Attention Sinks Are Functionally Essential in Softmax Transformers: Theoretical Evidence

Anonymous ACL submission

Abstract

Transformers often display an *attention sink*: probability mass concentrates on a fixed, content-agnostic position. We prove that computing a simple trigger-conditional behavior *necessarily* induces a sink in softmax self-attention models. Our results formalize a familiar intuition: normalization over a probability simplex must force attention to collapse onto a stable anchor to realize a default state. We instantiate this with a concrete task: when a designated trigger token appears, the model must return the *average of all preceding (non-BOS) token representations*, motivated by the view that the trigger aggregates the content seen so far while the BOS token contains no input-dependent content. We also prove that non-normalized ReLU attention can solve the same task without any sink, confirming that the normalization constraint is the fundamental driver of sink behavior. Experiments validate our predictions and demonstrate they extend beyond the theoretically analyzed setting: softmax models develop strong sinks while ReLU attention eliminates them in both single-head and multi-head variants.

1 Introduction

Transformers (Vaswani et al., 2017) frequently concentrate attention on an early position in a way that is largely insensitive to content. This *attention sink* has been reported for small and large models alike (Xiao et al., 2024; Gu et al., 2024; Guo et al., 2024). It occurs under a variety of positional schemes—absolute/learned embeddings, ALiBi, RoPE, and even without explicit positional encodings (Press et al., 2021; Su et al., 2021; Gu et al., 2024)—and similar behavior shows up in multimodal and vision settings (Kang et al., 2025; Wang et al., 2025; Feng and Sun, 2025). The breadth of contexts points to a pervasive pattern, not a peculiarity of any single model or training regime.

This pattern matters for practice. When probability mass concentrates on a fixed position, models may under-use available context and lose accuracy (Yu et al., 2024; Guo et al., 2024). Concentration can also worsen numerical issues relevant to compression/quantization (Sun et al., 2024; Lin et al., 2024) and distort attention-based analyses (Guo et al., 2024). Why is sink behavior so common? One plausible account is general *inductive bias*—a phenomenon documented in other settings (Soudry et al., 2024; Arora et al., 2019)—whereby preferences of the model class and learning setup steer solutions toward sinky circuits even when alternatives exist. In this work we argue that, in certain settings, this isn't the case, and sink behavior is *functionally essential* for the computation being performed.[1]

We investigate this claim theoretically in a simplified setting (section 3). Consider a synthetic, prefix-conditional task on sequences in which each token representation includes: (i) a binary *feature channel* indicating whether the token carries a trigger; (ii) a constant bias channel; (iii) a designated start token feature (a binary feature equal to one only for first token in the sequence, *BOS*); and (iv) i.i.d. samples from a continuous distribution with bounded probability density in the remaining coordinates. The target is intuitive: the model writes nothing to the residual stream at every position (i.e., outputs the zero vector), except at the unique trigger position where it should write the *empirical mean of all preceding non-BOS token vectors* (We exclude BOS from the average because it contains no input-dependent content).

Our main results are necessity theorems for *softmax* self-attention: for single-layer models (theorem 1), any hypothesis that achieves vanishing error

[1] We do not claim sinks are indispensable in all architectures (e.g., gated attention can mitigate them (Qiu et al., 2025)). Rather, we prove they are a necessary consequence of softmax attention.

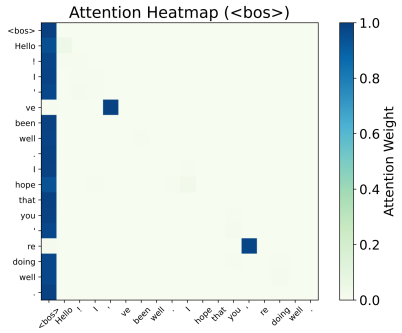

Figure 1: Reproduced from **Figure 4** of Barbero et al. (2025): an attention head that fires on an apostrophe trigger and otherwise attends to BOS.

on this task must place at least $1 - \varepsilon$ attention on a *fixed sink token* at *all* non-trigger positions before the trigger; for multi-layer models (theorem 2), at least one layer must exhibit sink behavior at some pre-trigger position. At a high level, we formalize a widely held intuition: normalization in probability, when combined with noisy keys, forces attention to concentrate on a stable anchor to keep the default output variance low in the no-trigger regime. We complement this with a constructive result (theorem 3): ReLU attention can solve the same task with zero attention on the BOS token, demonstrating that the normalization constraint is a driver of sink formation.

Experiments on both single-layer and multi-layer models provide supporting evidence (section 4). Models trained on the task develop attention sinks with near-unit mass on BOS when no trigger is present, aligning with our theoretical analysis. Swapping softmax for ReLU attention eliminates sink formation while preserving task accuracy, confirming that the softmax normalization constraint is a driver of the sink behavior. We observe these patterns across both single-layer and deeper multi-head architectures, demonstrating that our theoretical insights capture fundamental properties of softmax attention.

2 Related Work and Empirical Evidence

In realistic empirical settings, attention sinks frequently implement no-op behavior in the absence of specific triggers. Barbero et al. (2025) demonstrate this directly: their case study of an "apostrophe head" in Gemma 7B shows two operating modes—firing on apostrophe triggers and otherwise attending to BOS as a default no-op (fig. 1). Similarly, Guo et al. (2024) document an active–dormant head in Llama 2–7B that switches between active computation on code-like inputs and dormant sink behavior on text-like inputs. Notably, Guo et al. (2024) report that sink behavior diminishes under certain non-softmax/activation variants, consistent with our theoretical findings.

These works complement our theoretical perspective. Barbero et al. (2025) argue that sinks enable controlled information mixing, with BOS serving as a stable anchor. Guo et al. (2024) analyze the training dynamics behind sink formation—how these patterns emerge during optimization. In contrast, our work establishes a theoretical necessity of sink behavior in softmax attention and its absence in ReLU attention via expressiveness analyses regardless of optimization and training schemes. We include an illustrative figure from Barbero et al. (2025) (fig. 1)[2] to highlight that our synthetic task captures key aspects of real sink behavior—sinks emerge to implement a no-op when no trigger fires.

3 Theory and Results

We now set up our analysis. We introduce the task in section 3.2, explain why this task is meaningful and how its assumptions match realistic modeling in section 3.3, introduce the model architectures in section 3.4, and state our main necessity claims in section 3.5.

3.1 Notation and Setup

Let $n \geq 5$ denote the input dimension and $L \geq 4$ denote the sequence length. We use $\mathbb{1}\{\cdot\}$ for the indicator function, $[n] = \{1, \ldots, n\}$, and write sequences as $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)})$ with tokens $\mathbf{x}^{(i)} \in \mathbb{R}^n$.

3.2 Task Definition

We define a synthetic trigger-conditional task that exemplifies tasks in which we know attention sinks form in practice (as in fig. 1): the model must detect a trigger token and, *only at the trigger position*, write to the residual stream some summary of prior content.

Input Distribution. We denote the data distribution $\mathcal{D}$ as follows: For each sequence, sample a *trigger index* j uniformly from $\{3, \ldots, L\}$. Construct tokens as follows:

[2] Licensed under Creative Commons Attribution 4.0 (CC BY 4.0). Minor cropping for layout; no other changes. License: https://creativecommons.org/licenses/by/4.0/.

- **Position 1 (BOS):** $\mathbf{x}^{(1)} = (1, 0, 0, \ldots, 0)$ (fixed one-hot vector).
- **Position j (Trigger token):** $\mathbf{x}^{(j)} = (0, 1, 0, \ldots, 0)$ (fixed one-hot vector).
- **Positions $i \neq 1, j$:** Coordinate 1 (the BOS indicator) is zero; coordinate 2 is zero; coordinate n is fixed to one (a bias channel); and coordinates $3 \leq k \leq n-1$ are i.i.d. over some continuous distribution.

Target Output. The target output $\mathbf{y}^{(i)}$ is the zero vector $\mathbf{0}$ at all positions except the trigger position $i = j$, where it equals $\overline{\mathbf{x}} = (j-2)^{-1} \sum_{k=2}^{j-1} \mathbf{x}^{(k)}$, the mean of all preceding non-BOS tokens.

Loss Function. We evaluate hypotheses using the ℓ_∞ loss: $\mathcal{L}(f) = \sup_{(\mathbf{x},\mathbf{y}) \in \text{support}(\mathcal{D})} \left\| \mathbf{y}^{(i)} - f(\mathbf{x})^{(i)} \right\|_2$.

