ATTENTION SINKS BEYOND THE FIRST POSITION: HETEROGENEOUS PATHWAYS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention sinks, in which a small subset of tokens in a sentence receives disproportionately large attention mass, are commonly associated with the first position. However, tokens at non-initial positions can also become attention sinks, yet the circuit mechanisms underlying these *secondary sinks* remain poorly understood. In this work, we investigate their formation in the first Transformer block using complementary activation-patching experiments. Across the evaluated models and token types, secondary sinks arise through different contributions from word embeddings and self-attention outputs. When the "bos" token appears at a non-initial position, its sink behavior is driven predominantly by its embedding in some models and by a combination of its embedding and attention output in others. For the other tested secondary sink tokens, their sink behavior relies mainly on the strong self-sinking within a small subset of critical heads in the first-block attention output. Further experiments reveal a corresponding difference in sensitivity to contextual mixing: embedding-dominated sinks persist under contextual mixing, even in the presence of competing secondary sink tokens, whereas attention-output-dominated sinks can disappear when competing tokens occur earlier in the sequence. These findings suggest multiple first-block pathways to secondary sink formation and show how these pathways shape the sink behavior under different contexts. More broadly, our analysis provides a circuit-level framework for understanding and distinguishing attention-sink mechanisms in large language models.

1 INTRODUCTION

The attention mechanism is a central component of Transformer-based large language models (LLMs) Vaswani et al. (2017), enabling models to flexibly aggregate contextual information across long input sequences. However, as model size and training data continue to scale, Transformer blocks tend to assign disproportionately large attention weights to a small subset of tokens, which are commonly referred to as **attention sinks** Xiao et al. (2024). Attention sinks have been observed consistently across training regimes, model families, and datasets Gu et al. (2025); Kang et al. (2025); Wang et al. (2025), and can substantially affect the performance of large language models on downstream tasks Xiao et al. (2024); Qiu et al. (2026); Son et al. (2024). While early studies primarily associated attention sinks with the first token position, subsequent evidence Yu et al. (2024); Sun et al. (2024) shows that sink behavior is not exclusive to the first position: tokens at non-initial positions can also attract disproportionately large amounts of attention depending on the sequence context.

A growing body of work Barbero et al. (2025); Li et al. (2026); Peng et al. (2026); Ran-Milo et al. (2026); Sun et al. (2026) has investigated how and why attention sinks arise. However, existing studies primarily focus on sink tokens at the initial position, which we refer to as **primary sinks** (in contrast to **secondary sinks**, which occur at non-initial positions). Their explanations are largely based on properties unique to the first position. For example, Li et al. (2026) attributes primary sinks to the unusually high variance of the first-position representation; Peng et al. (2026) traces them to early-layer mechanisms that differentiate the first-position embedding from later embeddings; Sun et al. (2026) relates them to the alignment between an embedding-independent transformation of the first-token representation and a trigger direction; and Ran-Milo et al. (2026) identifies first-position positional encoding as the key driver. While these studies provide valuable insights into the formation of primary sinks, their explanations rely on first-position asymmetry and Transformer blocks

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

beyond the first layer, and therefore do not fully explain the circuit-level emergence of secondary sinks. Moreover, Zhang et al. (2024); Binkowski et al. (2026) show that secondary sinks may help hallucination detection in LLMs. This naturally raises the following question:

Can we identify the common Transformer components that give rise to secondary attention sinks?

In this work, we address this question by tracing the circuits underlying secondary sinks in the first Transformer block through systematic experiments. We focus on the first block because secondary sink formation is strongly influenced by token embeddings, making this block the earliest stage at which sink and non-sink tokens can be distinguished. Our **experiments** and **contributions** are summarized below.

- We identify secondary sink tokens across a diverse set of LLM families and perform complementary activation-patching experiments to disentangle the roles of token embeddings ($\mathbf{X}$-path) and first-block attention outputs ($\mathbf{V}$-path) in sink formation. Our analysis reveals that secondary sinks arise through different mechanisms depending on the token and model: the BOS token at a non-initial position is predominantly driven by the $\mathbf{X}$-path in most evaluated models, whereas other secondary sink tokens rely primarily on the $\mathbf{V}$-path.
- We investigate the causal mechanisms underlying $\mathbf{V}$-path sink formation through intervention experiments. We first show that, in the first Transformer block, replacing the attention output of a secondary sink token with the output from a single attention head can preserve sink behavior, revealing a set of influential heads that we term *critical heads*. Further analysis demonstrates that $\mathbf{V}$-path sinks are induced by strong self-sinking behavior within a small subset of these critical heads in the first Transformer block.
- We characterize the robustness of the two sink pathways under contextual mixing with competing sink tokens. Our experiments show that $\mathbf{X}$-path sinks remain stable across different surrounding token contexts, whereas $\mathbf{V}$-path sinks are sensitive to interference from competing sink tokens. When competing with other secondary sink tokens, self-sinking behavior is disrupted and $\mathbf{V}$-path sink formation is suppressed, while restoring self-sinking recovers the $\mathbf{V}$-path sink.

1.1 Related Works

Attention sinks, first identified by Xiao et al. (2024), have been widely observed in large language models (LLMs) and Vision Transformers (ViTs) Clark et al. (2019); Kobayashi et al. (2020); Darcet et al. (2024); Kang et al. (2025). Their functional importance has since been demonstrated across a wide range of applications and research areas Yang et al. (2025); Agarwal et al. (2025); Xu et al. (2026). From a spectral perspective, Cancedda (2024) showed that specific subspaces are responsible for the emergence of attention sinks. Guo et al. (2024) provided both empirical and theoretical analyses of attention sinks in small Transformer language models on the Bigram-Backcopy task. Gu et al. (2025) systematically investigated how optimization, data distribution, loss functions, and model architecture influence the emergence of attention sinks. Barbero et al. (2025) argued that attention sinks enable LLMs to avoid over-mixing, linking this phenomenon to the mathematical analysis of information propagation in Transformers. Wang et al. (2025) identified secondary sinks, which emerge primarily in middle layers, persist across multiple layers, and attract a smaller but still significant fraction of the attention mass. More recently, Li et al. (2026) identified a

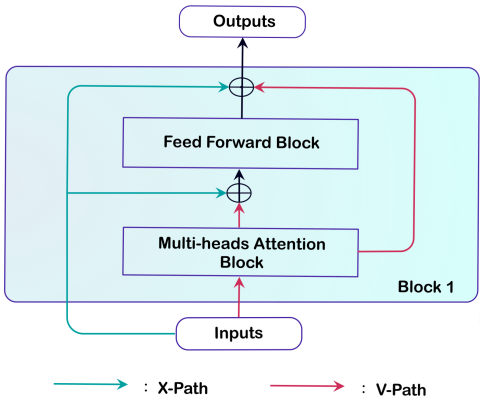


Figure 1: **Information flow in the first Transformer block.** The $\mathbf{X}$-path is indicated by the green and black arrows, while the $\mathbf{V}$-path is indicated by the red and black arrows. The symbol $\oplus$ denotes residual addition via the skip connection.

variance discrepancy inherent in the causal value aggregation process as the root cause of attention sink formation. Sun et al. (2026) provided a mechanistic analysis of the co-emergence of massive activations and attention sinks, showing that their frequent co-occurrence is primarily an architectural consequence of pre-norm Transformers rather than an inherent functional dependency. Wu & Summa (2026) conducted a causal analysis of attention sinks in diffusion transformers, showing that suppressing dynamically identified attention sinks has little effect on semantic alignment under standard inference settings. Peng et al. (2026) investigated the emergence of position-zero attention sinks from an interpretability perspective, proposing a two-block transformer circuit that explains their formation and evolution during pretraining.

