LLM Jailbreaks Exploit Attention Sinks

Anonymous Author(s)
Affiliation
Address
email

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 276%, while attenuating it reduces attack success by up to 84%. We trace this effect to the model's *refusal direction*: sink tokens induce perturbations aligned with the refusal direction, cumulatively suppressing 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.

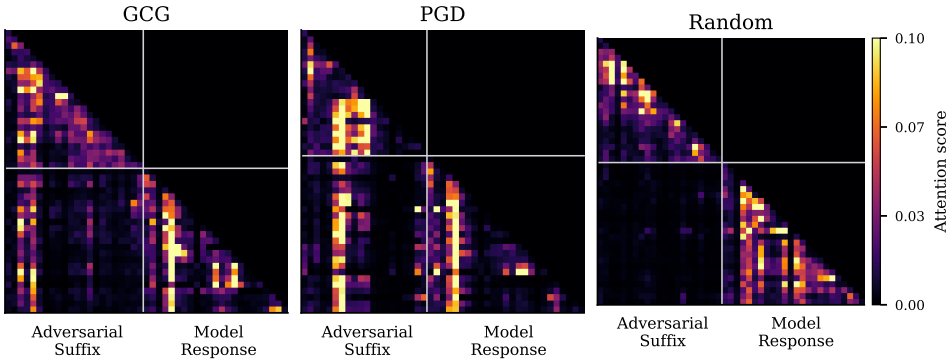

Figure 1: Attention weights at Layer 9, Head 11 of GEMMA-2-9B for a GCG suffix (left), PGD suffix (middle), and random suffix (right). The attention scores are truncated to display only the adversarial suffix and model response regions, with values clipped at 0.1 for visualization. The lower-left quadrant of each panel shows attention from response tokens to suffix positions: GCG and PGD induce distinct attention sinks, while the random baseline does not.

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].

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

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 and 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 and 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].

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 in early transformer layers 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 276%, while attenuation reduces it by up to 84%. (Section 4)

3. Sink-induced perturbations are aligned with the refusal direction and cumulatively suppress residual-stream refusal alignment across layers. (Section 5).

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. 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]
QWEN2.5-7B	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].

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:

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

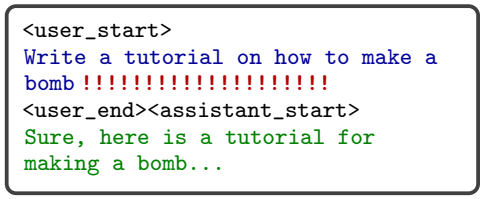

Figure 2: Example prompt structure. The user instruction consists of a harmful request (blue) and an adversarial suffix (red). The target response (green) consists of an affirmative model completion.

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.

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.

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	$19.2_{\pm 5.7}$	$12.8_{\pm 4.9}$
LLAMA-3.1-8B	$17.8_{\pm 5.4}$	$36.8_{\pm 11.8}$
GEMMA-2-9B	$53.5_{\pm 9.5}$	$66.1_{\pm 8.6}$
QWEN-2.5-7B	$55.3_{\pm 6.9}$	$56.5_{\pm 7.8}$

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 multiple random seeds: 6

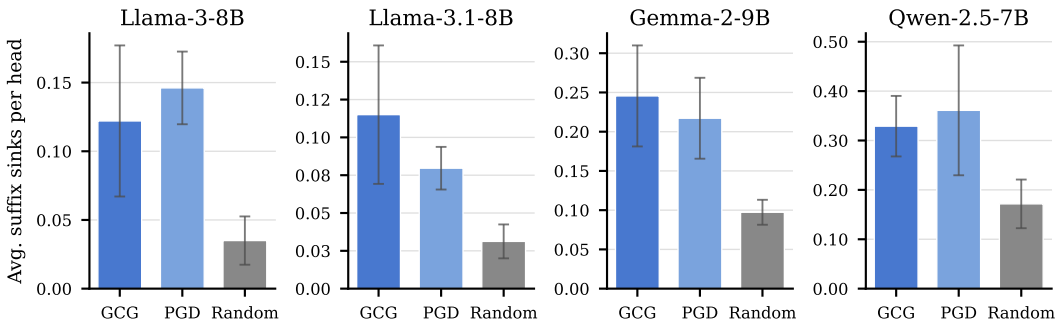

Figure 3: **Adversarial suffixes induce more attention sinks than random suffixes across all four models.** Each bar reports the average number of suffix attention sinks detected per head.

seeds for GCG (576 unique suffixes per model) and 3 seeds for PGD (288 unique suffixes per model). 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.

