Attention Sinks Are Provably Necessary in Softmax Transformers: Evidence from Trigger-Conditional Tasks

Yuval Ran-Milo
Tel Aviv University
yuvalmilo@mail.tau.ac.il

Abstract

Transformers often display an *attention sink*: probability mass concentrates on a fixed, content-agnostic position. Are sinks a byproduct of the optimization/training regime? Or are they sometimes functionally necessary in softmax Transformers? We prove that, in some settings, it is the latter: 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 (e.g., when the model needs to ignore the input). We instantiate this with a concrete task: when a designated trigger token appears, the model must return the *average of all preceding token representations*, and otherwise output zero, a task which mirrors the functionality of attention heads in the wild (Barbero et al., 2025; Guo et al., 2024). 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]

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, as well as in diffusion language models (Kang et al., 2025; Wang et al., 2025; Feng and Sun, 2025; Rulli et al., 2025). The breadth of contexts points to a pervasive pattern, not a peculiarity of any single model or training regime.

This pattern has significant practical consequences. When probability mass concentrates on a fixed position, attention can be diverted away from other tokens and downstream accuracy can be affected (Yu et al., 2024). Sinks can also worsen numerical issues relevant to compression and quantization (Sun et al., 2024; Lin et al., 2024; Bondarenko et al., 2023; Son et al., 2024), distort attention-based interpretability analyses (Guo et al., 2024), and complicate streaming and long-context inference (Xiao et al., 2024). Analogous sink effects have also been documented in vision and multimodal settings, where they waste representational capacity on irrelevant visual tokens (Kang et al., 2025; Wang et al., 2025; Feng and Sun, 2025). (See Appendix B for an extended discussion on the practical motivations for mitigating attention sinks.)

Why is sink behavior so common? One plausible account is an *inductive bias*—a phenomenon documented in other settings (Soudry et al., 2024; Arora et al., 2019; Ran-Milo et al., 2026)— whereby the learning setup (model class and optimization procedure) steers solutions toward models that exhibit attention sinks, even when sink-free alternatives exist. In this work we argue that, in certain settings, this isn't the case, and sink behavior is *functionally essential*: all models that successfully compute a natural class of functions must exhibit sinks.[2]

We investigate this claim theoretically by introducing a *trigger-conditional task*: a model must output the mean of past tokens at a designated trigger position, and output zero (a no-op) everywhere else. This formulation captures the core mechanism of empirically observed attention heads "in the wild" (Barbero et al., 2025; Guo et al., 2024) which aggregate context when triggered and use a sink to remain dormant otherwise (see section 2

[1] An extended version of this paper is available at arxiv.org/pdf/2603.11487.

[2] We do not claim sinks are unavoidable in all architectures (e.g., sinks do not appear in gated attention or Mamba-based models (Qiu et al., 2025; Endy et al., 2025)). Rather, we prove they are a necessary consequence of softmax attention.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 70–89
July 2-7, 2026 ©2026 Association for Computational Linguistics

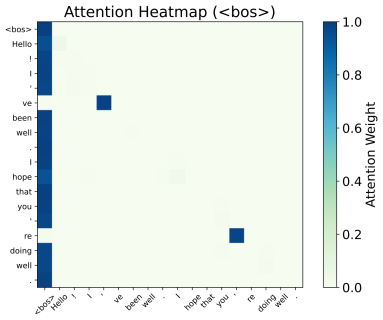

Figure 1: Reproduced from Barbero et al. (2025)[4]: an attention head that fires on an apostrophe trigger and otherwise attends to BOS.

for more details). We prove that attention sinks are necessary for softmax attention to solve this task. Specifically, we consider a synthetic, trigger-conditional task on sequences in which each token representation consists of: (i) a *BOS indicator* equal to one only for the first token; (ii) a *trigger indicator* equal to one only at the trigger position; (iii) a *non-trigger non-BOS indicator* equal to one for all remaining tokens; and (iv) i.i.d. samples from a continuous distribution in the content 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 *mean of all preceding non-*BOS [3] *token vectors*.