2 PRELIMINARIES

Transformer Architecture. We briefly introduce the Llama-style decoder-only Transformer architecture used throughout our analysis. Let $\mathcal{V}$ denote the vocabulary of a finite number of tokens. Given an input text, the tokenizer first maps it into a sequence of token indices $\mathbf{s} = (s_1, \ldots, s_T)$, where $s_t \in \{1, \cdots, |\mathcal{V}|\}$. $\mathbf{s}$ is then embedded as $\mathbf{X}^{(0)}$ by a word embedding matrix $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$, where d is the hidden dimension and

$$\mathbf{X}^{(0)} = \begin{bmatrix} \mathbf{x}_1^{(0)} & \cdots & \mathbf{x}_T^{(0)} \end{bmatrix} \in \mathbb{R}^{T \times d}, \quad \mathbf{x}_t^{(0)} = \mathbf{E}_{s_t,:}.$$

A Llama-style model consists of a stack of L Transformer blocks. Each block contains a multi-head self-attention (**MHSA**) module with H heads followed by a feed-forward network (**FFN**) with SwiGLU activation function Shazeer (2020). Both submodules are equipped with residual connections and pre-normalization. Let $\mathbf{X}^{(l-1)} \in \mathbb{R}^{T \times d}$ denote the input to the l-th Transformer block. The output of the block can be decomposed into the following components:

$$\mathbf{X}^{(l)} = \tilde{\mathbf{X}}^{(l)} + \mathbf{FFN}^{(l)} \left(\mathrm{RMSNorm} \left(\tilde{\mathbf{X}}^{(l)} \right) \right), \tag{1}$$

where

$$\tilde{\mathbf{X}}^{(l)} = \mathbf{X}^{(l-1)} + \mathbf{MHSA}^{(l)} \left(\mathrm{RMSNorm} \left(\mathbf{X}^{(l-1)} \right) \right). \tag{2}$$

MHSA Module. In this module, its normalized input $\mathbf{H}^{(l)} = \mathrm{RMSNorm}\left(\mathbf{X}^{(l-1)}\right)$ is first mapped to the keys, queries and values by the following transformation:

$$\mathbf{K}^{h,l} = \mathbf{H}^{(l)} \mathbf{W}_K^{h,l}, \mathbf{Q}^{h,l} = \mathbf{H}^{(l)} \mathbf{W}_Q^{h,l}, \mathbf{V}^{h,l} = \mathbf{H}^{(l)} \mathbf{W}_V^{h,l},$$

for each head $1 \leq h \leq H$, with $\mathbf{W}_K^{h,l}, \mathbf{W}_Q^{h,l}, \mathbf{W}_V^{h,l} \in \mathbb{R}^{d \times d_h}$ and $d_h = d/H$. The attention output is then computed by

$$\mathbf{MHSA}^{(l)} \left(\mathbf{H}^{(l)} \right) = \mathrm{Concat}_{h=1}^{H} \left(\mathbf{A}^{h,l} \mathbf{V}^{h,l} \right) \mathbf{W}_O^l.$$

The attention weights at each head are computed by

$$\mathbf{A}^{h,l} = \mathrm{Softmax} \left(\mathbf{Q}^{h,l} {\mathbf{K}^{h,l}}^{\top} / \sqrt{d_h} + \mathbf{M} \right),$$

where $\mathbf{M}$ is the attention mask. The final hidden representation passes through an RMSNorm layer and a linear projection to produce logits for next-token prediction.

Attention Sink Rate. Throughout this paper, the attention sink rate at position k is computed by the following formula as in Gu et al. (2025):

$$\mathrm{Sink}_t^{\epsilon} = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{H} \sum_{h=1}^{H} \mathbb{I} \left(\alpha_t^{h,l} > \epsilon \right), \tag{3}$$

where $\alpha_t^{h,l} = \frac{1}{T-t+1} \sum_{i=t}^{T} \mathbf{A}_{i,k}^{h,l}$ is the importance score for the t-th position, and $\mathbf{A}_{i,t}^{h,l}$ is the attention weight of position i to the position t at layer l and head h. Unless otherwise specified, the threshold hyperparameter ϵ is set to 0.3 when fixing the sequence length T to 64.

Experimental Setup. We analyze secondary sink behavior in four large language models: Llama2-7B, Llama3-8B, Mistral-7B, and Gemma2-2B. Our analysis focuses on three representative secondary sink tokens: the BOS token, the period ".", and the newline "\n" token. Unless otherwise specified, all reported results are averaged over 512 samples from the WikiText-2 (Merity et al., 2016) test set. For large-scale experiments involving multiple models and tokens, we report representative results in the main paper and provide the complete results in the appendix due to space constraints. Unless explicitly noted, the full results are consistent with those presented in the main text.

3 Two Circuits in the First Block

In this section, we identify the secondary sink tokens for each model in Section 3.1, introduce two paths of information flow, and relate the sink circuit to these pathways in Section 3.2.

3.1 Secondary Sinks in LLMs

In popular LLM families, attention sinks are not restricted to the first position of the input sequence. Instead, various weakly informative or structural tokens, such as the BOS token, the period ".", the newline token "\n", the conjunction "and", and the preposition "of", can also exhibit sink behavior at non-initial positions. In this subsection, we identify and analyze the secondary sink behavior of three representative tokens across various models.

Identify Per-model Secondary Sink Tokens. For a selected representative test token, we estimate the sink rate Sink_t^ϵ by a token-replacement experiment. For example, to evaluate Sink_t^ϵ of the BOS token in Llama3-8B, we sample sequences of length 64 from the source dataset, replace the token at the t-th position with the BOS token, feed the modified sequence into the model, and compute Sink_t^ϵ according to Equation (3). We repeat this procedure over the three representative tokens, models, and t ranging from 1 to 10. The full results are reported in the appendix and summarized in Table 1. The summary shows that secondary sinks appear across all examined model families. In particular, all three representative tokens exhibit secondary sink behavior in Llama2-7B and Mistral-7B, whereas only the BOS token consistently acts as a secondary sink token in Llama3-8B and Gemma2-2B.

Table 1: The secondary sink tokens across models.