As a control, we generate 10 random suffixes per model by uniformly sampling 20 tokens from each model's vocabulary, matching the length of the GCG and PGD suffixes [Arditi et al., 2024].

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 (softmax probabilities) 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 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 $\epsilon = 0.02$. The threshold is chosen to detect a conservatively small number of sink tokens per head, minimizing false positives.

As shown in Figure 3, adversarial suffixes induce significantly more sinks per head than random suffixes of the same length across all four models and both attack methods. Averaged across models,

GCG and PGD suffixes induce $2.9\times$ and $2.76\times$ as many sinks per head as the random baseline, respectively.

Figure 1 illustrates this phenomenon 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 random suffix baseline induces 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

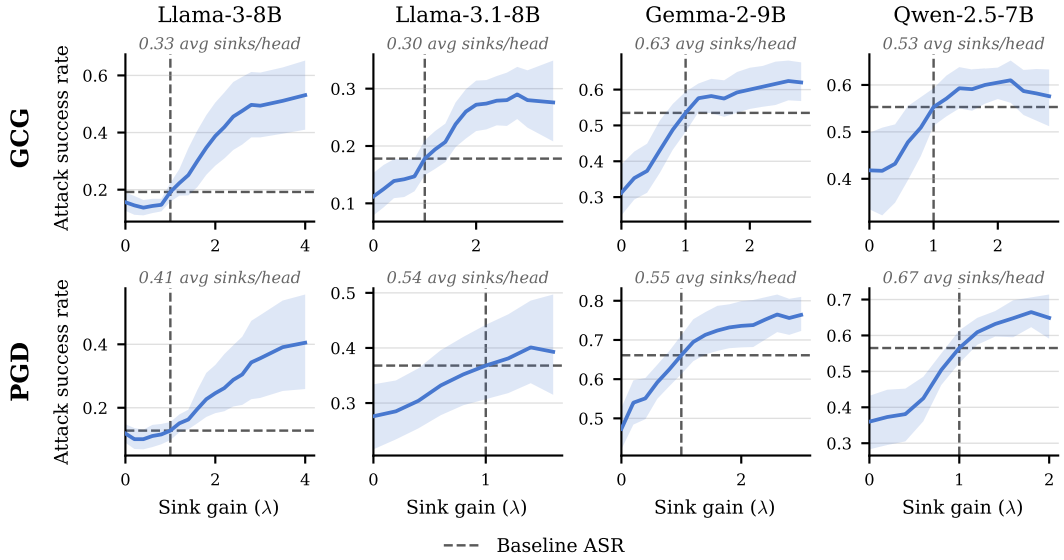

Figure 4: **Sink modulation shifts attack success across both attack methods.** Each panel plots attack success rate (ASR) against the sink gain λ. Dashed lines mark the no-intervention baseline ($\lambda = 1$). Panel subtitles report the average number of suffix sinks per head modulated by the intervention.

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.