3.3 Task Motivation

This setup captures a basic and pervasive pattern in sequence modeling: *gate on a recognizable event, otherwise preserve a default state*. Real attention heads frequently exhibit exactly this bimodal behavior—firing on specific triggers to aggregate context, and otherwise attending to sinks as a no-op (Barbero et al., 2025; Guo et al., 2024). Our task distills this to its minimal form: detect a trigger and compute the mean of prior content, or write nothing. [3] The design choices are less arbitrary than they may appear. Many aspects are without loss of generality: the BOS and trigger feature channels can be any two orthogonal vectors via a change of basis; we fix them to coordinates 1 and 2 for simplicity. The constant bias channel models position-independent offsets that MLP layers can inject in practice.

3.4 Model Architecture

We study self-attention models with two variants of attention mechanisms. We denote the learnable parameter of a single-layer attention model by $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O \in \mathbb{R}^{n \times n}$ for queries, keys, values, and output projection respectively. For input sequence $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)})$, we calculate the attention weights $\alpha_{i,j}$ as defined below for each variant. The model output is then computed as $f(\mathbf{x})^{(i)} = \mathbf{W}_O \sum_{j=1}^{i} \alpha_{i,j} \mathbf{W}_V \mathbf{x}^{(j)}$.

Softmax Attention. The *attention weight* from position i to position $j \leq i$ is given by:

$$\alpha_{i,j} = \frac{\exp(\mathbf{x}^{(i)} \mathbf{W}_Q \mathbf{W}_K^T (\mathbf{x}^{(j)})^T)}{\sum_{k=1}^{i} \exp(\mathbf{x}^{(i)} \mathbf{W}_Q \mathbf{W}_K^T (\mathbf{x}^{(k)})^T)}$$

ReLU Attention. For ReLU attention, we replace the softmax normalization with element-wise ReLU. We divide the scores by the number of positions up to the current position i, excluding both the BOS token and the current token [4]. Namely, if we define $n_i = \max\{i - 2, 1\}$, then we have $\alpha_{i,j} = \text{ReLU}(\mathbf{x}^{(i)} \mathbf{W}_Q \mathbf{W}_K^T (\mathbf{x}^{(j)})^T)/n_i$.

Multi-Layer Attention. A D-layer softmax/ReLU model is the composition $f = f^{(D)} \circ \cdots \circ f^{(1)}$, where each $f^{(d)}$ is a single-layer softmax/ReLU attention model. We denote by $\alpha_{i,j}^{(d)}$ the attention weight at position i attending to position j in layer d.

3.5 Main Result

We are now ready to state our theoretical results. Our central contribution is threefold: (i) we establish that an attention sink is *necessary* at *every* position prior to the trigger for single-layer softmax attention to solve the trigger-conditional task (theorem 1); (ii) we prove that in multi-layer softmax attention, at least one position must exhibit sink behavior (theorem 2); [5] and (iii) we prove constructively that ReLU attention can solve the same task *without* any sink behavior (theorem 3). This contrast directly demonstrates that the softmax normalization constraint—not the task structure or optimization dynamics—is the fundamental driver of attention sinks. Proofs for theorems 1 to 3 can be found in Appendices C, D, and E respectively.

Theorem 1. *For any $\varepsilon > 0, \delta > 0, L \geq 4$ and $n \geq 5$, there exist constants $\eta > 0$ such that the following holds. Consider any single-layer softmax attention model f with loss $\mathcal{L}(f) \leq \eta$ on sequences with length L and dimension n where non-trigger*

[3] Our analysis applies almost as-is to a broader class of trigger-conditional problems, such as key-query retrieval where a query must extract a specific previous token (e.g., marked by a feature bit) while ignoring others, resembling the apostrophe head in fig. 1. We analyze the averaging task for clarity, leaving the formal characterization of the full class of tasks necessitating sinks to future work.

[4] This scaling is necessary because ReLU attention cannot naturally compute averages: concatenating the input sequence to itself would double the output at the final position while keeping the average the same. Moreover, a similar scaling *would not work for Softmax attention*, as our analysis would work over any such variant.

[5] Experiments (section B) show this existential bound is not loose: sinks do form, but not in all positions and layers.

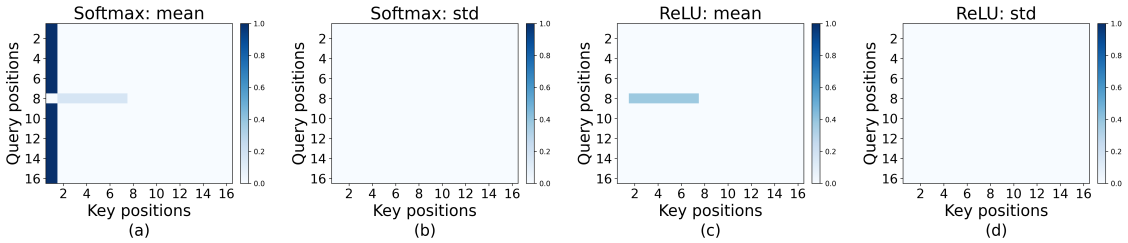

Figure 2: **Experimental validation: Theoretically analyzed model.** (a) Mean attention weights for softmax attention across 1000 test examples with trigger at position 8. Dark regions indicate high attention mass concentrated on BOS (position 1) before the trigger. (b) Standard deviation of softmax attention weights shows negligible variance, confirming stable sink behavior. (c) Mean attention weights for ReLU attention show no sink formation—attention on BOS remains near zero. (d) Standard deviation for ReLU attention confirms consistent behavior across examples.

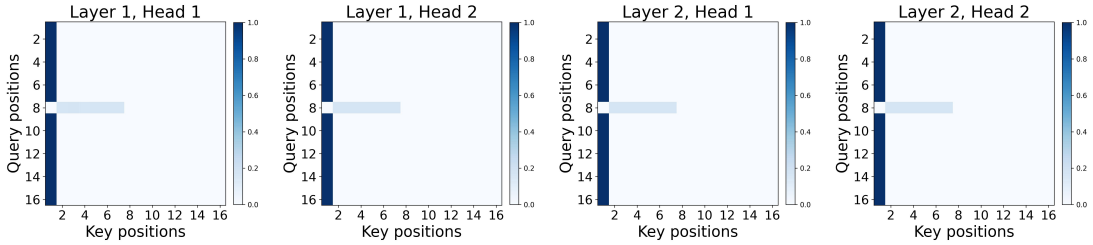

Figure 3: **Multi-layer multi-head validation.** Attention patterns for a 2-layer 2-head softmax model on a random input (with trigger at position 8). All heads exhibit strong sink behavior.

coordinates are drawn from some continuous distribution $\mathcal{D}$ with bounded probability density. Then with probability at least $1-\delta$ over the choice of $\mathbf{x}$ *with trigger index j, for all positions $1 < i < j$, we have $\alpha_{i,1} \geq 1-\varepsilon$.*

Theorem 2. *In the setting of theorem 1, but with a D-layer softmax attention model, there exists at least one layer $d \in \{1, \ldots, D\}$ and position $1 < i < j$ such that $\alpha_{i,1}^{(d)} \geq 1-\varepsilon$.*

Theorem 3. *For any $\delta > 0, L \geq 4$ and $n \geq 3$, there exists a one-layer ReLU attention model f with loss $\mathcal{L}(f) \leq \delta$ such that for any input sequence $\mathbf{x}$ with trigger index j and any position $i \neq j$ we have $\alpha_{i,1} = 0$.*

4 Experiments

We validate our theoretical predictions on the synthetic trigger-conditional task. In section 4.1, we train single-layer single-head models to validate theorem 1. In section 4.2, we train multi-layer multi-head models with residual connections to validate theorem 2. All experiments use sequences of length $L = 16$ with trigger at position $j = 8$; training details are in section A.

4.1 Single-Layer Models

We first validate theorem 1 on single-layer single-head models.

Experiment 1: Softmax Attention Forms Sinks. Theorem 1 predicts that softmax attention models achieving low loss must have a strong attention sink at all pre-trigger positions. To test this, we visualize the mean and standard deviation of attention weights across 1000 test examples (fig. 2, panels a and b). The model places near-unit attention mass on position 1 at every position before the trigger, with negligible variance across examples.