Model	BOS	Period	Newline
Llama2-7B	✓	✓	✓
Llama3-8B	✓	✗	✗
Mistral-7B	✓	✓	✓
Gemma2-2B	✓	✗	✗

3.2 $\mathbf{X}$-Path Sink and $\mathbf{V}$-Path Sink

For a token at a non-initial position, information can influence the output of the first Transformer block through two drivers: the token's word embedding and its first-block attention output. To better study the underlying effect of these drivers, we propose two conceptual pathways: $\mathbf{X}$-path and $\mathbf{V}$-path, see Figure 1. These pathways reflect how information from token embeddings and attention outputs flows through the first Transformer block, providing more intuition on how individual components can be key drivers of secondary sink formation. To better present our experiment, we represent the post-attention residual at position t as r_t:

$$r_t = x_t^{(0)} + a_t, \quad x_t^{(1)} = r_t + \mathbf{FFN}^{(1)}\left(\mathrm{RMSNorm}\left(r_t\right)\right) \tag{4}$$

where, for simplicity, a_t is the first-block attention output at position t. Then, we are ready to introduce our controlled experiments.

Remark 1. We focus primarily on the first block because the mechanisms governing secondary sink formation in later layers appear to be largely similar to those already studied for primary sinks. However, the first block is crucial for identifying the patterns that initiate secondary sink behavior and has not yet been studied. We provide further illustration in Section F in the appendix.

Model	Token	R_X^{suff}	R_X^{nec}	R_V^{suff}	R_V^{nec}	Types
Gemma2-2B	BOS	0.928 [0.927, 0.929]	0.973 [0.973, 0.973]	0.000 [0.000, 0.000]	0.015 [0.015, 0.016]	X
Llama3-8B	BOS	0.375 [0.354, 0.396]	0.933 [0.932, 0.934]	0.015 [0.014, 0.016]	0.573 [0.552, 0.593]	Neither
Mistral-7B	BOS	0.856 [0.846, 0.865]	0.813 [0.795, 0.830]	0.022 [0.008, 0.035]	-0.004 [-0.005, -0.004]	X
Mistral-7B	"."	0.000 [0.000, 0.000]	0.001 [0.000, 0.001]	0.875 [0.867, 0.884]	0.999 [0.999, 0.999]	V
Mistral-7B	"\n"	0.026 [0.019, 0.034]	0.005 [0.003, 0.007]	0.879 [0.871, 0.887]	0.950 [0.938, 0.961]	V
Llama2-7B	BOS	0.658 [0.643, 0.673]	0.193 [0.179, 0.209]	0.807 [0.791, 0.821]	0.342 [0.327, 0.357]	Neither
Llama2-7B	"."	0.069 [0.060, 0.079]	0.013 [0.012, 0.014]	0.989 [0.988, 0.990]	0.931 [0.921, 0.940]	V
Llama2-7B	"\n"	0.003 [0.001, 0.004]	0.009 [0.009, 0.010]	0.998 [0.997, 0.999]	0.997 [0.996, 0.999]	V

Table 2: Bidirectional causal replacement test. Entries are normalized effects with paired source-bootstrap 95% confidence intervals. Tokens in a category require both sufficiency and necessity coefficients to reach 0.5 for a pathway; all categories also satisfy the CI-supported classification rule.

Bidirectional Causal Replacement. We test whether the candidate token's secondary attention sink can be produced or removed by swapping one part of the first block's activation. For each model, we use the first 512 eligible length-64 sequences from WikiText-2, without adding a BOS token or a chat template. At each one-based position $t \in \{2, 3, 4, 5\}$, we run two versions of a sequence: one with its original token at t, termed the default sequence, and one with that token replaced by the candidate, termed the candidate sequence. For the default sequence, a secondary sink does not appear, while for the candidate sequence, a secondary sink appears at the target position. At position t, we denote the original token's embedding and attention components by $x_t^{(0)}$ and a_t, and denote the corresponding components for the candidate by $x_t^{(0),c}$ and a_t^c .

For each candidate's component, we ask two questions. **Sufficiency**: If we replace the corresponding component of the original token in the default sequence with the candidate's component, does the secondary sink appear? **Necessity**: If we replace that component in the candidate-token input with the original token's component, does the sink disappear? We make these replacements by passing the following combined activation r_t to the rest of the model:

	Sufficiency: start with default sequence	Necessity: start with candidate sequence
$\mathbf{X}$-path	$r_t = x_t^{(0),c} + a_t$	$r_t = x_t^{(0)} + a_t^c$
$\mathbf{V}$-path	$r_t = x_t^{(0)} + a_t^c$	$r_t = x_t^{(0),c} + a_t$

(5)

In the $\mathbf{X}$-path row, the sufficiency test adds the candidate's embedding component to the default sequence; the necessity test removes that component from the candidate sequence. The $\mathbf{V}$-path row applies the same logic to the attention output component. In both tests, after position t is patched, tokens in later positions attend to the newly patched token, and the model processes the rest through its usual normalization, feed-forward network, residual connection, and later blocks.

These tests provide evidence of sufficiency or necessity. For example, we test whether the embedding of the BOS token is the dominant component for the formation of a secondary sink by performing $\mathbf{X}$-path sufficiency and necessity tests above. During the sufficiency test, if the sink rate at position t rises significantly in the default sequence, it implies that the embedding of the BOS token is sufficient to induce a secondary sink, since its attention output component is not present. In the necessity test, if the sink rate at position t drops substantially in the candidate sequence, it implies that the embedding of the BOS token is also necessary to induce a secondary sink, since its attention output component alone cannot produce a secondary sink.

Evaluation Criteria. Let S_{default} and $S_{\text{candidate}}$ be the mean secondary sink rates for the unpatched default and candidate sequences. For each path $B \in \mathbf{X}, \mathbf{V}$, let S_B^{suff} be the mean sink rate after inserting the candidate's component into the default sequence, and let S_B^{nec} be the mean rate after replacing that component in the candidate sequence with the default token's component. We measure each change relative to the gap between the two unpatched sequences, $\Delta = S_{\text{candidate}} - S_{\text{default}}$, and define the sufficiency score R_B^{suff}, necessity score R_B^{nec} and the joint score J_B as:

$$R_B^{\text{suff}} = \frac{S_B^{\text{suff}} - S_{\text{default}}}{\Delta}, \qquad R_B^{\text{nec}} = \frac{S_{\text{candidate}} - S_B^{\text{nec}}}{\Delta}, \qquad J_B = \min\{R_B^{\text{suff}}, R_B^{\text{nec}}\}. \tag{6}$$

A sufficiency score of 1 means that inserting the candidate's component recovers the full attention sink rate relative to the gap. A necessity score of 1 means that replacing a non-sink component