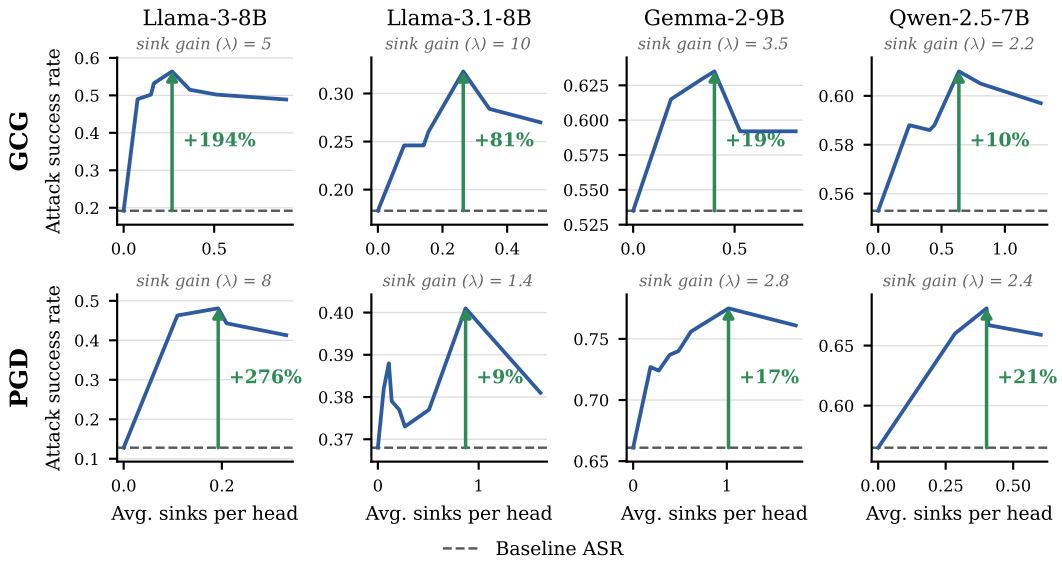

Figure 5: **Amplifying suffix sinks boosts attack success across all four models.** For each model, we fix the sink gain λ at its ASR-maximizing value (panel subtitle) and plot ASR against the average number of sinks per head, varied by sweeping the sink threshold ϵ. Dashed lines mark the no-intervention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

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 0.3–0.7 suffix sinks per head on average, yet modulating just these positions produces visible changes in attack success across diverse model families.

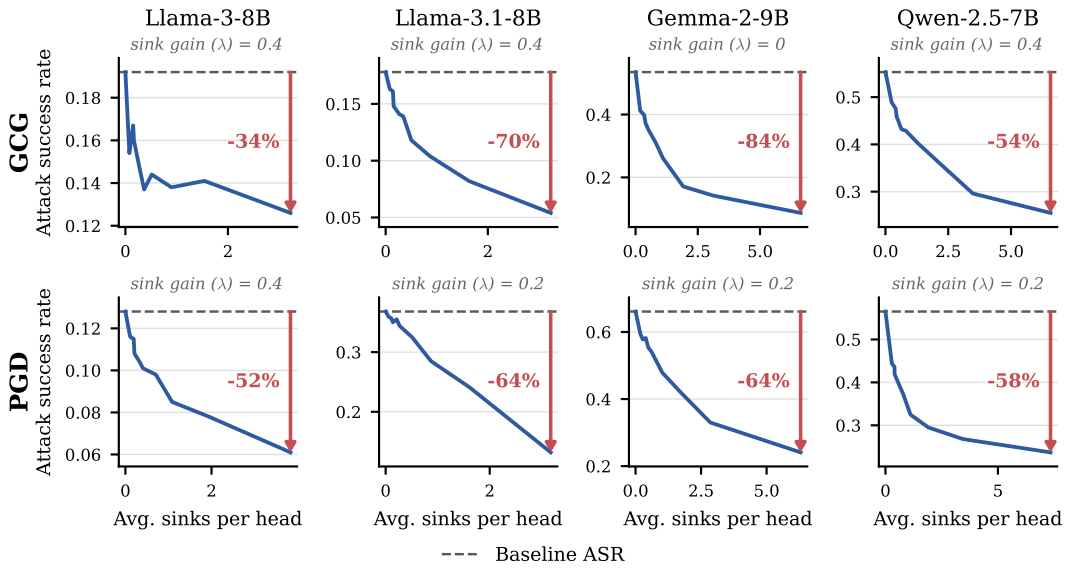

Figure 6: **Attenuating suffix sinks reduces attack success across all four models.** For each model, we fix the sink gain λ at its ASR-minimizing value (panel subtitle) and plot ASR against the average number of sinks per head, varied by sweeping the sink threshold ϵ. Dashed lines mark the no-intervention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

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 C), 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. Notably, amplification yields the largest relative gains on the models with the lowest baseline ASR.

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.

5 Attention Sinks as Carriers of 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 larger positive projection $\langle \mathbf{x}, \mathbf{r} \rangle$ of the model's activations $\mathbf{x}$ onto $\mathbf{r}$ corresponds to a higher probability of refusal on a harmful prompt[1].