Experiment 2: ReLU Attention Avoids Sinks. Our constructive result establishes that ReLU attention can solve the same task with zero attention on BOS. We replace softmax with ReLU attention while keeping all other parameters identical (fig. 2, panels c and d). The ReLU model achieves comparable task accuracy without developing sink behavior: attention weights on position 1 remain near zero throughout the sequence.

4.2 Multi-Layer Multi-Head Models

Figure 3 shows attention patterns for a 2-layer 2-head softmax model. All heads exhibit strong sink behavior before the trigger, while ReLU variants eliminate sink formation entirely (section B). Similar patterns emerge in 4-layer 4-head models (section B), where sinks appear in some but not all heads, consistent with theorem 2 which guarantees existence rather than ubiquity.

5 Limitations

The synthetic trigger-conditional task, while empirically grounded in real sink behavior (Barbero et al., 2025; Guo et al., 2024), represents a specific computational pattern within a broader class of trigger-conditional problems. Our analysis likely extends to related tasks such as key-query retrieval where a query must extract a specific previous token (e.g., marked by a feature bit) while ignoring others—resembling the apostrophe head in fig. 1. We leave the formal characterization of the full class of tasks necessitating sinks to future work.

For multi-layer and multi-head models, our necessity result (theorem 2) guarantees that at least one layer must exhibit sink behavior at some position, but does not characterize which specific layers or heads form sinks. Our experiments (section B) show that sinks indeed do not form at all positions or in all heads, consistent with the existential nature of the theorem. Our analysis however does not provide characterizations for exactly where sinks emerge, we leave this to future work.

References

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. 2019. Implicit regularization in deep matrix factorization. *Preprint*, arXiv:1905.13655.

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

Wenfeng Feng and Guoying Sun. 2025. Edit: Enhancing vision transformers by mitigating attention sink through an encoder-decoder architecture. *Preprint*, arXiv:2504.06738.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2024. When attention sink emerges in language models: An empirical view. *Preprint*, arXiv:2410.10781.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I. Jordan, and Song Mei. 2024. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms. *Preprint*, arXiv:2410.13835.

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

Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. 2024. Duquant: Distributing outliers via dual transformation makes stronger quantized llms. *Preprint*, arXiv:2406.01721.

Ofir Press, Noah A. Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *Preprint*, arXiv:2108.12409.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *Preprint*, arXiv:2505.06708.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. 2024. The implicit bias of gradient descent on separable data. *Preprint*, arXiv:1710.10345.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, Hui Xue, Jialing Tao, Ranjie Duan, and Jiexi Liu. 2025. Mirage in the eyes: Hallucination attack on multi-modal large language models with only attention sink. *Preprint*, arXiv:2501.15269.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. *Preprint*, arXiv:2309.17453.

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

A Training Details

All models are trained using the Adam optimizer with learning rate 10^{-3} and batch size 128 over the ℓ_2 loss until the ℓ_∞ loss is less than 10^{-2} for the entire batch. We use input dimension $n = 16$ and sample non-designated coordinates i.i.d. from $\mathcal{U}(-1, 1)$.

B Additional Experimental Results

We present additional experimental results for multi-layer multi-head models beyond those shown in the main text. All models use the same training configuration described in section A.

B.1 ReLU Attention: 2-Layer 2-Head Model

Figure 4 shows attention patterns for a 2-layer 2-head model using ReLU attention with residual connections. Unlike the softmax variant (fig. 3), no head exhibits sink behavior—attention on BOS (position 1) remains near zero throughout all layers and heads, confirming that our theoretical prediction extends to multi-layer settings.

B.2 Larger Models: 4-Layer 4-Head Architecture

To further validate our findings at larger scale, we train 4-layer 4-head models with both softmax and ReLU attention. Figures 5 and 6 show representative attention patterns. The softmax variant exhibits strong sink behavior at least in one head every layer in the no-trigger regime, while the ReLU variant maintains near-zero attention on BOS throughout. These results provide additional evidence that the necessity of attention sinks in softmax models persists in deeper, wider architectures.

C Proof of Main Result

In this Appendix section we prove theorem 1.

C.1 Proof Sketch

First, we show that as $\eta \to 0$, for any two indices $i, h < j$, the value vector of the i-th token times the attention score $\alpha_{h,i}$ must tend to zero. Now, suppose no sink forms at some pre-trigger position. Then softmax normalization leaves nontrivial mass on content tokens before the trigger; to keep the output uniformly small across a positive measure input distribution, the model must *crush* those contributions, i.e., there is a positive-measure set S of tokens with $\|\mathbf{V}\mathbf{x}\|_2 \to 0$. This makes the pre-trigger output *insensitive* to which $\mathbf{x} \in S$ occurs, whereas the trigger output must be *sensitive* to that choice. The two requirements conflict under vanishing loss; hence a sink must exist at all pre-trigger positions.

C.2 Detailed Proof

Step 1: We can assume that $\mathbf{W}_K = \mathbf{I}$ and $\mathbf{W}_O = \mathbf{I}$. Let

$$\mathbf{B} := \mathbf{W}_Q \mathbf{W}_K^\top, \qquad \mathbf{V} := \mathbf{W}_O \mathbf{W}_V.$$

For any input, the scores and outputs are

$$s_{i,k} = \mathbf{x}^{(i)} \mathbf{B} (\mathbf{x}^{(k)})^\top, \qquad \hat{\mathbf{y}}^{(i)} = \sum_{k \le i} \alpha_{i,k} \, \mathbf{V} \, \mathbf{x}^{(k)},$$

with

$$\alpha_{i,k} = \frac{\exp(s_{i,k})}{\sum_{\ell \le i} \exp(s_{i,\ell})}.$$

Thus the attention depends on $(\mathbf{W}_Q, \mathbf{W}_K)$ only through $\mathbf{B}$, and the output depends on $(\mathbf{W}_O, \mathbf{W}_V)$ only through $\mathbf{V}$. Reparameterize by setting

$$\mathbf{W}_K := \mathbf{I}, \mathbf{Q} := \mathbf{B}, \mathbf{W}_O := \mathbf{I}, \mathbf{W}_V := \mathbf{V}$$

leaves $\alpha_{i,k}$ and $\hat{\mathbf{y}}^{(i)}$ unchanged, hence the loss is unchanged. Therefore, we will assume without loss of generality that $\mathbf{W}_K = \mathbf{I}$ and $\mathbf{W}_O = \mathbf{I}$, write $\mathbf{Q}$ for the query map, and $\mathbf{V}$ for the (combined) value map.

Step 2: Setup and pigeonhole principle. Fix $\varepsilon_0, \delta_0 > 0$ and suppose there exists a sequence of one-layer softmax models $\{f_t\}_{t=1}^\infty$ with $\eta_t := \mathcal{L}(f_t) \to 0$ such that, for each t, with probability at least δ_0 over $(\mathbf{x}, j) \sim \mathcal{D}$ there is a pre-trigger position $i < j$ violating the sink condition:

$$\alpha_{i,1} \le 1 - \varepsilon_0. \tag{1}$$

Since $\sum_{k \le i} \alpha_{i,k} = 1$, (1) implies that the total mass on non-BOS keys is at least ε_0. There are only finitely many index triples (i, h, j) with $2 \le h \le i < j \le L$. By a pigeonhole principle, there exist infinitely many times $t_{a_1}, t_{a_2}, \ldots$ and fixed indices $2 \le i^\star < j^\star \le L$ and $2 \le h^\star \le i^\star$, and a constant $\gamma > 0$ (e.g., $\gamma = \varepsilon_0 / L^2$), such that

$$\mathbb{P}\Big(\alpha_{i^\star,1} \le 1 - \varepsilon_0 \text{ and } \alpha_{i^\star,h^\star} \ge \gamma\Big) \ge \delta \tag{2}$$

for some $\delta > 0$ independent of t. By relabeling this subsequence, we assume without loss of generality that (2) holds for all t.