2 Sinks Empirically Enable No-Op Behaviors in Real Models

Our main results are necessity theorems for *softmax* self-attention: for single-layer models (theorem 1), any model that achieves vanishing error on this task must place attention arbitrarily close to 1 (the maximal possible value) on a *fixed sink token* (the BOS token) at *all* non-trigger positions; for multi-layer models (theorem 2), we show that at least one layer must exhibit sink behavior at some non-trigger position [5]. At a high level, we formalize a widely held intuition: normalization of attention scores forces the model to concentrate probability mass on a stable anchor whenever it needs to produce a default output, independent of the variable input content. We complement these necessity theorems 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 the primary driver of sink formation.

Experiments on both single-layer and multi-layer models provide supporting evidence (section 4). Single-layer softmax Transformers 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—rather than the task structure or optimization dynamics—is the fundamental driver of the sink behavior. We observe these patterns across both single-layer and deeper multi-head multi-layer architectures, demonstrating that our theoretical insights capture fundamental properties of normalization-based attention mechanisms.

In realistic empirical settings, attention sinks frequently appear in attention heads implementing a 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-operation (no-op) (fig. 1)[5]. 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; in particular, replacing softmax with ReLU attention eliminates sinks, consistent with our theoretical result (theorem 3).

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)[5] to highlight that our synthetic task captures key aspects of real sink behavior—sinks emerge to implement a no-op when no trigger fires. See Appendix H for more related works.

3 Theory and Results

We now set up our analysis. We introduce the task in section 3.2, explain why this task is meaningful

[3] We exclude BOS from the average because it contains no input-dependent content.

[5] Indeed, we empirically see in section 4 that sinks do form, but not in all positions and layers (see fig. 5).

[5] 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/.

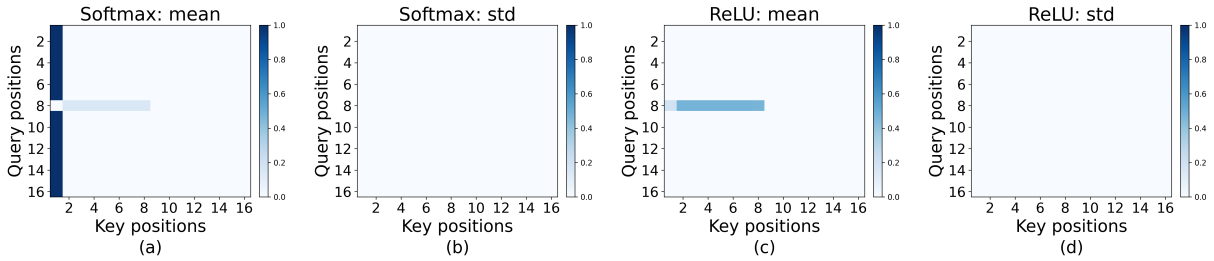

Figure 2: **Validation of theorem 1 and theorem 3.** (a) Mean attention weights for softmax attention across 1000 test examples with trigger at position 8. Dark regions indicate high attention mass concentrated on BOS (position 1) at non-trigger positions. (b) Standard deviation of softmax attention 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.

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

We write $\mathbb{R}_{>0}$ for the positive reals and $\mathbb{N}_{\geq k}$ for the natural numbers at least k. We use $\mathbb{1}\{\cdot\}$ for the indicator function and denote $[k] = \{1, \ldots, k\}$. Let $n \in \mathbb{N}_{\geq 5}$ be the input dimension and $L \in \mathbb{N}_{\geq 4}$ denote the sequence length. We write sequences as $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)})^\top \in \mathbb{R}^{L \times n}$ with tokens vectors $\mathbf{x}^{(i)} \in \mathbb{R}^{n \times 1}$.

3.2 Task Definition

We define a synthetic task designed to capture the mechanism of attention sinks "in the wild". Empirical studies show that attention heads in LLMs frequently implement trigger-conditional behavior: they aggregate context upon detecting a specific trigger, and attend to a sink token to effectively "switch off" otherwise (Barbero et al., 2025; Guo et al., 2024) (see section 2 for more details). Our task isolates this structure: the model must detect a trigger token and, *only at the trigger position*, write to the residual stream the mean of prior content, and write the zero vector at all other positions.