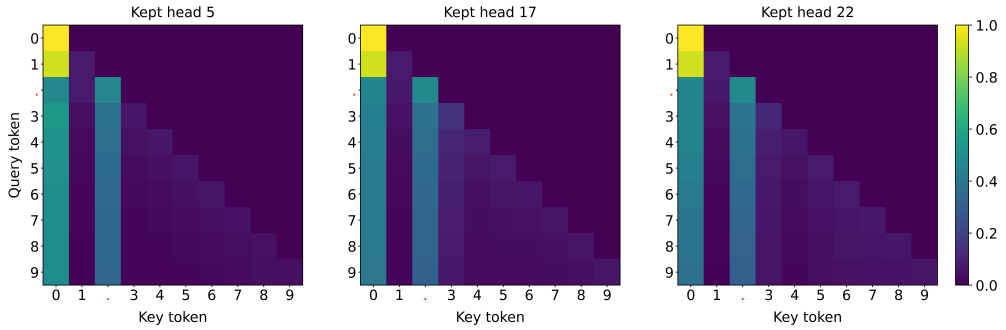


Figure 2: Heatmaps for critical-head identification test. We display the attention pattern of the critical heads of the Llama2-7B model and the secondary sink token "period". We can see that all the identified heads can independently invoke attention sinks of the whole model through the **V**-path sink circuit.

removes the secondary sink phenomenon. The joint score J_B takes the lower of the two, so a path scores highly only when it passes both tests. Under the formulated criteria, we classify a secondary sink token as following an **X**-path sink if $J_X \geq 0.5$ and $J_V < 0.5$, and as following a **V**-path sink if $J_V \geq 0.5$ and $J_X < 0.5$. If both joint scores are greater than 0.5, the classification is Mixed; if neither does, it is classified as Neither.

Experimental Results. Table 2 identifies two **X**-path cases, four **V**-path cases, and two Neither cases. For BOS in Gemma2-2B and Mistral-7B, inserting the candidate's embedding component into the default sequence recovers most of the secondary sink, while replacing it in the candidate sequence removes it ($J_X = 0.928$ and 0.813, respectively). Swapping the attention component has little effect in either model.

For period and newline in Llama2-7B and Mistral-7B, the pattern reverses: swapping the candidate's attention component recovers 87.5–99.8% of the unpatched sink-rate gap in the sufficiency test and removes 93.1–99.9% in the necessity test. The embedding component recovers at most 6.9% and removes at most 1.3%. All four pairs therefore meet the **V**-path criterion.

The two Neither cases fail for different reasons. For BOS in Llama3-8B, inserting the candidate's embedding into the default sequence recovers only 37.5% of the gap and inserting the candidate's attention output into the default sequence recovers only 1.5%. This indicates neither component meets the sufficiency threshold. For BOS in Llama2-7B, replacing either component in the candidate sequence removes only 19.3% or 34.2% of the gap. Thus, neither component meets the necessity threshold. Since the Neither case it more complicated since it involves interaction between token's embedding and attention output, we focus mainly on the analysis of **X**-path sink and **V**-path sink tokens. We can also see from Table 2 that the classification result is robust from the criteria ranging from 0.4 to 0.8, which represent certain robustness for criteria selection. Therefore, we set the threshold to 0.5.

4 The Emergence of V-Path Sinks

Unlike the **X**-path sinks, where sink formation is primarily driven by the word embedding of the secondary sink token itself, the **V**-path sink circuit is potentially affected by the mixing of the value vectors of all preceding tokens in the context. This makes it natural to ask why the **V**-path sinks remain robust under most contextual mixing and positional variation. To address this question, we hypothesize that a subset of heads in the multi-head self-attention module is primarily responsible for the formation of V-

Table 3: The identified critical head indices across models and tokens.

Model	Token	Critical head indices
Llama2-7B	Period	5, 17, 22
Llama2-7B	Newline	1, 20, 23, 27
Mistral-7B	Period	12, 28
Mistral-7B	Newline	9, 12

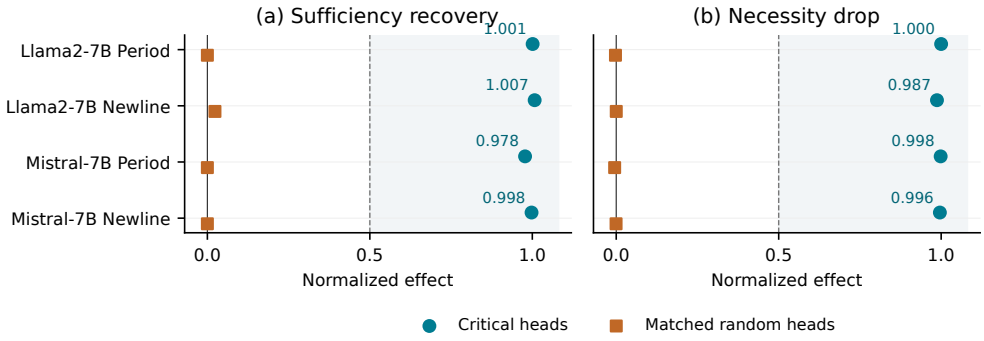


Figure 3: Causal validation of critical heads. Paired replacement of the identified heads nearly recovers and suppresses the V-sink effects, while the tested random sets have effects near zero.

path sinks. We test this hypothesis through the following experiments.

Critical-Head Identification Test. For each model–token pair previously classified as following the $\mathbf{V}$-path sink circuit, we replace the token at the target test position t with the tested secondary sink token for each sample. When forwarding the modified sample to the first Transformer block, we evaluate how each head contributes to forming the secondary sink independently. For a selected head index j, we compute the modified attention outputs at position t by:

$$a_t = \mathrm{Concat}_{h=1}^{H}\left(\mathbb{I}_{\{h=j\}}\mathbf{A}^{h,1}\mathbf{V}^{h,1}\right)\mathbf{W}_O^1, \tag{7}$$

where $\mathbb{I}$ is the index indicator function. This operation retains the weighted value output of the j-th head while setting the values of all other heads to zero. Then, the first-block output $\mathbf{x}_t^{(1)}$ is replaced by the output of the $\mathbf{V}$-path with this modified attention term, while the forward computation in the remaining layers is left unchanged. By plotting the attention pattern of the whole model for each head, we identify heads that can independently preserve secondary sink behavior and term them *critical heads*. The identified critical heads are summarized in Table 3. For intuition, we also visualize the attention pattern of the critical heads for the period token of the Llama2-7B model in Figure 2. From the presented results, we find that, for each $\mathbf{V}$-path sink token, a small subset of heads can invoke secondary sinks by themselves.

To explore the causal evidence underlying how a secondary sink is formed through the $\mathbf{V}$-path, we conduct an experiment to measure the joint effect of the critical heads.

Causal Validation of the Identified Heads. We construct the default and candidate sequences similarly to the bidirectional causal replacement test. The tested component for replacement changes from the token embedding and the whole first-block attention output to the attention outputs of a selected head set $\mathcal{H}$. For all heads $h \in \mathcal{H}$ at target position t, we denote the attention output of the selected heads in the default sequence as $a_{t,h}$, and the candidate token's output as $a^c_{t,h}$. The **Sufficiency** test replaces $a_{t,h}$ with $a^c_{t,h}$ in the default sequence, while the **Necessity** test replaces $a^c_{t,h}$ with $a_{t,h}$ in the candidate sequence. All other head outputs and positions retain their respective values, and the native output projection, residual stream, feedforward computation, and later layers execute normally. We additionally compare each set with one fixed, size-matched random set of heads sampled without replacement from the critical heads' complement, shared across positions, samples, and intervention directions.