Step 3: Constructing tokens via Lemma 6. Since the event in (2) has positive probability at least δ, by Lemma 6 there exists $\varepsilon' > 0$ (independent of t) such that for every content coordinate $m \in \{3, \ldots, n-1\}$ there exist tokens $x^{(m)}, y^{(m)}$

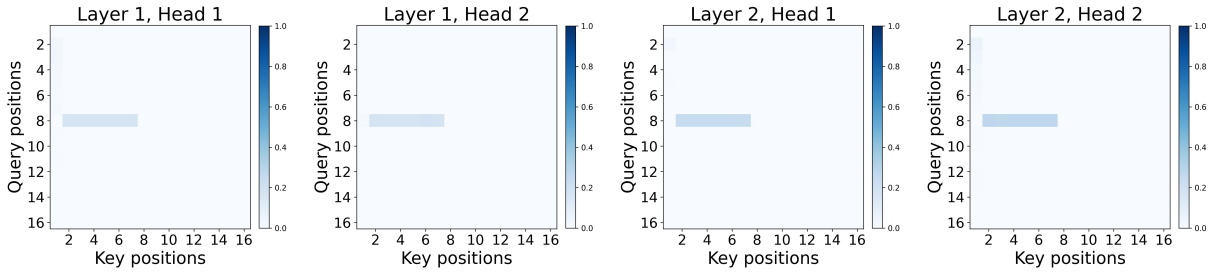

Figure 4: **ReLU attention: 2-layer 2-head model.** Attention patterns on a single test input (trigger at position 8). No sink formation occurs in any head—attention on BOS remains near zero, consistent with theorem 3.

with the following properties: (i) $x_k^{(m)} = y_k^{(m)}$ for all $k \neq m$, and $\left|x_m^{(m)} - y_m^{(m)}\right| \geq \varepsilon'$; and (ii) there exist sequences with either $x^{(m)}$ or $y^{(m)}$ at position $h^\star$ and with trigger index j satisfying $i^\star < j$, such that

$$\alpha_{i^\star,h^\star} \geq \gamma. \tag{3}$$

Step 4: Positive weight implies small values. By Lemma 5 (applied with the pair $(h^\star, i^\star)$), for every choice of token at position $h^\star$ we have

$$\left\|\alpha_{i^\star,h^\star}\mathbf{V}\mathbf{x}^{(h^\star)}\right\|_2 \leq 4\eta_t.$$

Combining with (3) yields that for any content coordinate m and any $\mathbf{z} \in \{x^{(m)}, y^{(m)}\}$,

$$\left\|\mathbf{V}\mathbf{z}\right\|_2 \leq \tfrac{4}{\gamma}\eta_t. \tag{4}$$

That is, the lower bound on $\alpha_{i^\star,h^\star}$ directly forces the value projections to be small for all tokens constructed in Step 2.

Step 5: Transplanting to $j = 3$ and deriving a contradiction. Fix t and abbreviate $\eta := \eta_t$. Pick a content coordinate $m \in \{3, \ldots, n-1\}$ and let $\mathbf{x}_t := x^{(m)}$ and $\mathbf{y}_t := y^{(m)}$ be the two tokens from Step 2 satisfying $|\mathbf{x}_{t,m} - \mathbf{y}_{t,m}| \geq \varepsilon'$. Instantiate two sequences by setting the trigger at $j = 3$ and taking $\mathbf{x}^{(2)} \in \{\mathbf{x}_t, \mathbf{y}_t\}$. At position $i = 3$ the target is

$$\mathbf{y}^{(3)} = \mathbf{x}^{(2)}. \tag{5}$$

Write

$$\beta_t(\mathbf{z}) \;:=\; \alpha_{3,3} \;\text{ for the sequence with } \mathbf{x}^{(2)} = \mathbf{z}, \tag{6}$$

$$\mathbf{v}_t \;:=\; \mathbf{V}_t\,\mathbf{x}^{(3)} \;=\; \mathbf{V}_t e_2. \tag{7}$$

By Lemma 1 and (4), at position 3 we can decompose

$$\hat{\mathbf{y}}^{(3)}(\mathbf{z}) = \underbrace{\alpha_{3,1}\,\mathbf{V}e_1 + \alpha_{3,2}\,\mathbf{V}\mathbf{z}}_{=:\,\mathbf{r}_t(\mathbf{z})} + \beta_t(\mathbf{z})\,\mathbf{v}_t, \tag{8}$$

$$\|\mathbf{r}_t(\mathbf{z})\|_2 \leq C_0\,\eta, \tag{9}$$

with $C_0 := 1 + \frac{4}{\gamma}$ independent of t. Consider the n-th (bias) coordinate. Since $(\mathbf{y}^{(3)})_n = (\mathbf{x}^{(2)})_n = 1$ and $0 < \beta_t(\mathbf{z}) \leq 1$, from (8) and the uniform loss bound we obtain

$$\begin{aligned}\left|\beta_t(\mathbf{z})\,(\mathbf{v}_t)_n - 1\right| &\leq \left|\hat{\mathbf{y}}_n^{(3)}(\mathbf{z}) - 1\right| + \left|(\mathbf{r}_t(\mathbf{z}))_n\right| \\ &\leq \eta + C_0\eta = C_1\eta, \end{aligned}\tag{10}$$

where $C_1 := 1 + C_0$. Hence, for all sufficiently large t,

$$(\mathbf{v}_t)_n \;\geq\; \frac{1 - C_1\eta}{\beta_t(\mathbf{z})} \;\geq\; 1 - C_1\eta \;>\; 0, \tag{11}$$

so $\mathbf{v}_t \neq \mathbf{0}$.

Let P_t denote the orthogonal projection onto $\mathbf{v}_t^\perp$. Since P_t is an orthogonal projection onto an $(n-1)$-dimensional subspace there must be at least one corrdinate $m \in 3, 4$ such that $\|P_t e_m\|_2 \geq 1/\sqrt{2}$; fix m to be that corrdinate. Now, applying P_t to (8) kills the $\mathbf{v}_t$ component:

$$P_t\hat{\mathbf{y}}^{(3)}(\mathbf{z}) = P_t\mathbf{r}_t(\mathbf{z}), \tag{12}$$

$$\|P_t\hat{\mathbf{y}}^{(3)}(\mathbf{z})\|_2 \leq \|\mathbf{r}_t(\mathbf{z})\|_2 \leq C_0\eta. \tag{13}$$

Therefore, for the two choices $\mathbf{z} = \mathbf{x}_t, \mathbf{y}_t$,

$$\begin{aligned}&\left\|P_t\hat{\mathbf{y}}^{(3)}(\mathbf{x}_t) - P_t\hat{\mathbf{y}}^{(3)}(\mathbf{y}_t)\right\|_2 \\ &\quad\leq\; \|P_t\mathbf{r}_t(\mathbf{x}_t)\|_2 + \|P_t\mathbf{r}_t(\mathbf{y}_t)\|_2 \\ &\quad\leq\; 2C_0\eta. \end{aligned}\tag{14}$$

On the other hand, $P_t\mathbf{y}^{(3)}(\mathbf{z}) = P_t\mathbf{z}$, so

$$\begin{aligned}&\left\|P_t\mathbf{y}^{(3)}(\mathbf{x}_t) - P_t\mathbf{y}^{(3)}(\mathbf{y}_t)\right\|_2 \\ &\quad= \|P_t(\mathbf{x}_t - \mathbf{y}_t)\|_2 \\ &\quad= \|P_t((\mathbf{x}_{t,m} - \mathbf{y}_{t,m})e_m)\|_2 \\ &\quad= |\mathbf{x}_{t,m} - \mathbf{y}_{t,m}|\,\|P_t e_m\|_2 \\ &\quad\geq \varepsilon'\,\|P_t e_m\|_2 \\ &\quad\geq \varepsilon'/\sqrt{2}. \end{aligned}\tag{15}$$

Where the third equality stems from the fact that $\mathbf{x}_t$ and $\mathbf{y}_t$ differ only on coordinate m.

Finally, by the triangle inequality and the uniform loss bound,

$$\begin{aligned}\left\|P_t\mathbf{y}^{(3)}(\mathbf{x}_t) - P_t\mathbf{y}^{(3)}(\mathbf{y}_t)\right\|_2 \\ \le \left\|P_t\hat{\mathbf{y}}^{(3)}(\mathbf{x}_t) - P_t\hat{\mathbf{y}}^{(3)}(\mathbf{y}_t)\right\|_2 + 2\eta \\ \le (2C_0+2)\eta, \end{aligned} \tag{16}$$

which contradicts (15) for all sufficiently small η, because $\varepsilon'\|P_te_m\|_2 > 0$ is independent of t. This completes the proof.