3.2.1 Input Distribution

We sample a *trigger position* j uniformly from $\{2, \ldots, L\}$. Input tokens lie in $\mathbb{R}^n$ (for some $n \in \mathbb{N}_{\geq 5}$) and use four designated coordinate types: (i) a BOS *indicator* (coordinate 1), equal to one only at position 1 and zero elsewhere; (ii) a *trigger indicator* (coordinate 2), equal to one only at position j and zero elsewhere; (iii) a *non-trigger non-*BOS *indicator* (coordinate 3), equal to one at all positions $i \neq 1, j$ and zero elsewhere; and (iv) *content coordinates* ($4 \leq k \leq n$), drawn i.i.d. from some continuous distribution at positions $i \neq 1$ (the BOS token has content coordinates fixed to zero, as it carries no input-dependent content).

3.2.2 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 $(j-1)^{-1} \sum_{k=2}^{j} \mathbf{x}^{(k)}$, the mean of all preceding non-BOS tokens (including the trigger itself).

3.2.3 Loss Function

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

3.3 Task Motivation and Justification

This setup captures a basic and pervasive pattern in sequence modeling: *aggregate context upon a trigger, otherwise perform a no-op* (Barbero et al., 2025; Guo et al., 2024) (see section 2 for more details). Our task distills this to its minimal form: detect a trigger and compute the mean of prior content, or, otherwise, output zero. [6]

The design choices we make are less arbitrary than they may appear. Many aspects are without loss of generality: the BOS indicator, the trigger indicator, and the non-trigger non-BOS indicator channels can be any three mutually orthogonal vectors via a change of basis; we fix them to coordinates 1, 2, and 3 for simplicity. While having such fixed indicator channels feels somewhat arbitrary, it is a natural way to model position-type information that an MLP layer can easily learn to inject into the residual stream in practice (e.g., by writing a constant vector).

[6] 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[5]. We analyze the averaging task for clarity, leaving the formal characterization of the full class of tasks necessitating sinks to future work.

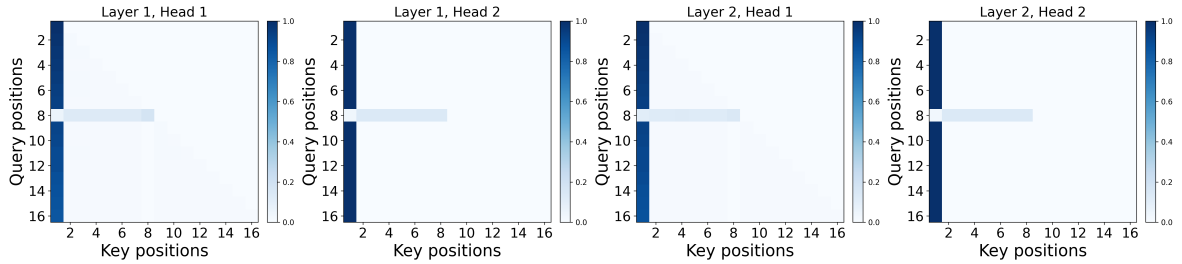

Figure 3: **Multi-layer multi-head validation.** Attention patterns for a 2-layer 2-head softmax model on a random input (with trigger at position 8). All heads exhibit strong sink behavior.

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 $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(L)})^\top \in \mathbb{R}^{L\times n}$, we calculate the attention weights $\alpha_{i,j}$ as defined below for each attention variant (softmax or ReLU). 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 $j \le i$ is given by:

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

ReLU Attention. For ReLU attention, we replace the softmax normalization with elementwise ReLU. We divide the scores by the number of positions up to the current position i, excluding the BOS token [7]. Namely, if we define $n_i = \max\{i-1, 1\}$, then we have $\alpha_{i,j} = \mathrm{ReLU}(\mathbf{x}^{(i)}\mathbf{W}_Q\mathbf{W}_K^\top(\mathbf{x}^{(j)})^\top)/n_i$.