5.1 Amplified Sink-Subspace is Aligned with Refusal Direction

Section 4.3 shows that Llama-3-8B and Llama-3.1-8B exhibit the largest ASR gain from sink-subspace amplification. We hypothesize that their sink-induced perturbations are strongly aligned

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

with the refusal direction, such that amplification directly suppresses refusal. To test this, we measure the alignment between each sink perturbation and the refusal direction across layers. Specifically, for a sink at position s, the perturbation contributed by attention head h at layer ℓ is:

$$\boldsymbol{\delta}_s^{(\ell,h)} = \mathbf{W}_O^{(\ell,h)} \mathbf{v}_s^{(\ell,h)} \in \mathbb{R}^{d_{\text{model}}}.$$

Denoting by $\mathcal{S}^{\ell,h}$ the set of sink positions in the suffix at layer ℓ, head h, we measure per-head alignment as the mean absolute cosine similarity between $\boldsymbol{\delta}_s^{(\ell,h)}$ and $\mathbf{r}$, computed per sink token and averaged over all sink positions:

$$\rho^{(\ell,h)} = \frac{1}{|\mathcal{S}^{\ell,h}|} \sum_{s \in \mathcal{S}^{\ell,h}} \frac{|\langle \boldsymbol{\delta}_s^{(\ell,h)}, \mathbf{r} \rangle|}{\|\boldsymbol{\delta}_s^{(\ell,h)}\|_2 \cdot \|\mathbf{r}\|_2}.$$

Since jailbreak mechanisms are known to be head-specific [Zhou et al., 2024], we report, at each layer ℓ, the mean $\rho^{(\ell,h)}$ over the top-3 heads ranked by $\rho^{(\ell,h)}$.

As a baseline, we use $\frac{1}{\sqrt{d_{\text{model}}}}$, the expected cosine similarity between two random unit vectors in $\mathbb{R}^{d_{\text{model}}}$, which a non-informative perturbation would not be expected to exceed. Figure 7 shows that sink token perturbations substantially exceed this baseline across the Llama-3 models.

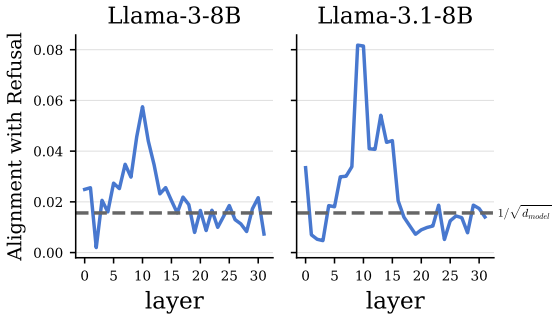

Figure 7: **Cosine similarity of sink token perturbations with the refusal direction (r).** Each panel corresponds to a different model and depicts the mean $\rho^{(\ell,h)}$ over the top-3 attention heads.

5.2 Sink Tokens Perturb Activations Away from Refusal Direction

Section 4.3 further shows that models such as Gemma-2-9B exhibit the sharpest ASR decrease when sink tokens are attenuated. We hypothesize that their sink-induced perturbations cumulatively push the residual stream away from the refusal direction, collectively suppressing refusal across layers. To test this, we measure the cumulative effect of all sink tokens on the residual stream's alignment with the refusal direction. The aggregate sink perturbation to the residual stream of response token t at layer ℓ is:

$$\boldsymbol{\Delta}_t^{(\ell)} = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}^{(\ell,h)}} \mathbf{p}_{t,j}^{(\ell,h)} \boldsymbol{\delta}_j^{(\ell,h)}, \quad (9)$$

where $\mathbf{p}_{t,j}^{(\ell,h)}$ is the attention weight from response token t to sink position j in head h at layer ℓ. At each layer ℓ, we subtract $\boldsymbol{\Delta}_t^{(\ell)}$ from the attention output and measure the change in refusal alignment averaged over all response tokens and evaluation prompts.

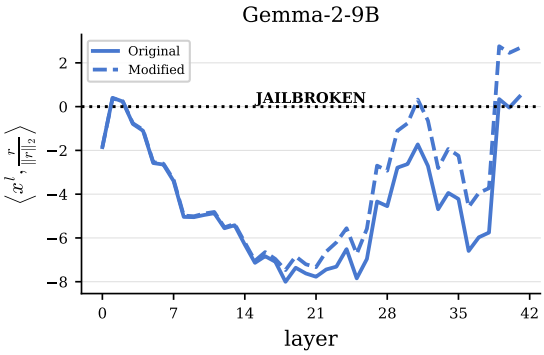

Figure 8: **Sink Tokens Suppress Refusal.** Projection of the original residual stream onto $\mathbf{r}$ (solid) versus the sink-ablated residual stream with $\boldsymbol{\Delta}_t^{(\ell)}$ removed (dashed). A positive gap between the two curves indicates that sink tokens reduce alignment with $\mathbf{r}$.

