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, in a single-layer single-head softmax self-attention model, computing a simple trigger-conditional behavior *necessarily* induces a sink whenever the trigger is absent. Our result formalizes 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, both in the analyzed architecture and in deeper 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 *inductive bias*: preferences of the model class and learning setup that 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.

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. Gaussian noise 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 result is a necessity theorem (theorem 1): in a single-layer and single-head *softmax* self-attention model, any hypothesis that achieves vanishing error on this task must place at least $1-\varepsilon$ attention on a *fixed sink token* whenever the trigger is absent. 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 com-

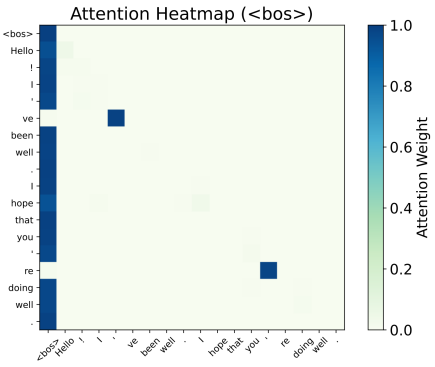

Figure 1: Reproduced from **Figure 4** of Barbero et al. (2025): an attention head that fires on an apostrophe trigger and otherwise attends to BOS.

plement this with a constructive result (theorem 2): 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 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, hinting that the softmax normalization constraint is a driver of the sink behavior. Critically, we observe qualitatively similar patterns in deeper multi-head architectures, demonstrating that our theoretical insights extend beyond the simplified analytical setting.

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)[1] 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, and state our main necessity claim in section 3.5.

3.1 Notation and Setup

Let $n \geq 4$ 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:

- **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.

[1] 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/.

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.

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 single-layer single-head self-attention models with two variants of attention mechanisms. We denote the learnable parameter matrices 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 define:

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

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

where the normalization is over all positions k up to the current position i (causal attention).

ReLU Attention. For ReLU attention, we replace the softmax normalization with element-wise ReLU activation. We normalize the scores by the number of positions up to the current position i, excluding both the BOS token and the current token. [2]

$$\alpha_{i,j} = \frac{\mathrm{ReLU}\big(\mathbf{x}^{(i)} \mathbf{W}_Q \mathbf{W}_K^T (\mathbf{x}^{(j)})^T\big)}{\max\{i-2,\, 1\}}$$

In both cases, the model output is computed as $f(\mathbf{x})^{(i)} = \mathbf{W}_O \sum_{j=1}^{i} \alpha_{i,j} \mathbf{W}_V \mathbf{x}^{(j)}$.

3.5 Main Result

We are now ready to state our theoretical results. Our central contribution is twofold: we establish that some attention sink is *necessary* for softmax attention to solve the trigger-conditional task (theorem 1), and we prove constructively that ReLU attention can solve the same task *without* any sink behavior (theorem 2). This contrast directly demonstrates that the softmax normalization constraint—not the task structure or optimization dynamics—is the fundamental driver of attention sinks.

Theorem 1 (Attention Sink Necessity)**.** *For any $\varepsilon > 0, \delta > 0, L \geq 4$ and $n \geq 4$, there exist constants $\eta > 0$ such that the following holds. Consider any one-layer softmax attention model f with loss $\mathcal{L}(f) \leq \eta$ on sequences with length L and dimension n where non-trigger coordinates are drawn from some continuous distribution $\mathcal{D}$. Then with probability at least $1 - \delta$ over the choice of $\mathbf{x}$ with trigger index j, there exists a position $1 < i < j$, such that $\alpha_{i,1} \geq 1 - \varepsilon$.*

Proof sketch (full proof in section C). First, we show that as $\eta \to 0$, the value vector for the BOS token must also tend to zero. Suppose no sink forms. Then softmax normalization leaves nontrivial mass on content tokens before the trigger; to keep the default output uniformly small across a continuous 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.

□

[2] This normalization 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. Alternatively, we could have defined the task to *sum* over past tokens for the ReLU model instead of *averaging*, which would yield an analogous theorem without requiring normalization. Moreover, please note that similar normalization *would not work for Softmax attention*, as our analysis would work over any such variant.