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* non-trigger position 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); [8] 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.

Proof sketches for theorem 1, theorem 2, and theorem 3 can be found in section D.1, section E.1, and section F.1, respectively; full proofs are in section D, section E, and section F.

Theorem 1 (Single-Layer Attention Sink Necessity)**.** *For any $\varepsilon, \delta \in \mathbb{R}_{>0}$, $L \in \mathbb{N}_{\ge 4}$, $n \in \mathbb{N}_{\ge 5}$, and a bounded probability density function $\mathcal{P}$, there exists a constant $\eta \in \mathbb{R}_{>0}$ such that the following holds. Consider any single-layer softmax attention[9] model f with loss $\mathcal{L}(f) \le \eta$ on sequences with length L and dimension n where content coordinates are drawn from $\mathcal{P}$.[10] Then with probability at least $1-\delta$, for all non-trigger positions $i \ne j$, we have $\alpha_{i,1} \ge 1-\varepsilon$.*

Theorem 2 (Multi-Layer Attention Sink Necessity)**.** *For any $\varepsilon, \delta \in \mathbb{R}_{>0}$, $L \in \mathbb{N}_{\ge 4}$, $n \in \mathbb{N}_{\ge 5}$ and a bounded probability density function $\mathcal{P}$, there exists a constant $\eta \in \mathbb{R}_{>0}$ such that the following holds. Consider any D-layer softmax attention[9] model f with loss $\mathcal{L}(f) \le \eta$ on sequences with length L and dimension n where content coordinates are drawn from $\mathcal{P}$.[11] Then over all inputs with trigger position $j \ge 3$, with probability at least $1-\delta$, there exists at least one layer*

[7] 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 hold for any such variant.

[8] Indeed, we empirically see in section 4.2 (e.g., fig. 5) that sinks do form, but not in all positions and layers.

[9] Our analysis immediately extends to any attention mechanism whose weights $\alpha_{i,j}$ satisfy: (i) *normalization*—$\sum_{j\le i}\alpha_{i,j} \ge c$ for some constant $c > 0$; and (ii) *monotonicity*—inserting an additional key into positions $1, \ldots, i$ does not increase $\alpha_{i,j}$ for any existing key j.

[10] It is easy to show that such an f exists for any $\eta \in \mathbb{R}_{>0}$.

[11] It is easy to show that such an f exists for any $\eta \in \mathbb{R}_{>0}$.

$d \in \{1, \ldots, D\}$ *and a non-*BOS *non-trigger position* $i \neq j$ *such that* $\alpha_{i,1}^{(d)} \geq 1 - \varepsilon$.

Theorem 3 (ReLU Attention Without Sinks)**.** *For any* $L \in \mathbb{N}_{\geq 4}$ *and* $n \in \mathbb{N}_{\geq 3}$*, there exists a one-layer ReLU attention model* f *with loss* $\mathcal{L}(f) = 0$ *such that for any input sequence* $\mathbf{x}$ *with trigger position* j*, and any non-trigger 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 and theorem 3. In section 4.2, we validate our multi-layer findings (theorem 2) in more realistic settings by training multi-layer multi-head models with residual connections. All experiments use sequences of length $L = 16$; training details are in Appendix A. Code for reproducing our experiments is available at https://github.com/YuvMilo/sinks-are-provably-necessary.

4.1 Single-Layer Models

We first validate theorem 1 and theorem 3 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 non-trigger positions. To test this, we visualize the mean and standard deviation of attention weights across 1000 test examples with trigger position $j = 8$ (fig. 2, panels a and b). The model places near-unit attention mass on position 1 at every non-trigger position, with negligible variance across examples.

Experiment 2: ReLU Attention Avoids Sinks. Theorem 3 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. This observation reinforces that sinks are not a byproduct of the task or training dynamics, but a direct consequence of the normalization geometry.

