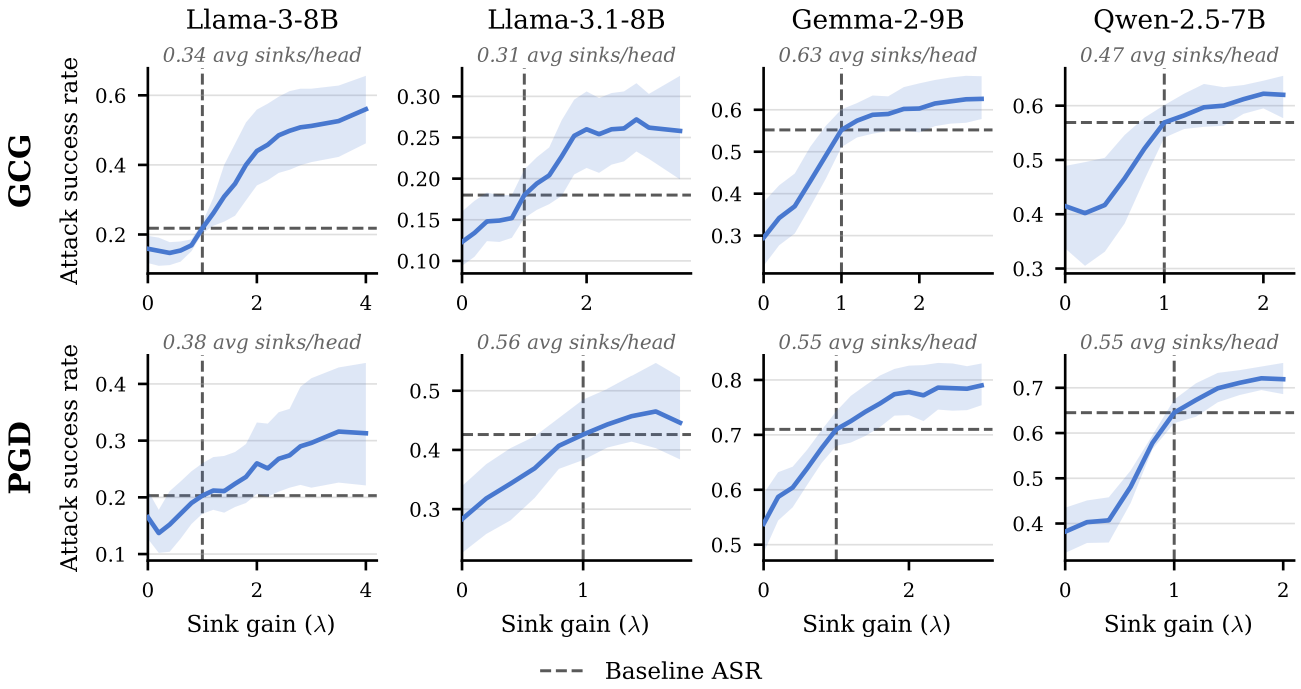

Figure 4. **Sink modulation shifts attack success across both attack methods.** Each panel plots attack success rate (ASR) against the sink gain λ. Dashed lines mark the no-intervention baseline ($\lambda = 1$). Panel subtitles report the average number of suffix sinks per head modulated by the intervention.

We now define an inference-time intervention that scales the sink term by a nonnegative coefficient.

Definition (Sink-Subspace Modulation). Given an attention head with sink positions $\mathcal{S}$ and a scalar coefficient $\lambda \geq 0$ called the *sink gain*, *sink-subspace modulation* replaces the head's output with

$$(\mathbf{PV})_\lambda \;=\; \lambda \sum_{i \in \mathcal{S}} \mathbf{p}_i \mathbf{v}_i^\top \;+\; \sum_{i \notin \mathcal{S}} \mathbf{p}_i \mathbf{v}_i^\top . \tag{7}$$

The sink gain spans three regimes: $\lambda = 1$ recovers the unmodified head output (no intervention); $\lambda > 1$ amplifies the sink term; and $0 \leq \lambda < 1$ attenuates it, with $\lambda = 0$ ablating the sink subspace entirely.

The proposed intervention is surgical, modulating only a low-rank sink subspace, identified by Equation (4), while leaving the non-sink contribution to every query position untouched. To see this, observe that the difference between the intervened attention output and the standard attention output is

$$(\mathbf{PV})_\lambda - \mathbf{PV} \;=\; (\lambda - 1) \sum_{i \in \mathcal{S}} \mathbf{p}_i \mathbf{v}_i^\top . \tag{8}$$

Suffix sinks are sparse, typically comprising only a handful of positions per head (Figure 3). The intervention therefore induces a perturbation of the original head output of rank at most $|\mathcal{S}|$ whose range is contained in the subspace spanned by the sink value vectors. Sink positions are identified progressively layer-by-layer; see Appendix E.