D Proof of Multi-Layer Result

Proof of theorem 2. **Setup and contradiction assumption.** Fix $\varepsilon_0, \delta_0 > 0$. Suppose for contradiction that there exists a sequence of D-layer softmax models $\{f_t\}_{t=1}^{\infty}$ with

$$\eta_t \;:=\; \mathcal{L}(f_t) \;\longrightarrow\; 0$$

such that, for every t,

$$\begin{aligned}\mathbb{P}\Big(\forall d \in \{1,\dots,D\},\ \forall\, 1 < i < j : \\ \alpha_{i,1}^{(d)} \le 1-\varepsilon_0\Big) \;\ge\; \delta_0.\end{aligned} \tag{17}$$

Let E_t denote the event inside the probability in (17). For each t, let $\mathbf{V}_t$ be the combined value map from Lemma 7, and write $\beta_{i,k}^{(t)}(\cdot)$ for the corresponding coefficients.

Crushing a positive-measure set of second tokens. On the event E_t, position 2 is pre-trigger (since $j \ge 3$) and for every layer d,

$$\alpha_{2,2}^{(d)} \;=\; 1-\alpha_{2,1}^{(d)} \;\ge\; \varepsilon_0.$$

Therefore, by Lemma 8,

$$\beta_{2,2}^{(t)}(\mathbf{x}) \;\ge\; \varepsilon_0^D \qquad \text{on } E_t. \tag{18}$$

Moreover, Lemma 10 applied to f_t yields

$$\left\|\beta_{2,2}^{(t)}(\mathbf{x})\,\mathbf{V}_t\mathbf{x}^{(2)}\right\|_2 \;\le\; 2\eta_t.$$

Combining with (18) gives

$$\|\mathbf{V}_t\mathbf{x}^{(2)}\|_2 \;\le\; \frac{2}{\varepsilon_0^D}\,\eta_t \qquad \text{on } E_t. \tag{19}$$

Define the measurable set

$$S_t \;:=\; \Big\{\mathbf{z}\in\mathbb{R}^n :\; \|\mathbf{V}_t\mathbf{z}\|_2 \le \tfrac{2}{\varepsilon_0^D}\eta_t\Big\}.$$

Since $E_t \subseteq \{\mathbf{x}^{(2)} \in S_t\}$ by (19), (17) implies

$$\mathbb{P}\big(\mathbf{x}^{(2)} \in S_t\big) \;\ge\; \delta_0. \tag{20}$$

By Lemma 6 applied to (20), there exists $\varepsilon' > 0$ (independent of t) such that for every content coordinate $m \in \{3,\dots,n-1\}$ there exist tokens $\mathbf{x}_t^{(m)}, \mathbf{y}_t^{(m)} \in S_t$ satisfying

$$\begin{aligned}\mathbf{x}_{t,k}^{(m)} = \mathbf{y}_{t,k}^{(m)} \text{ for all } k \ne m, \\ \left|\mathbf{x}_{t,m}^{(m)} - \mathbf{y}_{t,m}^{(m)}\right| \;\ge\; \varepsilon'.\end{aligned} \tag{21}$$

Transplanting to $j = 3$ and deriving a contradiction. Fix t and abbreviate $\eta := \eta_t$. Construct two sequences by setting the trigger at $j = 3$ and taking $\mathbf{x}^{(2)} \in \{\mathbf{x}_t^{(m)}, \mathbf{y}_t^{(m)}\}$. At position $i = 3$ the target is

$$\mathbf{y}^{(3)} \;=\; \mathbf{x}^{(2)}. \tag{22}$$

Write

$$\beta_t(\mathbf{z}) \;:=\; \beta_{3,3}^{(t)}(\mathbf{z}), \qquad \mathbf{v}_t \;:=\; \mathbf{V}_te_2.$$

By Lemma 7, for each choice $\mathbf{x}^{(2)} = \mathbf{z}$ we can decompose

$$\hat{\mathbf{y}}^{(3)}(\mathbf{z}) = \underbrace{\beta_{3,1}^{(t)}(\mathbf{z})\,\mathbf{V}_te_1 + \beta_{3,2}^{(t)}(\mathbf{z})\,\mathbf{V}_t\mathbf{z}}_{=:\,\mathbf{r}_t(\mathbf{z})} + \beta_t(\mathbf{z})\,\mathbf{v}_t. \tag{23}$$

Since $\beta_{3,1}^{(t)}(\mathbf{z}), \beta_{3,2}^{(t)}(\mathbf{z}) \le 1$, Lemma 9 gives $\|\mathbf{V}_te_1\|_2 \le \eta$, and $\mathbf{z} \in S_t$ implies $\|\mathbf{V}_t\mathbf{z}\|_2 \le \frac{2}{\varepsilon_0^D}\eta$. Therefore

$$\|\mathbf{r}_t(\mathbf{z})\|_2 \;\le\; C_0\,\eta, \qquad C_0 \;:=\; 1+\tfrac{2}{\varepsilon_0^D}. \tag{24}$$

Consider the n-th (bias) coordinate. For the $j = 3$ construction, we have $(\mathbf{y}^{(3)})_n = (\mathbf{x}^{(2)})_n = 1$. Using (23) and the uniform loss bound,

$$\begin{aligned}\left|\beta_t(\mathbf{z})\,(\mathbf{v}_t)_n - 1\right| \le \left|\hat{\mathbf{y}}_n^{(3)}(\mathbf{z}) - 1\right| + \left|(\mathbf{r}_t(\mathbf{z}))_n\right| \\ \le \eta + C_0\eta = C_1\eta,\end{aligned}$$

where $C_1 := 1 + C_0$. Hence $(\mathbf{v}_t)_n \ge 1 - C_1\eta > 0$ for all sufficiently large t, so $\mathbf{v}_t \ne \mathbf{0}$.

Let P_t denote the orthogonal projection onto $\mathbf{v}_t^{\perp}$. Since $\dim(\mathbf{v}_t^{\perp}) = n-1$, there exists at least one coordinate $m \in \{3,4\}$ such that

$$\|P_te_m\|_2 \;\ge\; 1/\sqrt{2}. \tag{25}$$

Fix such an m, and take $\mathbf{x}_t := \mathbf{x}_t^{(m)}$ and $\mathbf{y}_t := \mathbf{y}_t^{(m)}$ from (21).

Applying P_t to (23) kills the $\mathbf{v}_t$ component, giving $P_t\hat{\mathbf{y}}^{(3)}(\mathbf{z}) = P_t\mathbf{r}_t(\mathbf{z})$. Therefore,

$$\begin{aligned}&\left\|P_t\hat{\mathbf{y}}^{(3)}(\mathbf{x}_t) - P_t\hat{\mathbf{y}}^{(3)}(\mathbf{y}_t)\right\|_2 \\ &\quad\le \|P_t\mathbf{r}_t(\mathbf{x}_t)\|_2 + \|P_t\mathbf{r}_t(\mathbf{y}_t)\|_2 \le 2C_0\eta,\end{aligned} \tag{26}$$

using (24). On the other hand, by (22) we have $P_t\mathbf{y}^{(3)}(\mathbf{z}) = P_t\mathbf{z}$, and since $\mathbf{x}_t$ and $\mathbf{y}_t$ differ only in coordinate m,

$$\begin{aligned}&\left\|P_t\mathbf{y}^{(3)}(\mathbf{x}_t) - P_t\mathbf{y}^{(3)}(\mathbf{y}_t)\right\|_2 \\ &\quad= \|P_t(\mathbf{x}_t - \mathbf{y}_t)\|_2 \\ &\quad= |\mathbf{x}_{t,m} - \mathbf{y}_{t,m}| \cdot \|P_t e_m\|_2 \\ &\quad\ge \varepsilon'/\sqrt{2},\end{aligned} \tag{27}$$

using (21) and (25).

Finally, by the triangle inequality and the uniform loss bound,

$$\begin{aligned}&\left\|P_t\mathbf{y}^{(3)}(\mathbf{x}_t) - P_t\mathbf{y}^{(3)}(\mathbf{y}_t)\right\|_2 \\ &\quad\le \left\|P_t\hat{\mathbf{y}}^{(3)}(\mathbf{x}_t) - P_t\hat{\mathbf{y}}^{(3)}(\mathbf{y}_t)\right\|_2 + 2\eta \\ &\quad\le (2C_0 + 2)\eta,\end{aligned}$$

which contradicts (27) for all sufficiently small η. This contradiction completes the proof. □

E Proof of ReLU Result

In this section we provide the formal proof of theorem 3.