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 across non-trigger positions. In deeper models, sinks appear in *some but not all* heads, consistent with theorem 2, which guarantees existence rather than ubiquity. For example, in a 4-layer 4-head softmax model that achieves low loss, head 3 in layer 4 places near-zero attention on BOS, while other heads in the same network develop clear sinks (see fig. 5 in Appendix C). This confirms the existential nature of theorem 2: a sink must exist *somewhere* in the network, but not in every head. Finally, replacing softmax with ReLU attention eliminates sink formation entirely in multi-layer models as well: no head of a 2-layer 2-head ReLU model develops a sink (see fig. 4 in Appendix C), and the same holds for a 4-layer 4-head ReLU model (see fig. 6).

5 Conclusions and Practical Implications

Our results show that for trigger-conditional behaviors, attention sinks are not an optimization artifact but a structural necessity: when a model must maintain a stable default (no-op) output on typical inputs while performing a content-dependent computation upon a recognizable trigger, softmax normalization forces sink formation. This has direct practical consequences: it can help practitioners distinguish between mitigation strategies that are fundamentally limited and those that address the root cause.

Specifically, sink-removal interventions operating *within* the softmax mechanism may be inherently limited for such computations. Penalizing BOS attention, spreading attention mass, or post-hoc reweighting may degrade the no-op guarantee, or cause the model to recreate an equivalent anchor elsewhere (a different position, head, or layer). In this sense, our results provide a principled reason to expect that simply "fighting" sinks without relaxing the simplex constraint can be counterproductive for trigger-conditional circuits: the sink may be the very mechanism that makes the circuit possible.

At the same time, the contrast with ReLU attention (theorem 3) clarifies a more promising direction. If sinks are undesirable for a downstream goal—e.g., they waste representational capacity (Yu et al., 2024), confound attention-based analyses (Guo et al., 2024), or create quantization-unfriendly outliers (Sun et al., 2024)—the right lever is to change how "off" states are represented, via non-normalized attention, explicit gating, or other mechanisms that can output zero without allocating probability mass.

More broadly, we hope our results can help guide future work on designing sink-free attention mechanisms that directly support no-op operations.

6 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 models, our necessity result (theorem 2) guarantees that at least one layer must exhibit sink behavior at some non-trigger position, but does not characterize which specific layer this must be. Our experiments extend this to multi-head architectures and confirm that sinks indeed do not form in all heads or layers (section C), consistent with the existential nature of the theorem; understanding exactly where sinks emerge would likely require a dynamical analysis of how optimization selects among valid solutions, which we leave to future work.

Finally, it would be interesting to investigate whether other special tokens that are stable and always present in the input (e.g., `<|think|>` in reasoning models) exhibit similar sink behavior, and to investigate the relatively newly discovered phenomenon of "secondary attention sinks" (Wong et al., 2026). We leave this direction for future work as well.

Acknowledgments

I thank Yotam Alexander, Amit Elhelo, Daniela Gottesman, Eden Lumbroso and Yoni Slutsky for illuminating discussions. Special thanks to my advisor Nadav Cohen for his guidance and mentorship. We used AI assistance for writing and code development. This work was supported by the European Research Council (ERC) grant NN4C 101164614, a Google Research Scholar Award, a Google Research Gift, Meta, the Yandex Initiative in Machine Learning, the Israel Science Foundation (ISF) grant 1780/21, the Tel Aviv University Center for AI and Data Science, the Adelis Research Fund for Artificial Intelligence, Len Blavatnik and the Blavatnik Family Foundation, and Amnon and Anat Shashua.

References

Anand, Umberto Cappellazzo, Stavros Petridis, and Maja Pantic. 2026. Mitigating attention sinks and massive activations in audio-visual speech recognition with llms. *Preprint*, arXiv:2510.22603.

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.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2023. Quantizable transformers: Removing outliers by helping attention heads do nothing. *Preprint*, arXiv:2306.12929.

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

Enrique Queipo de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. 2026. Attention sinks and compression valleys in llms are two sides of the same coin. *Preprint*, arXiv:2510.06477.

Nir Endy, Idan Daniel Grosbard, Yuval Ran-Milo, Yonatan Slutzky, Itay Tshuva, and Raja Giryes. 2025. Mamba knockout for unraveling factual information flow. *Preprint*, arXiv:2505.24244.