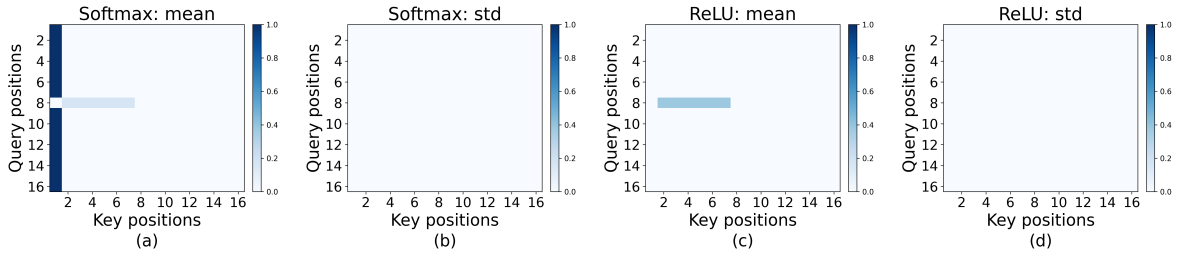

Figure 2: **Experimental validation: Theoretically analyzed model.** (a) Mean attention weights for softmax attention across 1000 test examples with trigger at position 8. Dark regions indicate high attention mass concentrated on BOS (position 1) before the trigger. (b) Standard deviation of softmax 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.

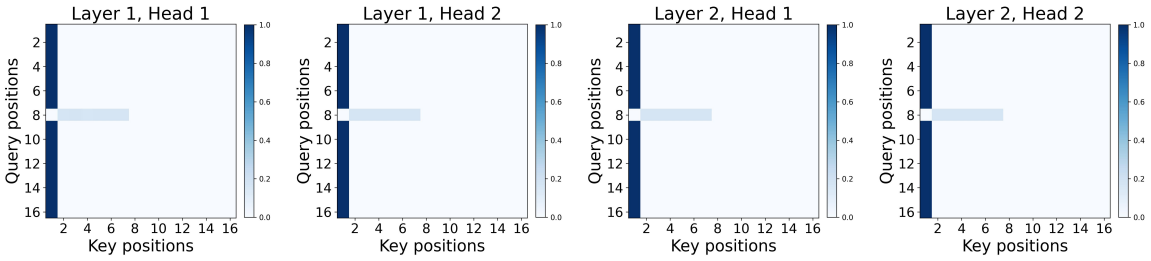

Figure 3: **Multi-layer multi-head validation.** Attention patterns for a 2-layer 2-head softmax model on a random input (with trigger at position 8). All heads exhibit strong sink behavior.

Theorem 2 (ReLU Attention Without Sinks)**.** *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 $\alpha_{i,1} = 0$.*

Proof sketch (full proof in section D). The proof is a simple construction that exploits the non-normalized nature of ReLU attention. □

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 our theoretical results. To examine whether our insights extend beyond the simplified analytical setting, in section 4.2 we train multi-layer multi-head models with residual connections. All experiments use sequences of length $L = 16$ with trigger at position $j = 8$; training details are in section A.

4.1 Theoretically Analyzed Model

We first validate our theoretical results on a single-layer single-head model analyzed in section 3.

Experiment 1: Softmax Attention Forms Sinks. Our necessity theorem predicts that softmax attention models achieving low loss must have a strong attention sink when no trigger is present. 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 in the no-trigger regime, 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 with near-unit attention mass on BOS before the trigger, while ReLU variants eliminate sink formation entirely (section B). Similar patterns emerge in 4-layer 4-head models (section B). These results demonstrate that the necessity of attention sinks in softmax models reflects a fundamental constraint that persists in deeper, more realistic architectures.

5 Limitations

Our theoretical analysis is based on a simplified single-layer single-head model. While our experiments with multi-layer multi-head models suggest the theoretical insights extend to more complex architectures, we have not proven necessity results for these settings. The synthetic trigger-conditional task, while illustrative and empirically grounded in real sink behavior (Barbero et al., 2025; Guo et al., 2024), represents a specific computational pattern. Future work should investigate whether formal necessity results can be established for deeper networks and whether similar mechanisms drive sink formation in the diverse real-world scenarios where they have been documented.