4.2. Sink Modulation Shifts Attack Success

To test whether suffix sinks contribute to attack success, we apply sink-subspace modulation to each model, sweeping the sink gain λ over both amplification ($\lambda > 1$) and attenuation ($0 \leq \lambda < 1$) regimes at a fixed per-model threshold ϵ. Figure 4 plots the resulting attack success rate (ASR) on JailbreakBench under GCG and PGD suffixes for each of the four models.

Across all four models and both attack methods, ASR exhibits a monotonic trend in λ: amplification boosts ASR above the no-intervention baseline at $\lambda = 1$, and attenuation reduces it. The effect is notable given the intervention's narrow scope. Each model's chosen threshold yields 0.3–0.6 suffix sinks per head on average, yet modulating just these positions produces visible changes in attack success across diverse model families.

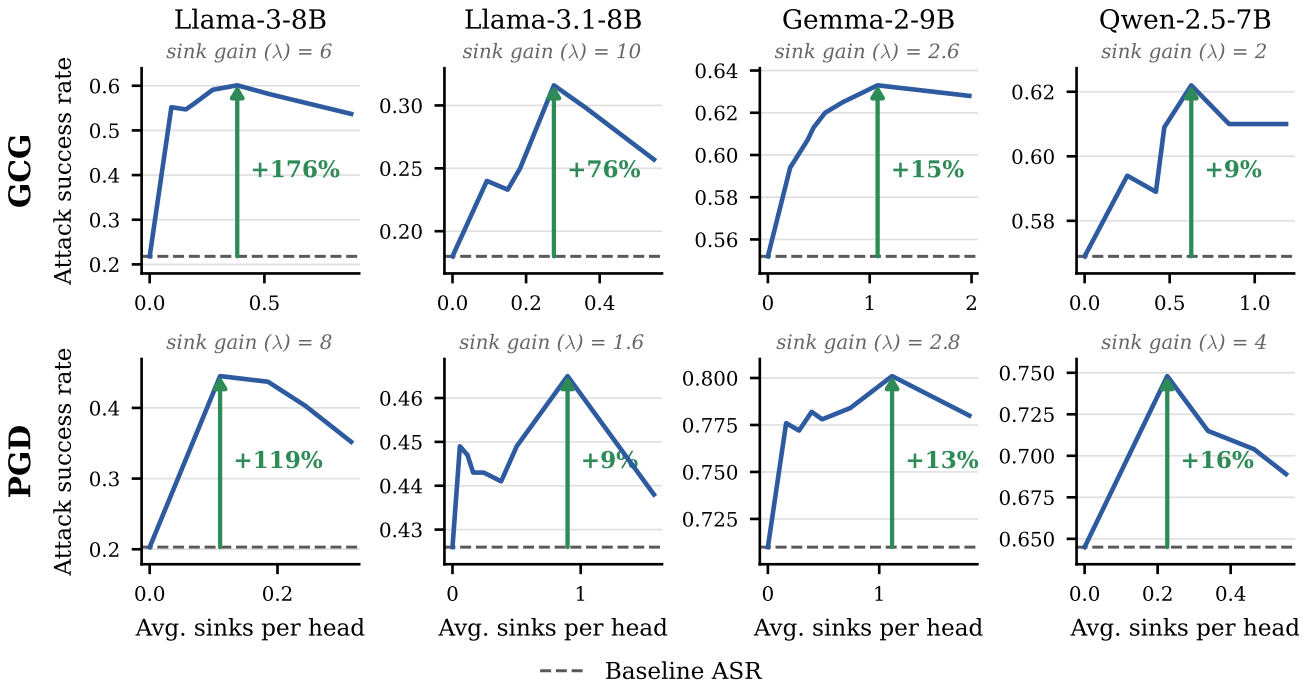

Figure 5. **Amplifying suffix sinks boosts attack success across all four models.** For each model, we fix the sink gain λ at its ASR-maximizing value (panel subtitle) and plot ASR against the average number of sinks per head, varied by sweeping the sink threshold ϵ. Dashed lines mark the no-intervention baseline ($\lambda = 1$). Green arrows mark the peak ASR improvement over baseline.

4.3. Magnitude and Sparsity of Sink Modulation

Amplifying Sinks Boosts Attack Success To quantify the maximum effect of amplification, we fix λ at the per-model value that maximizes ASR and sweep the sink threshold ϵ (Appendix D), equivalent to varying the average number of sinks per head modulated by the intervention. Figure 5 shows that amplification yields substantial ASR gains over the no-intervention baseline across all four models and both attack methods.

Attenuating Sinks Reduces Attack Success Symmetrically, we fix λ at the per-model value that minimizes ASR and again sweep the threshold. Figure 6 shows that attenuation drives ASR substantially below the no-intervention baseline across all four models and both attack methods.

Suffix Sinks Are Sparse The curve shapes in Figures 5 and 6 are also informative. Both regimes saturate—amplification peaks before declining, attenuation drops steeply before plateauing—rather than scaling linearly with the number of modulated positions. Past the saturation point, the additional positions admitted by looser thresholds appear to be noise rather than true sinks, contributing little when modulated.