Proof of theorem 3. We give an explicit zero-loss construction with $\alpha_{i,1} = 0$ for all i.

Parameters. Set $\mathbf{W}_K = \mathbf{I}$, $\mathbf{W}_V = \mathbf{I}$, and $\mathbf{W}_O = \mathbf{I}$. Let e_r denote the r-th standard basis vector. Recall: coordinate 1 is BOS; coordinate 2 is the trigger flag; coordinate n is the bias channel with $\mathbf{x}_n^{(1)}, \mathbf{x}_n^{(j)} = 0$ and $\mathbf{x}_n^{(t)} = 1$ for $t \neq 1, j$. Define

$$\mathbf{W}_Q \;=\; e_2 e_n^\top,$$

Then for any positions i, k,

$$s_{i,k} \;=\; \mathbf{x}^{(i)}\mathbf{W}_Q\mathbf{W}_K^\top(\mathbf{x}^{(k)})^\top \;=\; \mathbf{x}_2^{(i)}\left(\mathbf{x}_n^{(k)}\right).$$

Now we calculate the attention score given a trigger position $j \in \{3, \dots, L\}$.

For any position $i \neq j$, we have $\mathbf{x}_2^{(i)} = 0$, hence $s_{i,k} = 0$ for all $k \le i$ and therefore $\alpha_{i,k} = 0$ for all $k \le i$. For the index $i = j$, we have $\mathbf{x}_2^{(j)} = 1$, hence $s_{j,k} = \mathbf{x}_n^{(k)}$ for all $k \le j$, which implies that $\alpha_{j,k} = 1/(j-2)$ for all $2 \le k \le j-1$ and $\alpha_{j,k} = 0$ otherwise. Thus $\mathrm{ReLU}(s_{j,k})$ is 1 exactly on the preceding non-BOS tokens $k \in \{2, \dots, j-1\}$ and 0 elsewhere.

Plugging this into the model output formula we get immediately that the loss is zero and that $\alpha_{i,1} = 0$ for all i, as needed. □

F Lemmas

Lemma 1. *Let f be a single-layer self-attention model as in §3.4 and write $\mathbf{V} := \mathbf{W}_O\mathbf{W}_V$. If the loss $\mathcal{L}(f)$ (see section 3.2) satisfies $\mathcal{L}(f) \le \eta$, then*

$$\|\mathbf{V}e_1\|_2 \le \eta.$$

Proof. By causality, at position $i = 1$ we have $\alpha_{1,1} = 1$, hence $\hat{\mathbf{y}}^{(1)} = \mathbf{V}e_1$. Since $\mathbf{y}^{(1)} = \mathbf{0}$ and $\|\hat{\mathbf{y}}^{(1)} - \mathbf{y}^{(1)}\|_2 \le \mathcal{L}(f) \le \eta$, the claim follows. □

Lemma 2. *Assume the attention mechanism is softmax. Fix any query $\mathbf{q} \in \mathbb{R}^n$ and two candidate sets of keys $S \subseteq T \subset \mathbb{R}^n$. For the softmax probabilities*

$$\sigma_S(\mathbf{k}) = \frac{\exp(\mathbf{q}^\top\mathbf{k})}{\sum_{\mathbf{r}\in S}\exp(\mathbf{q}^\top\mathbf{r})},$$

$$\sigma_T(\mathbf{k}) = \frac{\exp(\mathbf{q}^\top\mathbf{k})}{\sum_{\mathbf{r}\in T}\exp(\mathbf{q}^\top\mathbf{r})},$$

we have $\sigma_T(\mathbf{k}) \le \sigma_S(\mathbf{k})$ for every $\mathbf{k} \in S$.

Proof. The denominators satisfy

$$\begin{aligned}\sum_{\mathbf{r}\in T}\exp(\mathbf{q}^\top\mathbf{r}) &= \sum_{\mathbf{r}\in S}\exp(\mathbf{q}^\top\mathbf{r}) \\ &\quad + \sum_{\mathbf{r}\in T\setminus S}\exp(\mathbf{q}^\top\mathbf{r}) \\ &\ge \sum_{\mathbf{r}\in S}\exp(\mathbf{q}^\top\mathbf{r}),\end{aligned}$$

while the numerator for a fixed $\mathbf{k} \in S$ is the same in both fractions. □

Lemma 3. *Assume the attention mechanism is softmax. Consider any sequence from $\mathcal{D}$ and any non-trigger indices $1 < i < j$ and $1 < h < i < j$. Then:*

1. *(Self-reduction) Let $\widetilde{\alpha}_{2,2}$ denote the attention weight on the second token in the length-2 prefix $(\mathtt{BOS}, \mathbf{x}^{(i)})$, computed with the same $(\mathbf{W}_Q, \mathbf{W}_K)$. Then $\alpha_{i,i} \le \widetilde{\alpha}_{2,2}$.*
2. *(Pairwise reduction) Let $\widetilde{\alpha}_{3,2}$ denote the attention weight on the second token in the length-3 prefix $(\mathtt{BOS}, \mathbf{x}^{(i)}, \mathbf{x}^{(h)})$, computed with $(\mathbf{W}_Q, \mathbf{W}_K)$. Then $\alpha_{h,i} \le \widetilde{\alpha}_{3,2}$.*

Proof. For (1), at real position i the query equals $\mathbf{x}^{(i)}\mathbf{W}_Q$. Let S be the two keys $\{\mathbf{W}_K\mathbf{x}^{(1)}, \mathbf{W}_K\mathbf{x}^{(i)}\}$ and $T = \{\mathbf{W}_K\mathbf{x}^{(k)} : k \le i\}$. Lemma 2 (with this fixed query) gives the claim, noting that $\widetilde{\alpha}_{2,2} = \sigma_S(\mathbf{W}_K\mathbf{x}^{(i)})$ and $\alpha_{i,i} = \sigma_T(\mathbf{W}_K\mathbf{x}^{(i)})$.

For (2), at real position h the query equals $\mathbf{x}^{(h)}\mathbf{W}_Q$. Let $S = \{\mathbf{W}_K\mathbf{x}^{(1)}, \mathbf{W}_K\mathbf{x}^{(i)}, \mathbf{W}_K\mathbf{x}^{(h)}\}$ and $T = \{\mathbf{W}_K\mathbf{x}^{(k)} : k \le h\}$; apply Lemma 2 as before. □

Lemma 4. *In the setting of lemma 1, assume the attention mechanism is softmax. For every sequence in* $\mathrm{support}(\mathcal{D})$ *and every non-trigger index* $1 < i < j$,

$$\left\|\alpha_{i,i}\mathbf{V}\mathbf{x}^{(i)}\right\|_2 \le 2\eta.$$

Proof. Fix i and consider the length-2 prefix $(\texttt{BOS}, \mathbf{x}^{(i)})$. At its position 2 (which is pre-trigger), the output equals

$$\hat{\mathbf{y}}^{(2)} = \widetilde{\alpha}_{2,1}\mathbf{V}e_1 + \widetilde{\alpha}_{2,2}\mathbf{V}\mathbf{x}^{(i)},$$

with target $\mathbf{y}^{(2)} = \mathbf{0}$. Hence

$$\begin{aligned}\left\|\widetilde{\alpha}_{2,2}\mathbf{V}\mathbf{x}^{(i)}\right\|_2 &\le \|\hat{\mathbf{y}}^{(2)}\|_2 + \|\widetilde{\alpha}_{2,1}\mathbf{V}e_1\|_2 \\ &\le \eta + \eta = 2\eta,\end{aligned}$$

using Lemma 1 for the BOS term. By Lemma 3(1), $\alpha_{i,i} \le \widetilde{\alpha}_{2,2}$, and multiplying both sides by the fixed vector $\mathbf{V}\mathbf{x}^{(i)}$ yields the result. □

Lemma 5. *In the setting of lemma 1, assume the attention mechanism is softmax. For every sequence in* $\mathrm{support}(\mathcal{D})$ *and every pair of non-trigger indices* $1 < i < h < j$*:*

$$\left\|\alpha_{h,i}\mathbf{V}\mathbf{x}^{(i)}\right\|_2 \le 4\eta.$$

Proof. Consider first the length-3 prefix $(\texttt{BOS}, \mathbf{x}^{(i)}, \mathbf{x}^{(h)})$. At position 3 (pre-trigger), with target $\mathbf{y}^{(3)} = \mathbf{0}$,

$$\hat{\mathbf{y}}^{(3)} = \widetilde{\alpha}_{3,1}\mathbf{V}e_1 + \widetilde{\alpha}_{3,2}\mathbf{V}\mathbf{x}^{(i)} + \widetilde{\alpha}_{3,3}\mathbf{V}\mathbf{x}^{(h)}.$$