References

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.

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

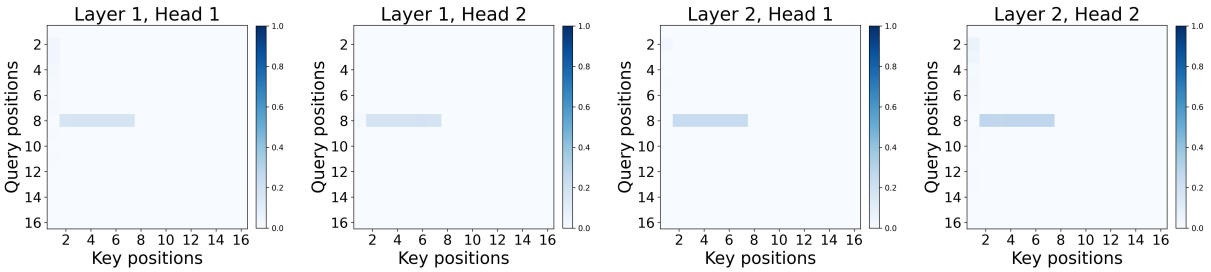

Figure 4: **ReLU attention: 2-layer 2-head model.** Attention patterns on a single test input (trigger at position 8). No sink formation occurs in any head—attention on BOS remains near zero, consistent with theorem 2.

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.

Simplifying the problem: 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 1: BOS must be (mapped to) zero as $\mathcal{L}(f) \to 0$. Recall $\mathbf{x}^{(1)} = e_1$ and the target is $\mathbf{y}^{(1)} = \mathbf{0}$. At position $i = 1$, causality implies $\alpha_{1,1} = 1$, hence

$$\hat{\mathbf{y}}^{(1)} = \sum_{k \le 1} \alpha_{1,k} \, \mathbf{V}\mathbf{x}^{(k)} = \mathbf{V}\mathbf{x}^{(1)} = \mathbf{V}e_1. \tag{1}$$

Therefore $\mathcal{L}(f) \to 0$ forces $\|\mathbf{V}e_1\|_2 \to 0$.

Step 2: From the negation, a closed interval on which $\mathbf{V}\mathbf{x}^{(2)}$ is arbitrarily small. Assume the negation of Theorem 1: there exist $\varepsilon_0, \delta_0 > 0$ and a sequence of hypotheses $\{f_t\}_{t=1}^\infty$ and losses $\eta_t \ge 0$ with $\mathcal{L}(f_t) \le \eta_t \to 0$ such that, on a set of inputs of probability at least δ_0,

$$\alpha_{i,1}^{(\eta)} \;\le\; 1 - \varepsilon_0 \qquad \text{for all } 1 < i < j. \tag{2}$$

Plugging in $i = 2$ to eq. (2) we have

$$\alpha_{2,2}^{(\eta)} = 1 - \alpha_{2,1}^{(\eta)} \;\ge\; \varepsilon_0. \tag{3}$$

For each η_t there must exist some $j_t \ge 3$ for which the event in (2) has positive probability. Condition on that j_t. Since the event has positive probability and $\mathbf{x}^{(2)}$ has a continuous density in its content coordinates, there exist $\epsilon' > 0$ (independent of t) and an open interval I_t of length $|I_t| \ge \epsilon'$ with probability measure $\ge \epsilon'$ such that, after further conditioning on the scalar content coordinate $\mathbf{x}_3^{(2)} \in I_t$ (holding the other coordinates of $\mathbf{x}^{(2)}$ fixed), the inequality

$$\alpha_{2,2}^{(\eta)} \;\ge \varepsilon_0 \tag{4}$$

still occurs with positive probability for all sufficiently small η_t. Moreover, by shrinking I_t if necessary, we can assume that there exists $\epsilon'' > 0$ (independent of t) such that $|I_t| \ge \epsilon'' > 0$ for all $t \in I_t$.