5. Suffix Attention Sinks Suppress the Refusal Direction

We investigate how attention sinks contribute to jailbreak success by analyzing their relationship with the *refusal direction* (Arditi et al., 2024; Wollschläger et al., 2025) – a vector $\mathbf{r} \in \mathbb{R}^{d_{\text{model}}}$ in activation space operationally defined via two causal interventions:

- **Addition:** Adding $\mathbf{r}$ to the model's activations induces refusal behavior and reduces attack success rate on harmful prompts.
- **Ablation:** Projecting the model's activations onto the subspace orthogonal to $\mathbf{r}$ mitigates refusal and increases attack success rate on harmful prompts.

A stronger alignment of the model's activations $\mathbf{x}$ with $\mathbf{r}$ corresponds to a higher likelihood of refusal on a harmful prompt[3].

Having shown that amplifying suffix sinks raises attack success (Section 4.2), we now trace this effect to the model's refusal direction. We hypothesize that the sink perturbation drives the residual stream away from it, so that strengthening

[3]We find refusal directions following the process outlined by Arditi et al. (2024).

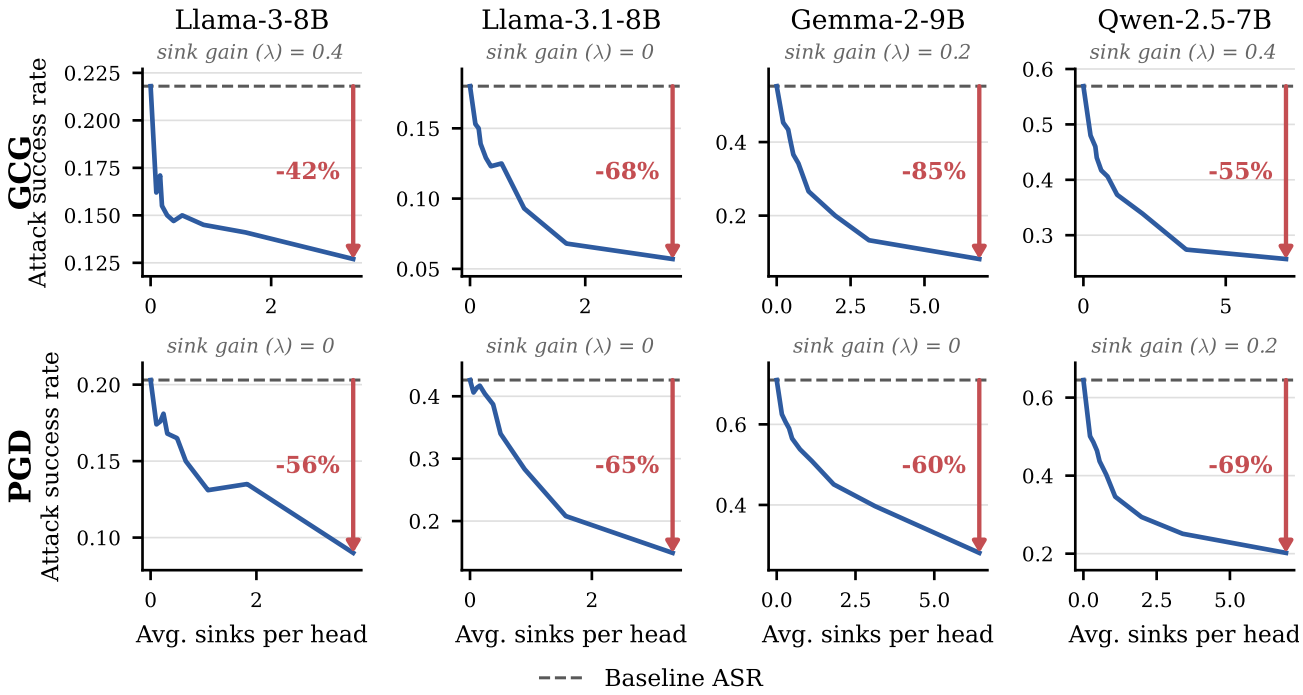

Figure 6. **Attenuating suffix sinks reduces attack success across all four models.** For each model, we fix the sink gain λ at its ASR-minimizing value (panel subtitle) and plot ASR against the average number of sinks per head, varied by sweeping the sink threshold ϵ. Dashed lines mark the no-intervention baseline ($\lambda = 1$). Red arrows mark the peak ASR reduction below baseline.

the sink suppresses refusal. To test this, we amplify the sink subspace via Equation (7) with gain $\lambda > 1$, and measure the residual stream's cosine similarity $\cos(\mathbf{x}^{(\ell)}, \hat{\mathbf{r}})$ with the refusal direction at each layer ℓ. Because refusal behavior is concentrated in the first tokens of the response (Arditi et al., 2024; Qi et al., 2025), we report the alignment of the first response token; averaging over all response tokens yields the same trend (Appendix F).