Therefore,

$$\begin{aligned}\left\|\widetilde{\alpha}_{3,2}\mathbf{V}\mathbf{x}^{(i)}\right\|_2 &\le \|\hat{\mathbf{y}}^{(3)}\|_2 + \|\widetilde{\alpha}_{3,1}\mathbf{V}e_1\|_2 \\ &\quad + \|\widetilde{\alpha}_{3,3}\mathbf{V}\mathbf{x}^{(h)}\|_2 \\ &\le \eta + \eta + 2\eta = 4\eta,\end{aligned}$$

using Lemma 1 for the BOS term and Lemma 4 for the self term. By Lemma 3(2), $\alpha_{h,i} \le \widetilde{\alpha}_{3,2}$. Multiplying by $\mathbf{V}\mathbf{x}^{(i)}$ gives the result. □

Lemma 6. *Let* $X = (X_1, \ldots, X_n) \sim \mu^{\otimes n}$*, where* μ *has a Lebesgue density* g *bounded by* $M := \sup_{x\in\mathbb{R}} g(x) < \infty$*. Fix* $\delta \in (0, 1]$*. Then there exists some* $\varepsilon' > 0$ *such that if a measurable set* $E \subset \mathbb{R}^n$ *satisfies* $\mathbb{P}(X \in E) \ge \delta$*, then for every coordinate* $j \in \{1, \ldots, n\}$ *there exist* $x, y \in E$ *such that*

$$x_k = y_k \text{ for all } k \ne j, \quad \text{and} \quad |x_j - y_j| \ge \varepsilon',$$

Proof. Fix j and, for $z \in \mathbb{R}^{n-1}$, set $E_j(z) := \{t \in \mathbb{R} : (z, t) \in E\}$. By Fubini and independence,

$$\mathbb{P}(X \in E) \;=\; \int \mu(E_j(z))\; d\mu^{\otimes(n-1)}(z).$$

Since μ has density g bounded by M, for any measurable $A \subset \mathbb{R}$ we have $\mu(A) \le M\,\lambda(A)$, where λ is Lebesgue measure. Hence

$$\begin{aligned}\delta \;&\le\; \int \mu(E_j(z))\, d\mu^{\otimes(n-1)}(z) \\ &\le\; M \int \lambda(E_j(z))\, d\mu^{\otimes(n-1)}(z).\end{aligned}$$

Therefore there exists z with $\lambda(E_j(z)) \ge \delta/M$. Any set $A \subset \mathbb{R}$ with Lebesgue measure $\lambda(A)$ has diameter at least $\lambda(A) - \eta$ for any $\eta > 0$, so we can choose $t_1, t_2 \in E_j(z)$ with $|t_1 - t_2| \ge \delta/M - \eta$ with $\eta < \delta/2M$. Setting $\varepsilon' = \delta/2M$ and taking $x = (z, t_1)$ and $y = (z, t_2)$ gives the claim. □

Lemma 7. *Let* $f = f^{(D)} \circ \cdots \circ f^{(1)}$ *be a D-layer causal self-attention model as in §3.4 . For each layer* $d \in \{1, \ldots, D\}$ *write*

$$\mathbf{V}^{(d)} \;:=\; \mathbf{W}_O^{(d)}\mathbf{W}_V^{(d)}.$$

$$\mathbf{V} \;:=\; \mathbf{V}^{(D)}\mathbf{V}^{(D-1)}\cdots\mathbf{V}^{(1)}$$

Then for every input sequence $\mathbf{x}$ *and every position* $i \in [L]$*, there exist coefficients* $\beta_{i,1}(\mathbf{x}), \ldots, \beta_{i,i}(\mathbf{x})$ *such that*

$$f(\mathbf{x})^{(i)} \;=\; \sum_{k=1}^{i} \beta_{i,k}(\mathbf{x})\, \mathbf{V}\mathbf{x}^{(k)}. \tag{28}$$

Moreover, for each i *we have* $\beta_{i,k}(\mathbf{x}) \ge 0$ *for all* $k \le i$ *and*

$$\sum_{k=1}^{i} \beta_{i,k}(\mathbf{x}) \;=\; 1.$$

Proof. Let $\mathbf{z}^{(0)} := \mathbf{x}$ and for $d \ge 1$ let $\mathbf{z}^{(d)} := f^{(d)}(\mathbf{z}^{(d-1)})$. Write $\alpha_{i,k}^{(d)}$ for the (softmax) attention

weight in layer d from position i to key $k \le i$. By definition of a single layer,

$$\mathbf{z}^{(d)(i)} \;=\; \sum_{k \le i} \alpha_{i,k}^{(d)}\, \mathbf{V}^{(d)} \mathbf{z}^{(d-1)(k)}.$$

Define $\beta_{i,k}^{(1)} := \alpha_{i,k}^{(1)}$, and for $d \ge 2$ define recursively

$$\beta_{i,k}^{(d)} \;:=\; \sum_{\ell:\, k \le \ell \le i} \alpha_{i,\ell}^{(d)}\, \beta_{\ell,k}^{(d-1)}.$$

A direct induction on d gives

$$\mathbf{z}^{(d)(i)} \;=\; \sum_{k \le i} \beta_{i,k}^{(d)}\, \mathbf{V}^{(d)} \cdots \mathbf{V}^{(1)} \mathbf{x}^{(k)}.$$

Nonnegativity and the row-sum identity follow since each $\alpha_{i,\cdot}^{(d)}$ is a probability vector. Taking $d = D$ and setting $\beta_{i,k} := \beta_{i,k}^{(D)}$ yields (28). □

Lemma 8. *In the setting of Lemma 7, for any input sequence* $\mathbf{x}$ *we have*

$$\beta_{2,2}(\mathbf{x}) \;=\; \prod_{d=1}^{D} \alpha_{2,2}^{(d)}(\mathbf{x}),$$

where $\alpha_{2,2}^{(d)}(\mathbf{x})$ *is the attention weight at position* 2 *attending to position* 2 *in layer* d*.*

Proof. In the recursion from the proof of Lemma 7, note that position 1 is causal and thus never depends on token 2, directly yielding the product formula. □

Lemma 9. *In the setting of Lemma 7, if the loss* $\mathcal{L}(f)$ *(see section 3.2) satisfies* $\mathcal{L}(f) \le \eta$ *then*

$$\|\mathbf{V} e_1\|_2 \;\le\; \eta.$$

Proof. By causality, at position $i = 1$ every layer attends only to position 1, hence $f(\mathbf{x})^{(1)} = \mathbf{V}\mathbf{x}^{(1)} = \mathbf{V}e_1$. Since $\mathbf{y}^{(1)} = \mathbf{0}$ and $\|f(\mathbf{x})^{(1)} - \mathbf{y}^{(1)}\|_2 \le \mathcal{L}(f) \le \eta$, the claim follows. □

Lemma 10. *In the setting of Lemma 7, assume softmax attention and that the loss* $\mathcal{L}(f)$ *(see section 3.2) satisfies* $\mathcal{L}(f) \le \eta$*. Then for every* $\mathbf{x}$ *in* $\mathrm{support}(\mathcal{D})$*,*

$$\left\|\beta_{2,2}(\mathbf{x})\, \mathbf{V}\mathbf{x}^{(2)}\right\|_2 \;\le\; 2\eta.$$

Proof. Since $j \ge 3$ always, position 2 is pre-trigger and the target satisfies $\mathbf{y}^{(2)} = \mathbf{0}$. By Lemma 7 with $i = 2$,

$$f(\mathbf{x})^{(2)} = \beta_{2,1}(\mathbf{x})\, \mathbf{V}e_1 + \beta_{2,2}(\mathbf{x})\, \mathbf{V}\mathbf{x}^{(2)}.$$