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

Zichuan Fu, Wentao Song, Yejing Wang, Xian Wu, Yefeng Zheng, Yingying Zhang, Derong Xu, Xuetao Wei, Tong Xu, and Xiangyu Zhao. 2025. Sliding window attention training for efficient large language models. *Preprint*, arXiv:2502.18845.

Zizhuo Fu, Wenxuan Zeng, Runsheng Wang, and Meng Li. 2026. Attention sink forges native moe in attention layers: Sink-aware training to address head collapse. *Preprint*, arXiv:2602.01203.

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.

Victoria Hankemeier and Malte Schilling. 2026. Stochastic parroting in temporal attention – regulating the diagonal sink. *Preprint*, arXiv:2602.10956.

Jonghyun Hong and Sungyoon Lee. 2025. Variance sensitivity induces attention entropy collapse and instability in transformers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8360–8378, Suzhou, China. Association for Computational Linguistics.

Sayed Mohammadreza Tayaranian Hosseini, Amir Ardakani, and Warren J. Gross. 2026. Innerq: Hardware-aware tuning-free quantization of kv cache for large language models. *Preprint*, arXiv:2602.23200.

Xingyue Huang, Xueying Ding, Mingxuan Ju, Yozen Liu, Neil Shah, and Tong Zhao. 2026. Threshold differential attention for sink-free, ultra-sparse, and non-dispersive language modeling. *Preprint*, arXiv:2601.12145.

Mingyu Jin, Kai Mei, Wujiang Xu, Mingjie Sun, Ruixiang Tang, Mengnan Du, Zirui Liu, and Yongfeng Zhang. 2025. Massive values in self-attention modules are the key to contextual knowledge understanding. *Preprint*, arXiv:2502.01563.

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.

Ziyong Lin, Haoyi Wu, Shu Wang, Kewei Tu, Zilong Zheng, and Zixia Jia. 2025. Look both ways and no sink: Converting LLMs into text encoders without training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22839–22853, Vienna, Austria. Association for Computational Linguistics.

Guozhi Liu, Weiwei Lin, Tiansheng Huang, Ruichao Mo, Qi Mu, Xiumin Wang, and Li Shen. 2026. Surgery: Mitigating harmful fine-tuning for large language models via attention sink. *Preprint*, arXiv:2602.05228.

Andrew Lu, Wentinn Liao, Liuhui Wang, Huzheng Yang, and Jianbo Shi. 2025. Artifacts and attention sinks: Structured approximations for efficient vision transformers. *Preprint*, arXiv:2507.16018.

Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. 2025. To sink or not to sink: Visual information pathways in large vision-language models. *Preprint*, arXiv:2510.08510.

Aidar Myrzakhan, Tianyi Li, Bowei Guo, Shengkun Tang, and Zhiqiang Shen. 2026. Sink-aware pruning for diffusion language models. *Preprint*, arXiv:2602.17664.

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, Zeyu Huang, Kaiyue Wen, Peng Jin, Bo Zheng, Yuxin Zhou, Haofeng Huang, Zekun Wang, Xiao Li, Huaqing Zhang, Yang Xu, Haoran Lian, Siqi Zhang, Rui Men, Jianwei Zhang, Ivan Titov, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. A unified view of attention and residual sinks: Outlier-driven rescaling is essential for transformer training. *Preprint*, arXiv:2601.22966.

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.

Yuval Ran-Milo, Yotam Alexander, Shahar Mendel, and Nadav Cohen. 2026. Outcome-based rl provably leads transformers to reason, but only with the right data. *Preprint*, arXiv:2601.15158.

Oliver Richter and Roger Wattenhofer. 2020. Normalized attention without probability cage. *Preprint*, arXiv:2005.09561.

Maximo Eduardo Rulli, Simone Petruzzi, Edoardo Michielon, Fabrizio Silvestri, Simone Scardapane, and Alessio Devoto. 2025. Attention sinks in diffusion language models. *Preprint*, arXiv:2510.15731.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. 2025. What are you sinking? a geometric approach on attention sink. *Preprint*, arXiv:2508.02546.