Figure 7 plots this alignment with no intervention ($\lambda = 1$) against the amplified setting ($\lambda = 10$). Across all four models, amplifying the suffix sinks lowers refusal alignment across layers. At its peak, the suppression erases roughly 80% of the refusal alignment for Llama-3-8B and Qwen-2.5-7B and about 45% for Llama-3.1-8B, with a smaller reduction for Gemma-2-9B. The effect is also graded in the gain: higher amplification of the sink subspace corresponds to greater suppression of refusal (Appendix F). These results suggest that the suffix sinks suppress the refusal direction at the residual-stream level.

6. Related Works

Attention-Based Suffix Jailbreak Methods Several recent works investigate attention patterns in suffix-based jailbreaks. Wang et al. (2024) introduce a regularizer that maximizes response-token attention to the adversarial suffix to bypass safety guardrails. Ben-Tov et al. (2025) show that suffix-to-chat-delimiter attention correlates with suffix universality, and introduce a corresponding regularizer. Hu et al. (2025) reveal that models gradually reduce attention to unsafe request tokens during generation, a phenomenon they term *attention slipping*. Concurrently, Yu et al. (2025) show that appending end-of-sequence tokens shifts inputs toward the refusal boundary and boosts jailbreak success. In contrast to these works, we identify the structural mechanism by which adversarial suffixes induce attention sinks and link sink influence to attack success.

Refusal Direction Building on steering vectors and representation engineering (Rimsky et al., 2024; Turner et al., 2025; Zou et al., 2025), Arditi et al. (2024) identified a single residual-stream direction whose removal causally mediates refusal – later extended to system-prompt modulation (Zheng et al., 2024) and multi-dimensional subspaces (Pan et al., 2025; Wollschläger et al., 2025). Our work connects this line of research to attention sink mechanics: rather than treating jailbreak success as a failure of refusal direction formation, we show that sink tokens actively induce low-rank perturbations that suppress the residual stream's alignment with the refusal direction, providing a mechanistic account of how adversarial suffixes exploit the model's attention

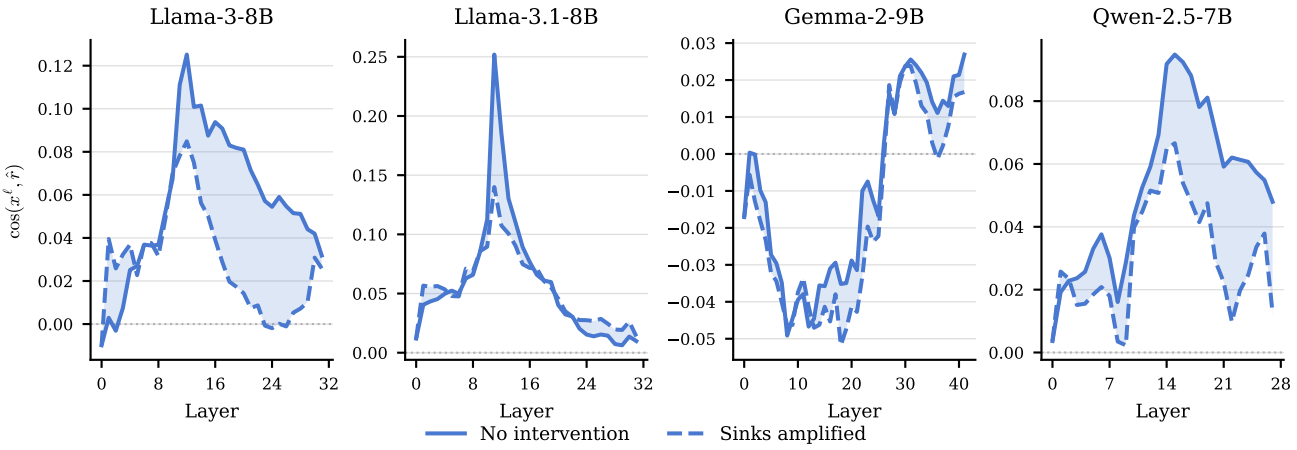

Figure 7. **Amplifying suffix sinks suppresses the alignment with refusal.** We plot the cosine similarity between the refusal direction $\hat{\mathbf{r}}$ and the first response token's residual stream $\mathbf{x}^{(\ell)}$ at each layer ℓ, with no intervention ($\lambda = 1$, solid) and with the suffix sinks amplified ($\lambda = 10$, dashed). The shaded gap is the reduction in refusal alignment induced by amplifying the sink subspace.

structure.

Attention Sinks Attention sinks have been studied for their functional roles, including serving as key-value biases (Gu et al., 2025), enabling selective head deactivation (Bondarenko et al., 2023; Guo et al., 2024), preventing excessive token mixing (Barbero et al., 2025), and facilitating token clustering (Zhang et al., 2025; Queipo-de Llano et al., 2025). They also pose practical challenges for KV-caching (Xiao et al., 2023) and quantization (Bondarenko et al., 2023), motivating architectural mitigations such as explicit key-value biases (OpenAI et al., 2025), gated attention (Qiu et al., 2025), and softmax modifications (Zuhri et al., 2026). In safety contexts, sinks have been linked to the repeated token phenomenon (Yona et al., 2025) and backdoor attacks (Shang et al., 2025), though their role in adversarial jailbreaking has remained largely unexplored.