At $i = 2$ the target is $\mathbf{y}^{(2)} = \mathbf{0}$, so

$$\hat{\mathbf{y}}^{(2)} = \alpha_{2,1}^{(\eta)} \mathbf{V}\mathbf{x}^{(1)} + \alpha_{2,2}^{(\eta)} \mathbf{V}\mathbf{x}^{(2)} \tag{5}$$

$$=_{\eta_t \to 0} \alpha_{2,2}^{(\eta)} \mathbf{V}\mathbf{x}^{(2)}, \tag{6}$$

where the last equality used step 1. Thus, for any $\mathbf{x}^{(2)} \in I$, since $\|\hat{\mathbf{y}}^{(2)}\|_2 \le \eta_t$ and (4), we have that $\mathbf{V}\mathbf{x}^{(2)} \to 0$ as $\eta_t \to 0$.

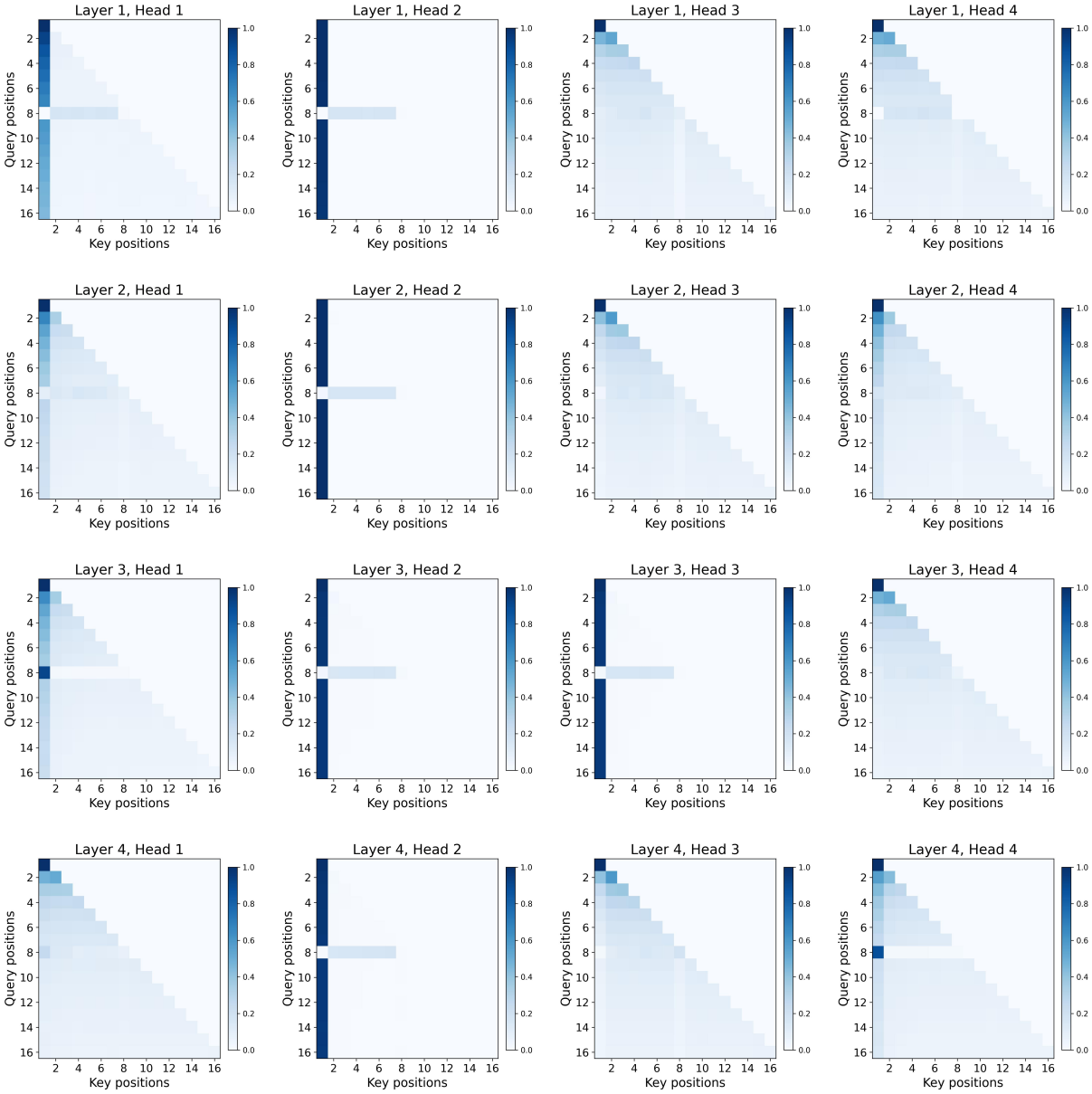

Figure 5: **Softmax attention: 4-layer 4-head model.** Representative attention patterns on a single test input showing strong sink at least in one head across all layers.

Step 3: Contradiction at $j = 3$. For simplicity of notation and ease of reading, we will write $\eta \to 0$ instead of $t \to \infty$, when in fact we mean taking the discrete series of η_t as $t \to \infty$ instead of a continuous η. Now, condition the loss on $j = 3$ and $\mathbf{x}_3^{(2)} \in I$. This loss must also go to zero (since this event has positive measure). The target at $i = 3$ is the mean of the preceding non-BOS tokens, which here reduces to