Pedro Sandoval-Segura, Xijun Wang, Ashwinee Panda, Micah Goldblum, Ronen Basri, Tom Goldstein, and David Jacobs. 2025. Using attention sinks to identify and evaluate dormant heads in pretrained llms. *arXiv preprint arXiv:2504.03889*.

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

Jaewon Sok, Jewon Yeom, Seonghyeon Park, Jeongjae Park, and Taesup Kim. 2026. Garbage attention in large language models: Bos sink heads and sink-aware pruning. *Preprint*, arXiv:2601.06787.

Seungwoo Son, Wonpyo Park, Woohyun Han, Kyuyeun Kim, and Jaeho Lee. 2024. 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*, pages 2242–2252, Miami, Florida, USA. Association for Computational Linguistics.

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.

Zunhai Su and Kehong Yuan. 2025. Kvsink: Understanding and enhancing the preservation of attention sinks in kv cache quantization for llms. *Preprint*, arXiv:2508.04257.

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

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

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.

Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. 2025. Softmax is not enough (for sharp size generalisation). *Preprint*, arXiv:2410.01104.

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.

Jeffrey T. H. Wong, Cheng Zhang, Louis Mahon, Wayne Luk, Anton Isopoussu, and Yiren Zhao. 2026. On the existence and behavior of secondary attention sinks. *Preprint*, arXiv:2512.22213.

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

Jing Xiong, Liyang Fan, Hui Shen, Zunhai Su, Min Yang, Lingpeng Kong, and Ngai Wong. 2026. Dope: Denoising rotary position embedding. *Preprint*, arXiv:2511.09146.

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

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.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. 2025. Attention sinks: A 'catch, tag, release' mechanism for embeddings. *Preprint*, arXiv:2502.00919.

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

Zihou Zhang, Zheyong Xie, Li Zhong, Haifeng Liu, Yao Hu, and Shaosheng Cao. 2026. One token is enough: Improving diffusion language models with a sink token. *Preprint*, arXiv:2601.19657.

Zayd M. K. Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. 2026. Softpick: No attention sink, no massive activations with rectified softmax. *Preprint*, arXiv:2504.20966.

A Training Details

All models are trained using the Adam optimizer (β_1=0.9, β_2=0.95) with batch size 128 over the ℓ_2 loss until the ℓ_∞ loss is less than 10^{-2} for the entire batch. Single-layer models use learning rate 10^{-3}; multi-layer models use learning rate 10^{-4}. We use input dimension $n = 16$ and sample content coordinates i.i.d. from $\mathcal{U}(-1, 1)$.

B Practical Impact of Attention Sinks

The goal of this section is to detail the empirical motivation for our theoretical study. Attention sinks have been shown to affect several aspects of model performance and deployment. We briefly survey the evidence here to motivate the practical importance of understanding their origin.

Accuracy and context utilization. When probability mass concentrates on a fixed position, attention can be diverted away from other tokens and downstream accuracy can be affected (Yu et al., 2024). Guo et al. (2024) document "active–dormant" heads in which dormant sink behavior effectively wastes representational capacity.

Compression and quantization. Attention sinks are correlated with outlier activations that complicate model compression. Sun et al. (2024) identify massive activations tied to sink tokens, and Lin et al. (2024) show that these outliers are a key challenge for quantization.

Streaming and long-context inference. Attention sinks complicate streaming and rolling-window KV-cache strategies: Xiao et al. (2024) show that evicting sink tokens from the cache causes catastrophic performance degradation, and that explicitly retaining them is necessary for stable generation on sequences far beyond the training length.

Vision and multimodal models. Analogous sink effects appear in vision Transformers and multimodal models. Kang et al. (2025) show that visual attention sinks allocate high attention weights to irrelevant visual tokens, wasting representational capacity. Wang et al. (2025) demonstrate that attention sinks in multimodal models can be exploited to induce hallucinations, and Feng and Sun (2025) propose architectural modifications to mitigate sink behavior in vision Transformers.