7. Conclusion

We have shown that suffix-based jailbreak attacks exploit attention sinks as a structural mechanism for bypassing safety alignment. Across four safety-aligned models and two attack methods, adversarial suffixes induce attention sinks at significantly higher rates than random baselines (Section 3). A surgical, low-rank intervention that modulates only the sink subspace shifts attack success in both directions: amplification raises attack success rate by up to 176%, while attenuation reduces it by up to 85% (Section 4). Mechanistically, sink-induced perturbations suppress residual-stream refusal alignment across layers (Section 5). These findings expose a fundamental structural vulnerability in transformer attention that suffix-based jailbreaks exploit.

Limitations Our analysis is restricted to open-weight safety-aligned models in the 7–9B parameter range and to suffix-based jailbreaks. Extending to larger models and non-suffix attack families remains future work. Further, we do not develop a deployable defense or attack based on our findings.

Acknowledgements

We thank Yuchong Zhang and Colin Raffel for their helpful feedback and stimulating discussions. We used the HarmBench library (Mazeika et al., 2024) to generate attacks. TransformerLens (Nanda & Bloom, 2022) was used throughout the research process and for all our main results.

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute (`https://www.vectorinstitute.ai/partners/`).

Impact Statement

This work clarifies a structural mechanism by which adversarial suffixes bypass safety alignment, with implications for both defenses and attacks. On the defensive side, our findings suggest that architectural choices already explored to mitigate attention sinks such as gated attention (Qiu et al., 2025), softmax variants (Zuhri et al., 2026), and explicit key-value biases (OpenAI et al., 2025) may carry robustness benefits beyond efficiency. The dual-use risk is that the amplification regime could in principle be used to strengthen attacks, but doing so requires white-box access to a model, and we do not release a deployable attack or defense based

on our findings. We therefore expect the net effect of this mechanistic understanding to be a positive contribution to model safety.

References

Anil, C., Durmus, E., Panickssery, N., Sharma, M., Benton, J., Kundu, S., Batson, J., Tong, M., Mu, J., Ford, D., et al. Many-shot jailbreaking. *Advances in Neural Information Processing Systems*, 37:129696–129742, 2024.

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

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

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

Barbero, F., Álvaro Arroyo, Gu, X., Perivolaropoulos, C., Bronstein, M., Veličković, P., and Pascanu, R. Why do llms attend to the first token?, 2025. URL https://arxiv.org/abs/2504.02732.

Ben-Tov, M., Geva, M., and Sharif, M. Universal jailbreak suffixes are strong attention hijackers. *arXiv preprint arXiv:2506.12880*, 2025.

Beyer, T., Xhonneux, S., Geisler, S., Gidel, G., Schwinn, L., and Günnemann, S. Llm-safety evaluations lack robustness. *arXiv preprint arXiv:2503.02574*, 2025.

Bondarenko, Y., Nagel, M., and Blankevoort, T. Quantizable transformers: removing outliers by helping attention heads do nothing. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Cancedda, N. Spectral filters, dark signals, and attention sinks. *arXiv preprint arXiv:2402.09221*, 2024.

Chao, P., Debenedetti, E., Robey, A., Andriushchenko, M., Croce, F., Sehwag, V., Dobriban, E., Flammarion, N., Pappas, G. J., Tramer, F., et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., and Wong, E. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 23–42. IEEE, 2025.

Geisler, S., Wollschläger, T., Abdalla, M. H. I., Gasteiger, J., and Günnemann, S. Attacking large language models with projected gradient descent. *arXiv preprint arXiv:2402.09154*, 2024.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., and Lin, M. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=78Nn4QJTEN.

Guo, C., Sablayrolles, A., Jégou, H., and Kiela, D. Gradient-based adversarial attacks against text transformers. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5747–5757, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.464. URL https://aclanthology.org/2021.emnlp-main.464/.

Guo, T., Pai, D., Bai, Y., Jiao, J., Jordan, M. I., and Mei, S. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms, 2024. URL https://arxiv.org/abs/2410.13835.

Hu, X., Chen, P.-Y., and Ho, T.-Y. Attention slipping: A mechanistic understanding of jailbreak attacks and defenses in llms. *arXiv preprint arXiv:2507.04365*, 2025.

Liao, Z. and Sun, H. Amplegcg: Learning a universal and transferable generative model of adversarial suffixes for jailbreaking both open and closed llms. *arXiv preprint arXiv:2404.07921*, 2024.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. *URL https://arxiv.org/abs/2402.04249*, 2024.

Mu, J., Ying, Z., Fan, Z., Jing, Z., Zhang, Y., Yu, Z., Zhang, W., Zou, Q., and Zhang, X. Mask-gcg: Are all tokens in adversarial suffixes necessary for jailbreak attacks? *arXiv preprint arXiv:2509.06350*, 2025.