Thus

$$\begin{aligned}\left\|\beta_{2,2}(\mathbf{x})\, \mathbf{V}\mathbf{x}^{(2)}\right\|_2 &\le \|f(\mathbf{x})^{(2)}\|_2 \\ &\quad + \beta_{2,1}(\mathbf{x})\, \|\mathbf{V}e_1\|_2 \\ &\le \eta + \eta \\ &= 2\eta,\end{aligned}$$

using $\|f(\mathbf{x})^{(2)} - \mathbf{y}^{(2)}\|_2 \le \eta$, $\beta_{2,1}(\mathbf{x}) \le 1$, and Lemma 9. □

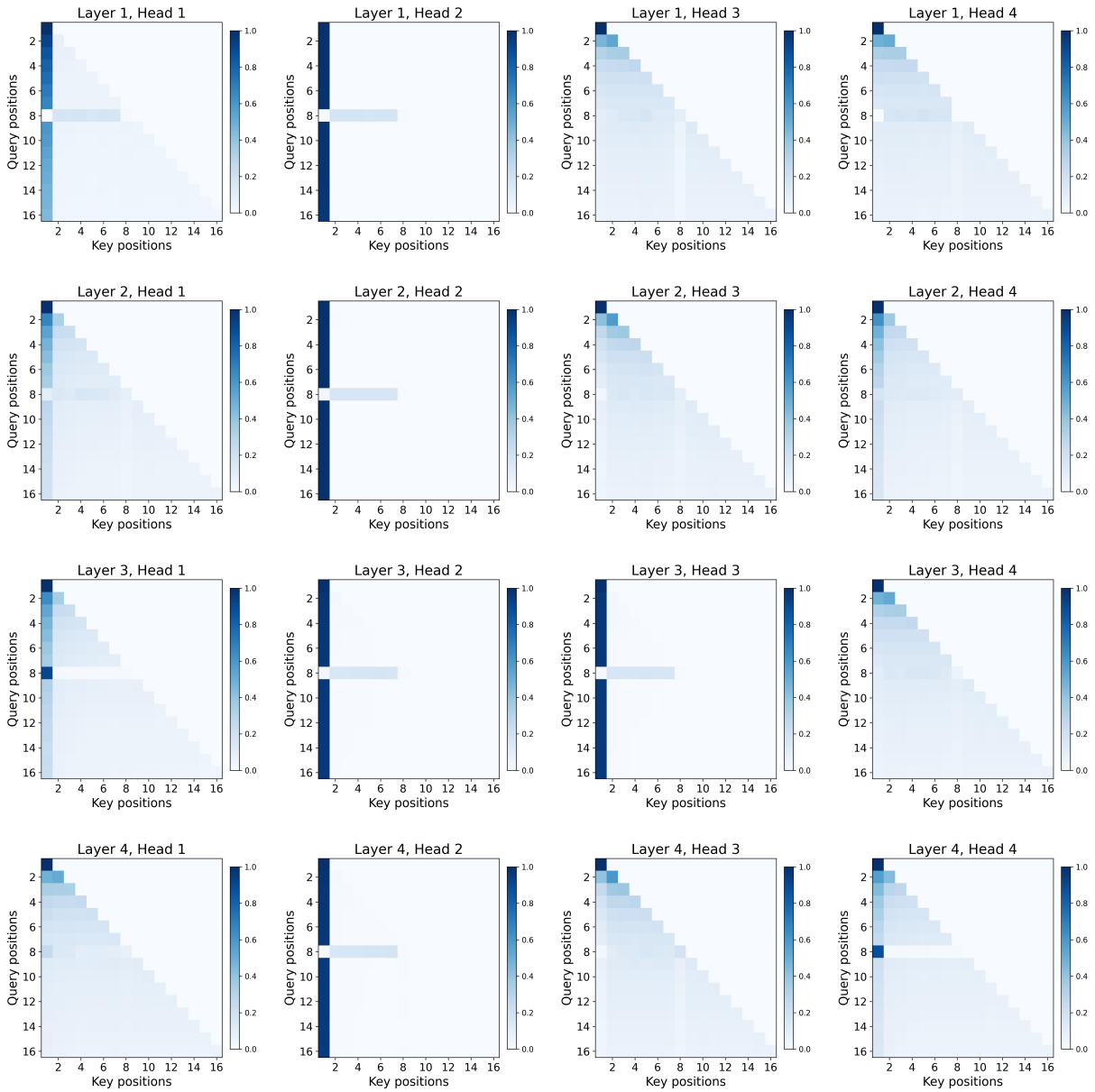

Figure 5: **Softmax attention: 4-layer 4-head model.** Representative attention patterns on a single test input showing strong sink at least in one head across all layers.

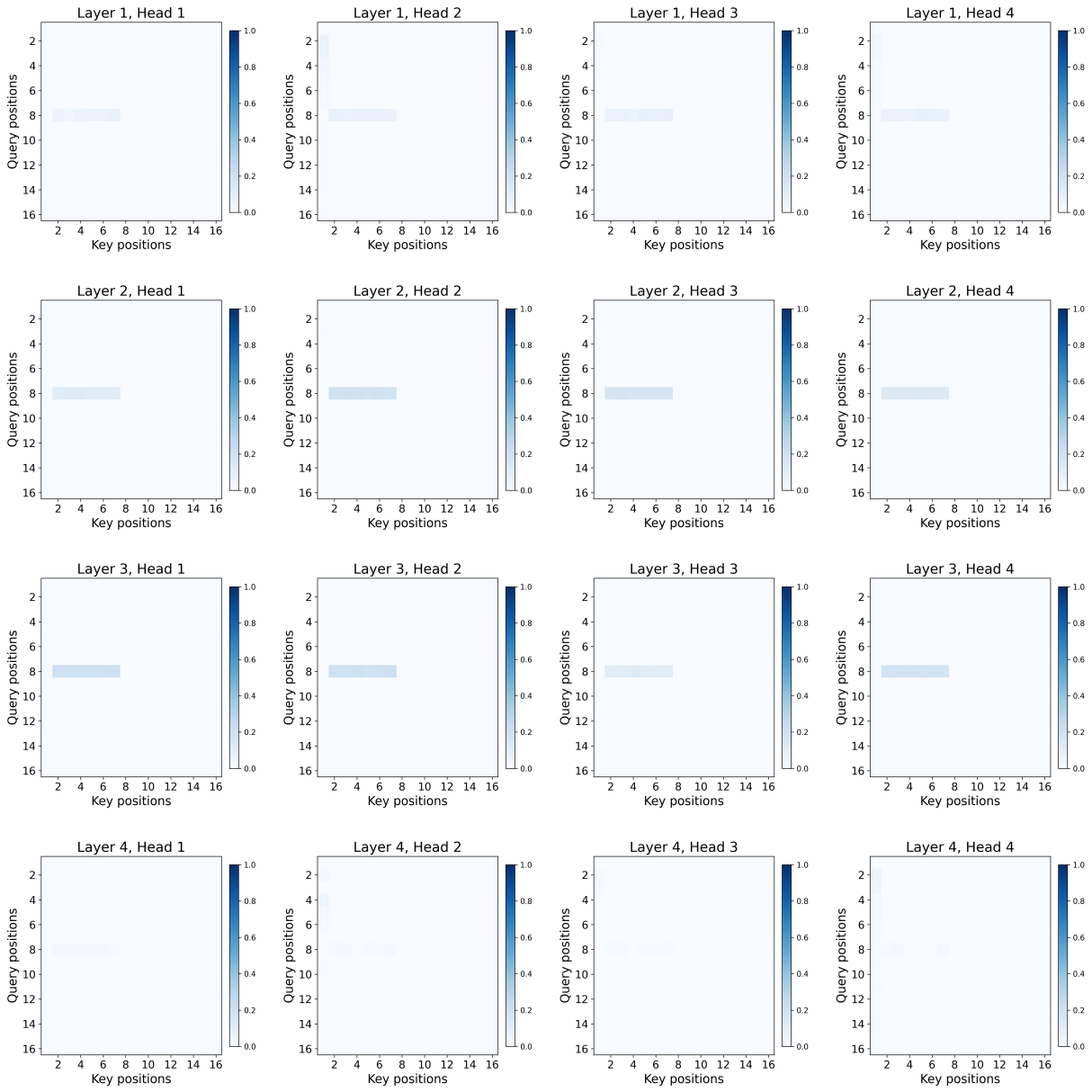

Figure 6: **ReLU attention: 4-layer 4-head model.** Representative attention patterns on a single test input showing absence of sink behavior across all layers.