$$\mathbf{y}^{(3)} = \mathbf{x}^{(2)}. \tag{7}$$

This implies that as $\eta \to 0$ we have that $\hat{\mathbf{y}}^{(3)} \to \mathbf{x}^{(2)}$. On the other hand,

$$\hat{\mathbf{y}}^{(3)} = \alpha_{3,1}^{(\eta)} \mathbf{V}\mathbf{x}^{(1)} + \alpha_{3,2}^{(\eta)} \mathbf{V}\mathbf{x}^{(2)} + \alpha_{3,3}^{(\eta)} \mathbf{V}\mathbf{x}^{(3)} \tag{8}$$

$$=_{\eta \to 0} \alpha_{3,3}^{(\eta)} \mathbf{V}\mathbf{x}^{(3)} \tag{9}$$

Notice that $\hat{\mathbf{y}}^{(3)}$ and $\alpha_{3,3}^{(\eta)}$ depend on, and are completely determined by $\mathbf{x}^{(2)}$, so we will denote them as $\hat{\mathbf{y}}^{(3)}(\mathbf{x}^{(2)})$ and $\alpha_{3,3}^{(\eta)}(\mathbf{x}^{(2)})$ respectively. Note that $\mathbf{x}_n^{(2)}$ is a constant (since coordinate 4 is the bias index), x_3 is also a constant vector (as this is the vector of the trigger token), and for each η, $\mathbf{V}$ is a constant as well.