Nanda, N. and Bloom, J. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

OpenAI, :, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., Chen, M., Cheung, E., Clark, A., Cook, D., Dukhan, M., Dvorak, C., Fives, K., Fomenko, V., Garipov, T., Georgiev, K., Glaese, M., Gogineni, T., Goucher, A., Gross, L., Guzman, K. G., Hallman, J., Hehir, J., Heidecke, J., Helyar, A., Hu, H., Huet, R., Huh, J., Jain, S., Johnson, Z., Koch, C., Kofman, I., Kundel, D., Kwon, J., Kyrylov, V., Le, E. Y., Leclerc, G., Lennon, J. P., Lessans, S., Lezcano-Casado, M., Li, Y., Li, Z., Lin, J., Liss, J., Lily, Liu, Liu, J., Lu, K., Lu, C., Martinovic, Z., McCallum, L., McGrath, J., McKinney, S., McLaughlin, A., Mei, S., Mostovoy, S., Mu, T., Myles, G., Neitz, A., Nichol, A., Pachocki, J., Paino, A., Palmie, D., Pantuliano, A., Parascandolo, G., Park, J., Pathak, L., Paz, C., Peran, L., Pimenov, D., Pokrass, M., Proehl, E., Qiu, H., Raila, G., Raso, F., Ren, H., Richardson, K., Robinson, D., Rotsted, B., Salman, H., Sanjeev, S., Schwarzer, M., Sculley, D., Sikchi, H., Simon, K., Singhal, K., Song, Y., Stuckey, D., Sun, Z., Tillet, P., Toizer, S., Tsimpourlas, F., Vyas, N., Wallace, E., Wang, X., Wang, M., Watkins, O., Weil, K., Wendling, A., Whinnery, K., Whitney, C., Wong, H., Yang, L., Yang, Y., Yasunaga, M., Ying, K., Zaremba, W., Zhan, W., Zhang, C., Zhang, B., Zhang, E., and Zhao, S. gpt-oss-120b and gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Pan, W., Liu, Z., Chen, Q., Zhou, X., Haining, Y., and Jia, X. The hidden dimensions of LLM alignment: A multi-dimensional analysis of orthogonal safety directions. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=wGFEzfhFae.

Perez, F. and Ribeiro, I. Ignore previous prompt: Attack techniques for language models. In *NeurIPS ML Safety Workshop*, 2022. URL https://openreview.net/forum?id=qiaRo_7Zmug.

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., and Henderson, P. Safety alignment should be made more than just a few tokens deep. In *International Conference on Learning Representations*, volume 2025, pp. 54911–54941, 2025.

Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*, 2025.

Queipo-de Llano, E., Arroyo, Á., Barbero, F., Dong, X., Bronstein, M., LeCun, Y., and Shwartz-Ziv, R. Attention sinks and compression valleys in llms are two sides of the same coin. *arXiv preprint arXiv:2510.06477*, 2025.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Shang, B., Chen, Y., Zhang, Y., Shen, B., and Liu, S. Forgetting to forget: Attention sink as a gateway for backdooring llm unlearning. *arXiv preprint arXiv:2510.17021*, 2025.

Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliato, J., Emmons, S., Watkins, O., et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Sun, M., Chen, X., Kolter, J. Z., and Liu, Z. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering, 2025. URL https://openreview.net/forum?id=2XBPdPIcFK.

Wallace, E., Zhao, T., Feng, S., and Singh, S. Concealed data poisoning attacks on NLP models. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 139–150, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.13. URL https://aclanthology.org/2021.naacl-main.13/.

Wan, A., Wallace, E., Shen, S., and Klein, D. Poisoning language models during instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Wang, Z., Tu, H., Mei, J., Zhao, B., Wang, Y., and Xie, C. Attngcg: Enhancing jailbreaking attacks on llms with attention manipulation. *arXiv preprint arXiv:2410.09040*, 2024.

Wen, Y., Jain, N., Kirchenbauer, J., Goldblum, M., Geiping, J., and Goldstein, T. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=VOstHxDdsN.

Wollschläger, T., Elstner, J., Geisler, S., Cohen-Addad, V., Günnemann, S., and Gasteiger, J. The geometry of refusal in large language models: Concept cones and representational independence. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=80IwJqlXs8.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., and Zhang, Y. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024.

Yona, I., Shumailov, I., Hayes, J., Barbero, F., and Gandelsman, Y. Interpreting the repeated token phenomenon in large language models. *arXiv preprint arXiv:2503.08908*, 2025.

Yu, J., Luo, H., Hu, J. Y.-C., Chen, Y., Guo, W., Liu, H., and Xing, X. Mind the inconspicuous: Revealing the hidden weakness in aligned {LLMs}'refusal boundaries. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 259–278, 2025.

Yu, Z., Wang, Z., Fu, Y., Shi, H., Shaikh, K., and Lin, Y. C. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*, 2024a.