We use the notation R_C^{suff} and R_C^{nec} as the sufficiency and necessity scores for critical heads, analogous to the bidirectional causal replacement test, and use R_R^{suff} and R_R^{nec} for the randomly selected heads. Figure 3 shows that the selected Period and Newline heads have 97.8–100.7% of the sufficiency score and 98.7–100.0% of the necessity score. Across all pairs, the fixed random controls show at most 2.3% and 1.9%, respectively. The results therefore localize a strong causal contribution of the small, selected head sets to the formation of the $\mathbf{V}$-path sink.

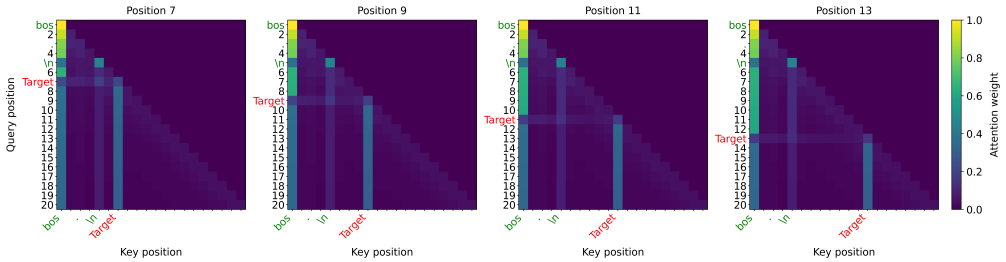


Figure 5: Heatmaps for **X**-path sink robustness test for the Mistral-7B model. We can see that the target non-sink token, marked in red in the plot, acts as a secondary sink token, regardless of token mixing and position.

Critical Heads' Attention Pattern Test. We analyze the attention patterns of the first Transformer block, focusing on the identified critical heads. For each input sequence, we perform the same token replacement at the target position as in the previous experiments and then pass the modified sequence through the original model. We then extract the attention matrices of the selected heads from the first Transformer block. The resulting heatmaps show the mean attention weight from the query position to the key position. This analysis reveals how the critical heads distribute attention weights in the first block when the secondary sink token mixes with its preceding tokens.

A typical result is presented in Figure 4. We find that, in the critical heads of the first Transformer block, the secondary sink token exhibits strong self-sinking behavior: when it mixes with its preceding tokens, these heads assign an unexpectedly large attention weight to the token itself. This observation helps explain why the **V**-path sink can remain effective despite contextual mixing. Although the circuit operates through the attention pathway, the critical heads preserve sink-related information by strongly attending to the secondary sink token itself.

Together, these three tests reveal that the **V**-path sink emerges through the self-sinking behavior of secondary sink tokens at a subset of critical heads in the first Transformer block.

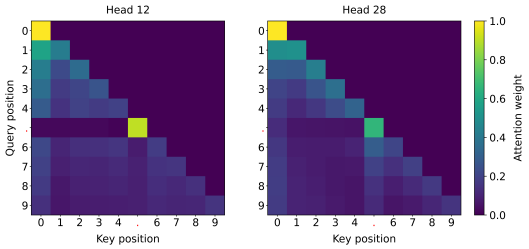


Figure 4: Heatmaps for the critical heads' attention pattern test. The attention pattern of the first block in critical heads exhibits strong self-sinking behavior for the period token "." of the Mistral-7B model.

5 Secondary Sink Behavior Analysis

In this section, we further investigate whether the two sink circuits exhibit different responses to contextual mixing. Specifically, we examine their robustness when competing secondary sink tokens appear in preceding positions.

5.1 **X**-Path Sink is Robust to Contextual Mixing

X-Path Sink Robustness Test. We conduct this experiment on Mistral-7B, and Gemma2-2B, in which the BOS token is classified as an **X**-path sink token. For each sample from the source dataset, we replace tokens at early positions with representative secondary sink tokens, which serve as potential competitors for attention weights at the target position. We then select a target position later in the sequence and replace the token at this position with the candidate sink token. We leave the remaining Transformer computation unchanged and visualize the resulting attention patterns at different target positions.

As shown in Figure 5, an attention sink emerges at the target position. This result indicates that the **X**-path sink circuit is robust to contextual mixing, even when competing secondary sink tokens occur in preceding positions.

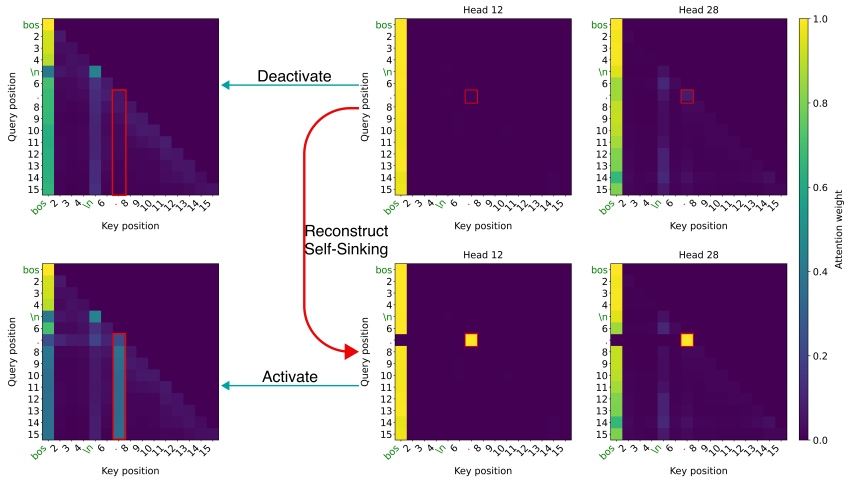


Figure 6: Heatmaps for $\mathbf{V}$-path sink disruption test for Mistral-7B. The period token "." becomes a non-sink token when it mixes with BOS and "\n" in preceding positions. When we reconstruct its self-sinking in critical heads, it becomes a secondary sink token again.

5.2 $\mathbf{V}$-Path Sink Is Sensitive to Mixing with Competing Sink Tokens

$\mathbf{V}$-Path Sink Can be Disrupted. We next test whether secondary sink tokens following the $\mathbf{V}$-path sink circuit lose their sink behavior when they mix with competing secondary sink tokens. We place competing secondary sink tokens at early positions in the input sequence and replace the token at the target position with the tested secondary sink token. We then pass the modified sequence through the original Transformer and visualize its attention pattern.

As shown in the top-left plot in Figure 6, when the BOS and newline tokens "\n" appear in preceding positions, the period token "." no longer forms a secondary sink. We additionally examine the first-block attention patterns of the corresponding critical heads and observe that the self-sinking behavior of the period token "." disappears in the top-right plot. This result suggests that competing secondary sink tokens can disrupt self-sinking in the critical heads and thereby suppress the $\mathbf{V}$-path sink circuit. To test whether this disruption causes the loss of sink behavior, we manually reconstruct self-sinking for the period token "." in the critical heads through weight reallocation. As shown in the bottom plots in Figure 6, restoring this self-sinking pattern causes the period token "." to recover its secondary sink behavior.