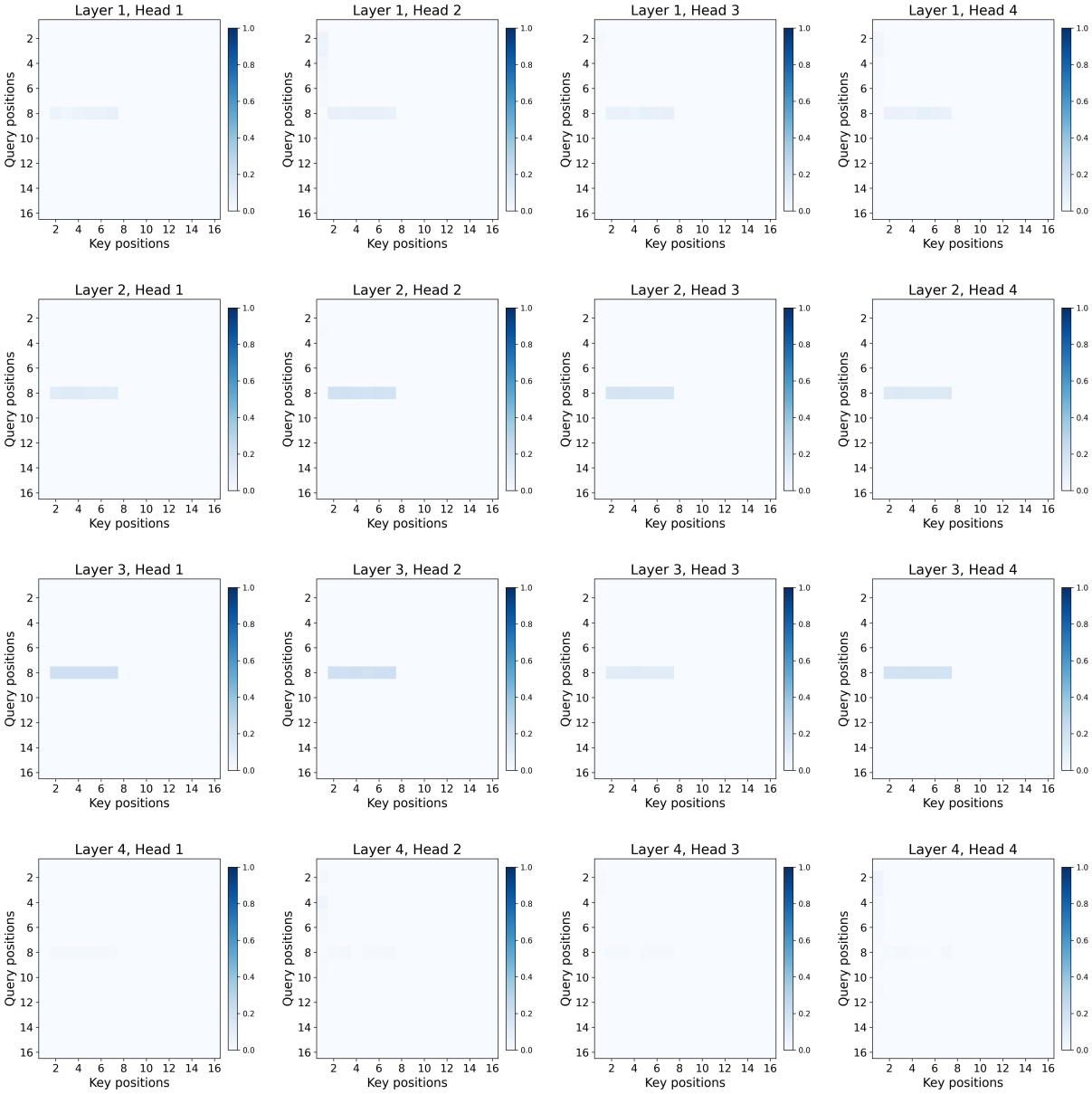

Figure 6: **ReLU attention: 4-layer 4-head model.** Representative attention patterns on a single test input showing absence of sink behavior across all layers.

This implies that as $\eta \to 0$, for any two inputs $\mathbf{x}^{(2)}, \mathbf{x}^{(2)\prime} \in I$ such that $\mathbf{x}_3^{(2)}, \mathbf{x}^{(2)\prime 3} \in I$, we have that $\alpha_{3,3}^{(\eta)}(\mathbf{x}^{(2)}) - \alpha_{3,3}^{(\eta)}(\mathbf{x}^{(2)\prime}) \to 0$ (one can simply subtract (9) for $\mathbf{x}^{(2)\prime}$ from (9) for $\mathbf{x}^{(2)}$ and take the limit $\eta \to 0$); This implies that

$$\hat{\mathbf{y}}^{(3)}(\mathbf{x}^{(2)}) - \hat{\mathbf{y}}^{(3)}(\mathbf{x}^{(2)\prime}) \to 0$$

but this contradicts the fact that the loss must go to zero! Since

$$y^{(3)}(\mathbf{x}^{(2)})_3 - y^{(3)}(\mathbf{x}^{(2)\prime})_3 =_{\eta \to 0} \mathbf{x}_3^{(2)} - \mathbf{x}_3^{(2)\prime}.$$

but for some $\mathbf{x}^{(2)}, \mathbf{x}^{(2)\prime} \in I$, we have that LHS $\geq \epsilon'$. Therefore, the negation (2) is false, completing the proof of the n=4 case.

D Proof of ReLU Result

In this section we provide the formal proof of theorem 2.

Proof of theorem 2. 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 $j \in \{3, \ldots, L\}$.

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