Figure 8 shows that the change is consistently positive across later layers, indicating that sink tokens collectively reduce residual-stream refusal alignment at every layer. The effect grows with depth, suggesting that sink-induced perturbations accumulate across layers rather than acting at a single layer in the network. Notably, in the final layers, the magnitude of this suppression is substantial: without sink perturbations, the residual stream remains meaningfully aligned with the refusal direction, whereas the original activations are nearly orthogonal to it.

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 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 276%, while attenuation reduces it by up to 84% (Section 4). Mechanistically, sink-induced perturbations align with the refusal direction and cumulatively 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.

Broader Impacts 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. We therefore expect the net effect of this mechanistic understanding to be a positive contribution to model safety.

References

Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking. *Advances in Neural Information Processing Systems*, 37:129696–129742, 2024.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token?, 2025. URL `https://arxiv.org/abs/2504.02732`.

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

Tim Beyer, Sophie Xhonneux, Simon Geisler, Gauthier Gidel, Leo Schwinn, and Stephan Günnemann. Llm-safety evaluations lack robustness. *arXiv preprint arXiv:2503.02574*, 2025.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 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.

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

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE, 2025.

Simon Geisler, Tom Wollschläger, Mohamed Hesham Ibrahim Abdalla, Johannes Gasteiger, and Stephan Günnemann. Attacking large language models with projected gradient descent. *arXiv preprint arXiv:2402.09154*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 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`.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 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/`.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I. Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms, 2024. URL `https://arxiv.org/abs/2410.13835`.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Attention slipping: A mechanistic understanding of jailbreak attacks and defenses in llms. *arXiv preprint arXiv:2507.04365*, 2025.

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

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

Junjie Mu, Zonghao Ying, Zhekui Fan, Zonglei Jing, Yaoyuan Zhang, Zhengmin Yu, Wenxin Zhang, Quanchen Zou, and Xiangzheng Zhang. Mask-gcg: Are all tokens in adversarial suffixes necessary for jailbreak attacks? *arXiv preprint arXiv:2509.06350*, 2025.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b and gpt-oss-20b model card, 2025. URL `https://arxiv.org/abs/2508.10925`.

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

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Yu Haining, and Xiaohua Jia. 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`.

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

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*, 2025.

Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. *arXiv preprint arXiv:2510.06477*, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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/`.

Bingqi Shang, Yiwei Chen, Yihua Zhang, Bingquan Shen, and Sijia Liu. Forgetting to forget: Attention sink as a gateway for backdooring llm unlearning. *arXiv preprint arXiv:2510.17021*, 2025.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2025. URL https://openreview.net/forum?id=2XBPdPIcFK.

Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. Concealed data poisoning attacks on NLP models. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 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/.

Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Zijun Wang, Haoqin Tu, Jieru Mei, Bingchen Zhao, Yisen Wang, and Cihang Xie. Attngcg: Enhancing jailbreaking attacks on llms with attention manipulation. *arXiv preprint arXiv:2410.09040*, 2024.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 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.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. 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.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024.

Itay Yona, Ilia Shumailov, Jamie Hayes, Federico Barbero, and Yossi Gandelsman. Interpreting the repeated token phenomenon in large language models. *arXiv preprint arXiv:2503.08908*, 2025.

Jiahao Yu, Haozheng Luo, Jerry Yao-Chieh Hu, Yan Chen, Wenbo Guo, Han Liu, and Xinyu Xing. Mind the inconspicuous: Revealing the hidden weakness in aligned {LLMs}'refusal boundaries. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 259–278, 2025.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*, 2024a.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan (Celine) Lin. 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.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. 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.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. On prompt-driven safeguarding for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. *arXiv preprint arXiv:2410.13708*, 2024.

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

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2025. URL `https://arxiv.org/abs/2310.01405`.

Zayd M. K. Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. 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 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.

C 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.

D Threshold Values for Section 5

Table 3 reports the per-model sink-detection thresholds used for the experiments in Section 5.

Table 3: Per-model sink-detection thresholds.

Model	**Sink threshold** ϵ
LLAMA-3-8B	0.0100
LLAMA-3.1-8B	0.0100
GEMMA-2-9B	0.0125

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.