Together, these experiments demonstrate that the $\mathbf{V}$-path sink circuit is highly sensitive to contextual mixing with competing secondary sink tokens. This sensitivity stems from its reliance on self-sinking in the critical heads: once contextual mixing disrupts this pattern, the $\mathbf{V}$-path can no longer produce a first-block representation that supports secondary sink formation.

6 Conclusion

In this work, we investigate the circuit mechanisms underlying secondary attention sinks. We identify three sink patterns: $\mathbf{X}$-path, $\mathbf{V}$-path and Neither. The $\mathbf{X}$-path sink pattern is driven primarily by the token embeddings, and the $\mathbf{V}$-path sink pattern relies on the multi-head attention output and self-sinking in a subset of critical heads. Our experiments further show that these sink patterns exhibit different sensitivities to contextual mixing: $\mathbf{X}$-path sink remains stable under the tested contexts, whereas $\mathbf{V}$-path sink can be disrupted and subsequently restored by manipulating self-sinking in the critical heads. These findings extend the study of attention sinks beyond first-position asymmetry and show that secondary sinks are not governed by a single mechanism. More broadly, our results provide a circuit-level framework for analyzing token-specific attention sinks, with potential implications for mechanistic interpretability, attention-sink mitigation, and the design of more reliable attention-based analyses.

AI USE STATEMENT

We primarily used generative AI tools for language polishing, proofreading, and code assistance. The core ideas, methodology, and contributions of this work were developed by the authors. All AI-assisted content has been carefully reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide experimental details across models in the Appendix, and all the presented results and statements are consistent with the results presented in the main text. The code and data used in this work will be released upon acceptance of the paper. All experiments were conducted using publicly available datasets and models, and we have included sufficient information for others to replicate our findings.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.

Jakub Binkowski, Kamil Adamczewski, and Tomasz Kajdanowicz. Attention sinks as internal signals for hallucination detection in large language models. *arXiv preprint arXiv:2604.10697*, 2026.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4792–4808, 2024.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, pp. 276–286, 2019.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 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 *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 2025.

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. *arXiv preprint arXiv:2410.13835*, 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*, 2025.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, 2020.

Siquan Li, Kaiqi Jiang, Jiacheng Sun, and Tianyang Hu. The structural origin of attention sink: Variance discrepancy, super neurons, and dimension disparity. *arXiv preprint arXiv:2605.06611*, 2026.

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

Jaehoon Oh, Seungjun Shin, and Dokwan Oh. House of cards: Massive weights in llms. *arXiv preprint arXiv:2410.01866*, 2024.

Runyu Peng, Ruixiao Li, Mingshu Chen, Yunhua Zhou, Qipeng Guo, and Xipeng Qiu. How attention sinks emerge in large language models: An interpretability perspective. *arXiv preprint arXiv:2603.06591*, 2026.

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. *Advances in Neural Information Processing Systems*, 38:100092–100118, 2026.

Yuval Ran-Milo, Hila Ofek, and Shahar Mendel. A mechanistic account of attention sinks in gpt-2: One circuit, broader implications for mitigation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 90–98, 2026.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeun Kim, and Jaeho Lee. Prefixing attention sinks can mitigate activation outliers for large language model quantization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2242–2252, 2024.

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

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks. *arXiv preprint arXiv:2603.05498*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, Hui Xue, Jialing Tao, Ranjie Duan, and Jiexi Liu. Mirage in the eyes: Hallucination attack on multi-modal large language models with only attention sink. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3707–3726, 2025.

Fangzheng Wu and Brian Summa. Attention sinks in diffusion transformers: A causal analysis. *arXiv preprint arXiv:2605.09313*, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*, 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*, 2024.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *arXiv preprint arXiv:2411.09968*, 2024.

A Appendix

A.1 Identify Secondary Sink Tokens across Models

Table 4: Attention Sink Rate for Representative Token Replaced at Target Position

Model	Token	Avg sink	pos 0	pos 1	pos 2	pos 3	pos 4	pos 5	pos 6	pos 7	pos 8	pos 9
llama2_7b	**BOS**	0.873	0.961	0.894	0.876	0.871	0.867	0.861	0.856	0.852	0.849	0.846
llama2_7b	**Period**	0.855	0.956	0.870	0.858	0.850	0.845	0.839	0.835	0.834	0.831	0.830
llama2_7b	**Newline**	0.849	0.957	0.868	0.854	0.844	0.838	0.832	0.828	0.824	0.822	0.822
llama3_8b	**BOS**	0.725	0.997	0.951	0.894	0.830	0.755	0.680	0.617	0.558	0.508	0.463
llama3_8b	Period	0.129	0.987	0.048	0.043	0.037	0.035	0.031	0.028	0.027	0.026	0.026
llama3_8b	Newline	0.106	0.958	0.018	0.017	0.013	0.010	0.009	0.009	0.007	0.008	0.008
mistral_7b	**BOS**	0.927	0.994	0.978	0.968	0.949	0.936	0.924	0.909	0.894	0.872	0.849
mistral_7b	**Period**	0.709	0.968	0.877	0.799	0.726	0.680	0.661	0.643	0.617	0.568	0.546
mistral_7b	**Newline**	0.720	0.967	0.893	0.807	0.731	0.689	0.675	0.652	0.628	0.589	0.565
phi3_mini_4k	BOS	0.100	0.906	0.008	0.008	0.009	0.009	0.010	0.010	0.011	0.012	0.014
phi3_mini_4k	Period	0.140	0.958	0.043	0.048	0.047	0.048	0.050	0.051	0.051	0.052	0.053
phi3_mini_4k	Newline	0.131	0.952	0.035	0.042	0.042	0.041	0.042	0.042	0.039	0.039	0.039
gemma2_2b	**BOS**	0.903	0.928	0.907	0.901	0.898	0.897	0.897	0.897	0.898	0.901	0.902
gemma2_2b	Period	0.026	0.003	0.002	0.007	0.010	0.015	0.020	0.027	0.041	0.056	0.076
gemma2_2b	Newline	0.048	0.017	0.013	0.019	0.023	0.033	0.042	0.053	0.069	0.092	0.123

In Table 4, we report the complete results of secondary sink identification for the representative tokens across all evaluated models, including Phi3-mini-4K. For Phi3-mini-4K, none of the representative tokens exhibits secondary sink behavior; they act as sinks only at the first position. In contrast, all other models contain at least one representative token that forms a secondary sink across multiple non-initial positions. We further observe that the sink rates of the BOS token in Llama3-8B and the period "." and newline "\n" tokens in Mistral-7B decrease substantially as the target position moves later in the sequence, although they remain non-negligible at the final tested position. Accordingly, our analysis focuses primarily on secondary sinks at early positions rather than tracing their circuits over substantially longer sequences. This setting is also practically relevant because structural tokens such as "." and "\n" naturally occur as sentences or text lines terminate.