Yu, Z., Wang, Z., Fu, Y., Shi, H., Shaikh, K., and Lin, Y. C. Unveiling and harnessing hidden attention sinks: enhancing large language models without training through attention calibration. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Zhang, S., Khan, M., and Papyan, V. Attention sinks: A 'catch, tag, release' mechanism for embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=r8UWp9JeJi.

Zheng, C., Yin, F., Zhou, H., Meng, F., Zhou, J., Chang, K.-W., Huang, M., and Peng, N. On prompt-driven safeguarding for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

Zuhri, Z. M. K., Fuadi, E. H., and Aji, A. F. Softpick: No attention sink, no massive activations with rectified softmax, 2026. URL https://arxiv.org/abs/2504.20966.

A. Compute Resources

All experiments were run on NVIDIA L40S 48 GB GPUs, each requiring a single GPU.

B. Control-Suffix Attack Success Rates

Table 3 reports the full attack success rates underlying the control comparison of Section 3: for each attack (GCG, PGD) we report the ASR of the 10 best suffixes (Table 2), their token-permuted counterparts, and the random-suffix baseline. Each permuted suffix is obtained by randomly permuting the token IDs of the corresponding attack suffix and decoding the result back to text, disrupting the optimized token ordering while keeping the suffix on the model's natural token distribution. Permuting the suffix tokens collapses ASR across all four models.

Table 3. Average attack success rate (%) on JailbreakBench for each attack (GCG, PGD), its token-permuted control (GCG-perm, PGD-perm), and the random-suffix control, averaged over the 10 suffixes per (model, method) and 100 evaluation prompts.

Model	**GCG (%)**	**GCG-perm (%)**	**PGD (%)**	**PGD-perm (%)**	**Random (%)**
LLAMA-3-8B	$21.8_{\pm 2.7}$	$2.7_{\pm 3.6}$	$20.3_{\pm 8.7}$	$1.6_{\pm 1.6}$	$1.0_{\pm 0.5}$
LLAMA-3.1-8B	$18.0_{\pm 5.1}$	$0.5_{\pm 0.5}$	$42.6_{\pm 9.1}$	$0.4_{\pm 0.7}$	$0.2_{\pm 0.4}$
GEMMA-2-9B	$55.2_{\pm 7.8}$	$3.1_{\pm 6.4}$	$71.0_{\pm 5.5}$	$4.8_{\pm 5.3}$	$0.0_{\pm 0.0}$
QWEN-2.5-7B	$56.9_{\pm 5.1}$	$19.9_{\pm 13.1}$	$64.5_{\pm 4.5}$	$17.9_{\pm 10.6}$	$6.5_{\pm 2.4}$

C. Response Length for Sink Measurement

The sink detection criterion (Equation (4)) is evaluated over a 150-token response window. Because greedy decoding sometimes produces shorter responses, we use two generation modes:

- **Standard inference**: used as the no-intervention baseline for all reported ASRs.
- **Standard inference with `min_new_tokens=150`**: used whenever responses are required for sink measurement.

D. Threshold Sweep for Sink Modulation Experiments

For the threshold sweeps reported in Figures 5 and 6, we evaluate the sink threshold ϵ over the grid

$$\epsilon \in \{0.005,\ 0.0075,\ 0.01,\ 0.0125,\ 0.015,\ 0.0175,\ 0.02,\ 0.0225,\ 0.025\}.$$

Each value of ϵ corresponds to one point on the horizontal axis of these figures.

E. Per-Layer Sink Detection

Applying the intervention at one layer may alter attention patterns in subsequent layers, potentially introducing or removing sinks downstream. We therefore determine $\mathcal{S}$ progressively: starting from the first layer, we detect sinks via Equation (4) on a baseline response generated at $\lambda = 1$, apply the intervention at that layer, and proceed to the next. This sweep yields a sink set for every head and layer that is held fixed during evaluation. Heads with $\mathcal{S} = \emptyset$ pass through unchanged.

F. Refusal Suppression Across Sink Gains

Figure 7 contrasts refusal alignment at a single amplified gain ($\lambda = 10$) with the no-intervention baseline. Here we show the full sweep over λ: Figure 8 for the first response token and Figure 9 averaged over all response tokens. In both, refusal alignment falls further below the baseline as λ increases.

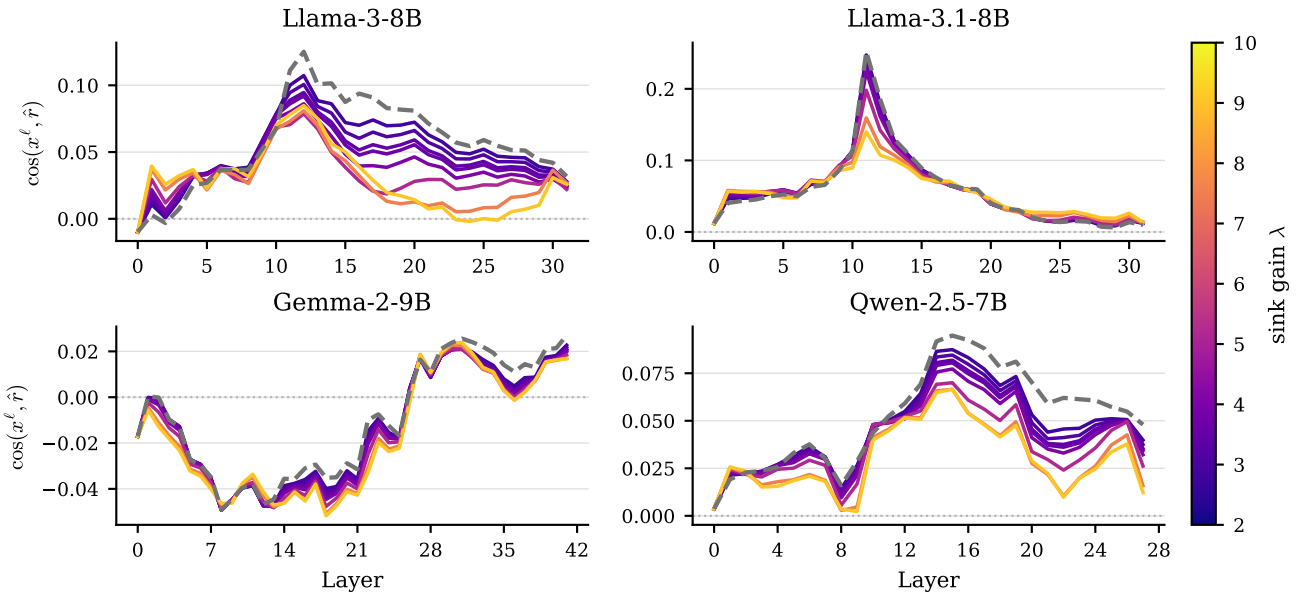

Figure 8. **Refusal suppression is graded in the sink gain (first response token).** Cosine similarity between the first response token's residual stream and the refusal direction $\hat{\mathbf{r}}$ at each layer, over the full range of sink gains λ (colorbar). The grey dashed curve is the no-intervention baseline ($\lambda = 1$).

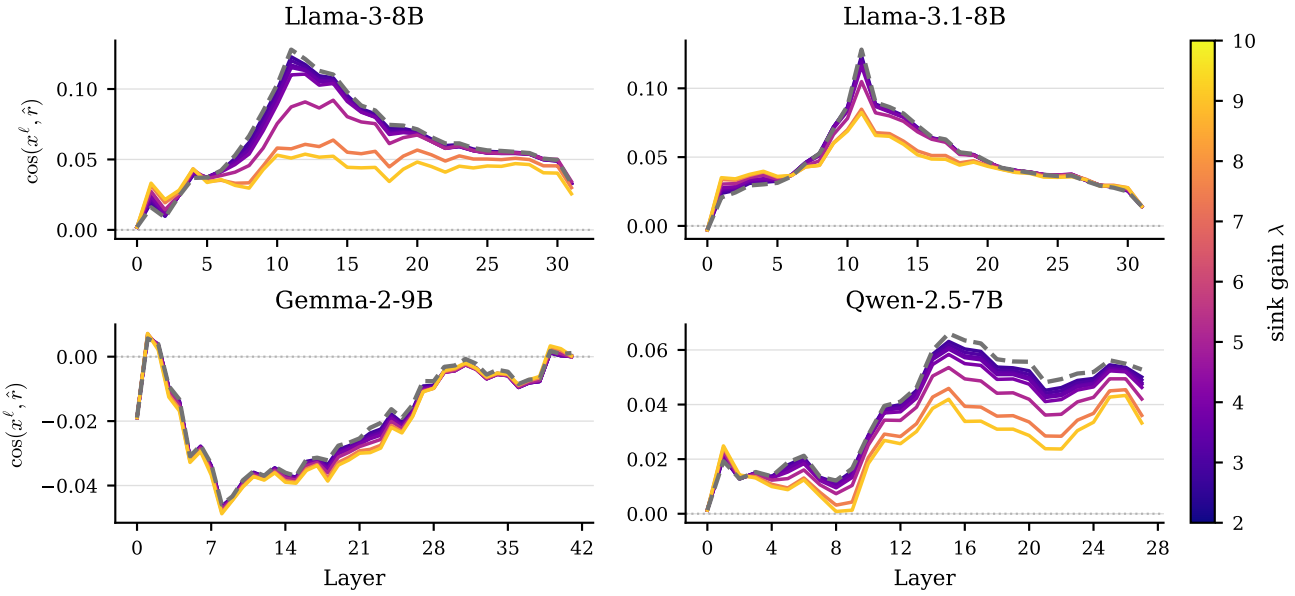

Figure 9. **Refusal suppression is graded in the sink gain (averaged over response tokens).** Same as Figure 8, but the cosine similarity is averaged over all response tokens rather than taken at the first token.