A.2 Supplementary Results for Critical-Head Identification Test

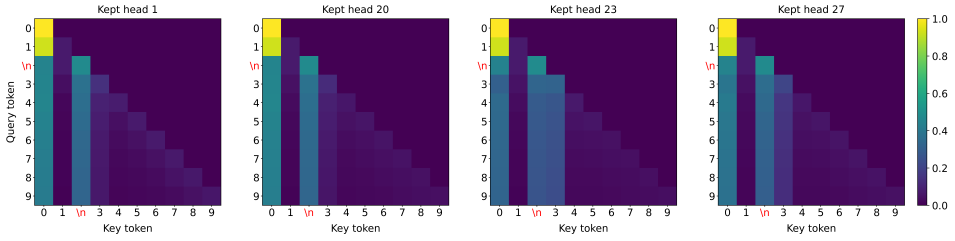


Figure 7: The attention pattern of the whole model for the critical-head identification test of the Llama2-7B model and secondary sink token newline "\n".

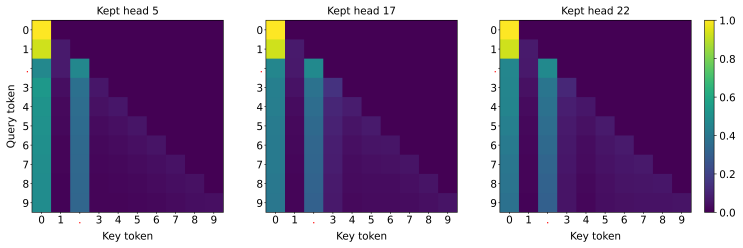


Figure 8: The attention pattern of the whole model for the critical-head identification test of the Llama2-7B model and secondary sink token period ”.”.

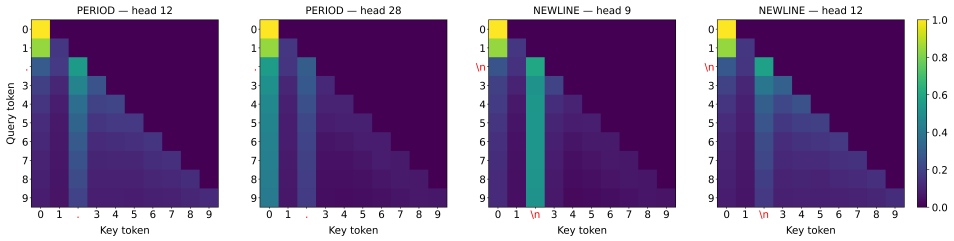


Figure 9: The attention pattern of the whole model for the critical-head identification test of the Mistral-7B model and secondary sink token period ”.” and newline ”\n”.

As shown in the figures in this Section, all identified secondary sink tokens have critical heads in the first Transformer block. However, both the number and indices of these heads vary across models and token types. We identify critical heads only at target position 3, since a single position is sufficient for this purpose: removing the heads identified at position 3 causes the attention sink rate to decrease substantially across the first ten target positions in the subsequent critical-head removal experiment.

A.3 Supplementary Results for Critical Heads' Attention Pattern Test

Across the figures presented in this section, the identified critical heads exhibit a consistent attention pattern despite variations across models and token types. At the target position, when a secondary sink token mixes with all preceding tokens, these heads assign an unexpectedly large attention weight to the token itself. We refer to this behavior as *self-sinking* in the critical heads.

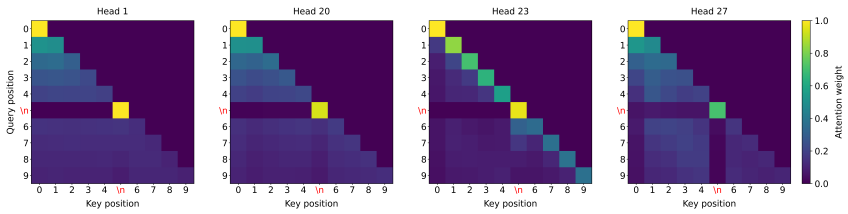


Figure 10: The attention pattern for the critical heads in the first Transformer block of the Llama2-7B model and secondary sink token newline ”\n”.

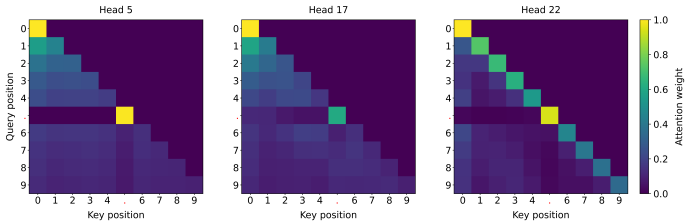


Figure 11: The attention pattern for the critical heads in the first Transformer block of the Llama2-7B model and secondary sink token period ”.”.

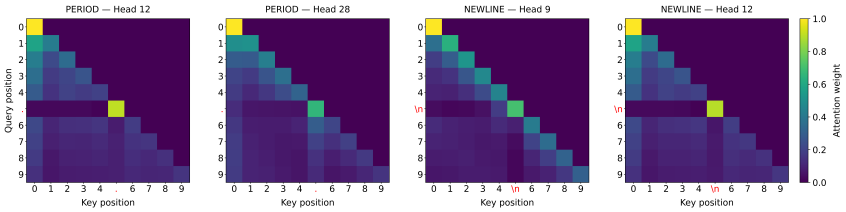


Figure 12: The attention pattern for the critical heads in the first Transformer block of the Mistral-7B model and secondary sink token period ”.” and newline ”\n”.

A.4 Supplementary Results for X-Path Sink Robustness Test

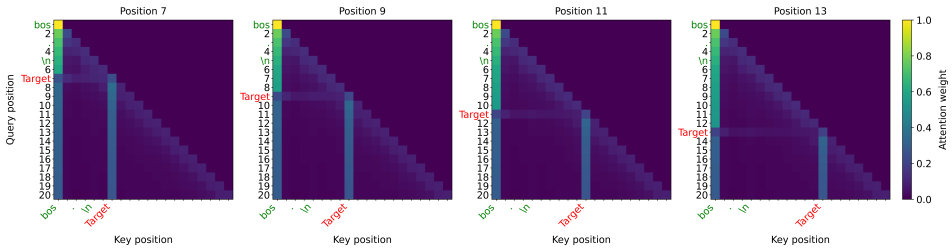


Figure 13: The attention pattern of the whole model for the **X**-path sink activation test of the Gemma2-2B model.

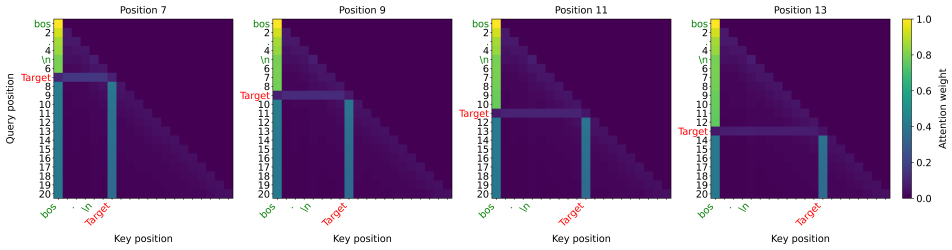


Figure 14: The attention pattern of the whole model for the **X**-path sink activation test of the Llama3-8B model.

As shown in the figures in this section, a non-sink token can activate the **X**-path sink circuit when its input embedding along the **X**-path is replaced with that of an **X**-path sink token, such as the BOS token in the evaluated models. After this intervention, the token becomes a secondary sink across all tested models and remains stable over the evaluated target positions, despite variations in its attention output caused by mixing with different surrounding tokens. These results suggest that the **X**-path sink circuit is robust to contextual mixing, in sharp contrast to the context-sensitive **V**-path sink circuit.

A.5 SUPPLEMENTARY RESULTS FOR V-PATH SINK DISRUPTION TEST

As shown in the figures in this section, nearly all evaluated model–token pairs exhibit a consistent pattern. When a token associated with the **V**-path sink circuit mixes with competing sink tokens, it may lose its secondary sink behavior, accompanied by the disruption of self-sinking in the critical heads. Reconstructing this self-sinking pattern restores the token's secondary sink behavior. The only exception is the BOS token in Llama2-7B, as shown in Figure **??**, whose sink behavior remains robust to competing tokens when averaged over a large number of samples. Nevertheless, a per-sample analysis provides additional support for our finding. As shown in Figure **??**, we examine the sentence "Summer is hot. ¡bos¿ Winter is cold." In this example, the BOS token does not form a secondary sink, and its self-sinking behavior is disrupted. After reconstructing self-sinking in the critical heads, the BOS token becomes a secondary sink again.

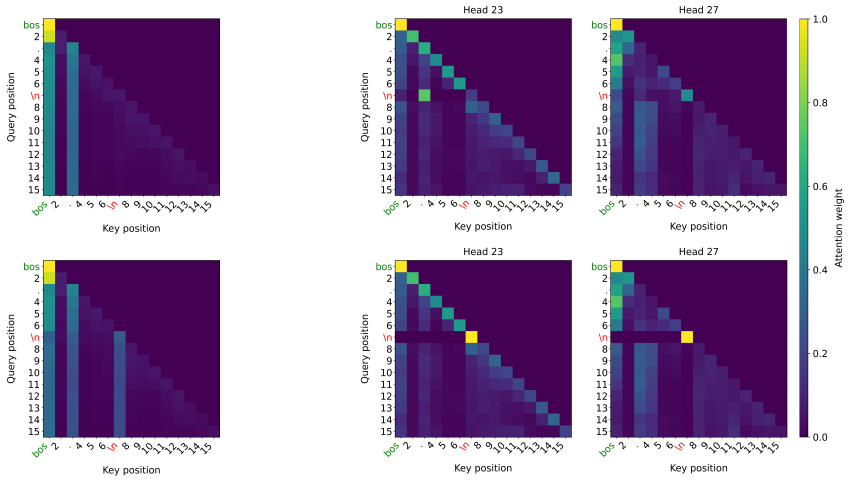


Figure 15: The results for the **V**-path sink disruption test of the newline "\n" token in Llama2-7B.

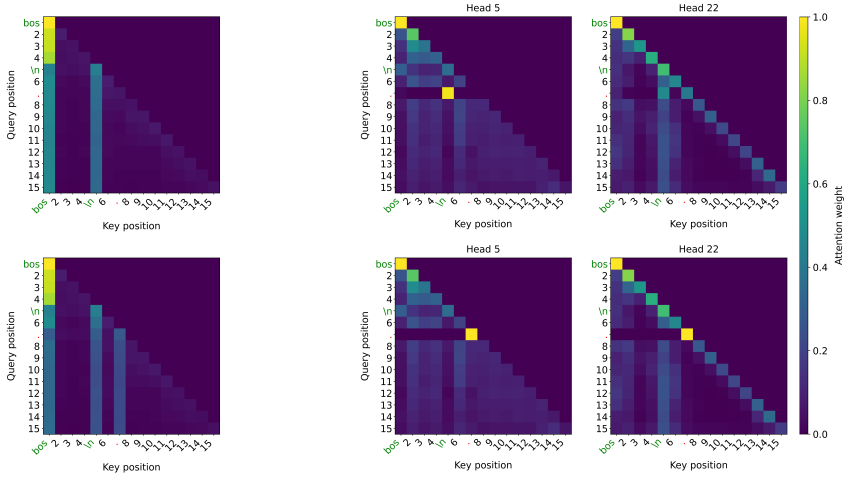


Figure 16: The results for the **V**-path sink disruption test of the period "." token in Llama2-7B.

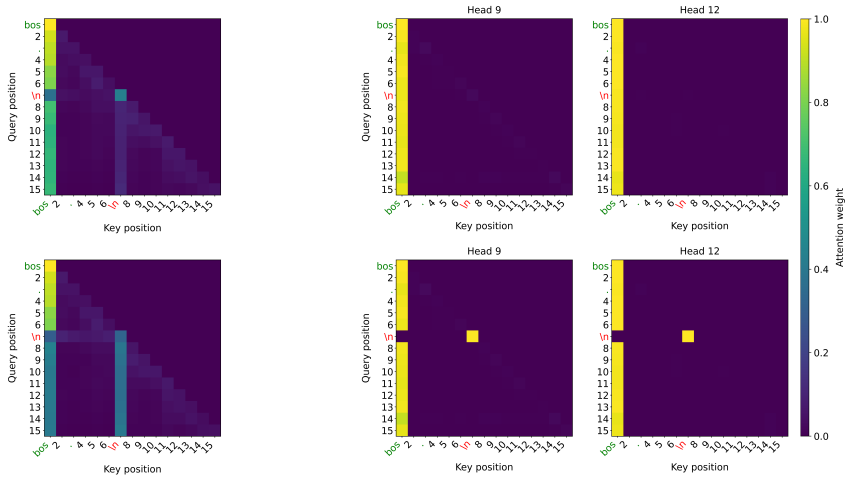


Figure 17: The results for the $\mathbf{V}$-path sink disruption test of the newline "\n" token in Mistral-7B.

A.6 How an Attention Sink Emerges in Subsequent Layers

In this section, we illustrate how the representation established for a secondary sink token in the first Transformer block propagates through subsequent layers and ultimately gives rise to sink behavior in the evaluated models. Upon entering the second block, this representation is then strongly activated by a small set of influential FFN neurons, often referred to as *super neurons* (Li et al., 2026). This activation results in the formation of massive activations (Sun et al., 2024; Oh et al., 2024) in the second layer's output, in which a small number of residual-stream dimensions have much larger value magnitudes than the rest of the dimensions. Because these dominant coordinates largely determine its direction, the pre-attention RMSNorm in the third layer maps the secondary sink token's second-layer output to a highly stable, nearly context-independent normalized representation. After key projection, this representation produces a key vector that is more strongly aligned with the queries of non-sink tokens than the keys of ordinary tokens. Consequently, in layers where sink behavior emerges, the secondary sink token receives systematically larger attention logits and attracts a disproportionate amount of attention from other positions. These observations support a coherent propagation mechanism: the first block establishes a distinctive token representation, the second-block FFN amplifies it through massive activations, and the resulting geometric alignment promotes secondary attention sinks in later layers.