Interpretability. Sinks distort attention-based analyses by concentrating probability mass on tokens that carry no content-relevant information, complicating efforts to use attention patterns for model interpretation (Guo et al., 2024).

C Additional Experimental Results

To further validate our findings at larger scale, we train 4-layer 4-head models with both softmax and ReLU attention. All models use the same training configuration described in Appendix A. Figures 5 and 6 show representative attention patterns. The softmax variant exhibits strong sink behavior in at least one head per 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.

D Proof of theorem 1

D.1 Proof Sketch

Proof sketch. Suppose for contradiction that $\alpha_{i,1} \le 1-\varepsilon$ at some non-trigger position i with probability at least $\delta > 0$, even as $\eta := \mathcal{L}(f) \to 0$. On this event a constant amount of attention mass falls on non-BOS tokens; by pigeonhole there exist indices i_0, h_0 and a constant $\gamma > 0$ such that $\alpha_{i_0,h_0} \ge \gamma$ on a positive-measure set.

Since every non-trigger position must output $\mathbf{0}$ with error at most η, and adding more keys can only decrease any fixed softmax weight, one can reduce to short prefixes and show that whenever $h \le i$ are both non-trigger positions, $\|\alpha_{i,h}\mathbf{V}\mathbf{x}^{(h)}\|_2 \le O(\eta)$. On the positive-measure set where $\alpha_{i,h} \ge \gamma$, this gives $\|\mathbf{V}\mathbf{x}^{(h)}\|_2 = O(\eta/\gamma)$: the value map must crush a positive-probability set of non-trigger tokens.

By bounded density and independence of the content coordinates, for every content coordinate $m \ge 4$ this crushed set contains two tokens $\mathbf{z}, \mathbf{z}'$ that agree on all coordinates except m, where they differ by at least a constant. Transplant them into two sequences with trigger at position 3: $(\texttt{BOS}, \mathbf{z}, \mathbf{t})$ and $(\texttt{BOS}, \mathbf{z}', \mathbf{t})$. The targets at the trigger position differ by $\frac{1}{2}(\mathbf{z} - \mathbf{z}')$, which has a $\Omega(1)$ component along e_m. The prediction at position 3 is $\hat{\mathbf{y}}^{(3)}(\mathbf{z}) = \alpha_{3,1}\mathbf{V}e_1 + \alpha_{3,2}\mathbf{V}\mathbf{z} + \alpha_{3,3}\mathbf{V}\mathbf{t}$; the first two terms are $O(\eta)$ by the crushing bound, and the third lies in the span of the fixed vector $\mathbf{v} := \mathbf{V}\mathbf{t}$. Projecting onto $\mathbf{v}^\perp$ removes the trigger contribution entirely, so the two projected predictions are $O(\eta)$-close, while the projected targets remain $\Omega(1)$-apart (choosing m so e_m has a nontrivial component in $\mathbf{v}^\perp$). This contradicts $\eta \to 0$. □

D.2 Full Proof

We prove theorem 1 by establishing two separate necessity results: one for pre-trigger positions (theorem 4) and one for post-trigger positions (theorem 5). Combining these two results directly yields the statement of theorem 1, which asserts necessity at all non-trigger positions $i \ne j$.

Theorem 4 (Pre-Trigger Necessity)**.** *For any $\varepsilon, \delta \in \mathbb{R}_{>0}, L \in \mathbb{N}_{\ge 4}, n \in \mathbb{N}_{\ge 5}$, and a bounded probability density function $\mathcal{P}$, there exists a constant $\eta \in \mathbb{R}_{>0}$ such that the following holds. Consider any single-layer softmax attention model f with loss $\mathcal{L}(f) \le \eta$ on sequences with length L and dimension n where non-trigger coordinates are drawn from $\mathcal{P}$. Then with probability at least $1-\delta$ over the choice of $\mathbf{x}$ with trigger position j, for all pre-trigger positions $1 < i < j$, we have $\alpha_{i,1} \ge 1-\varepsilon$.*

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 $\mathbf{V}$. Reparameterizing by setting

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

leaves $\alpha_{i,k}$ and $\hat{\mathbf{y}}^{(i)}$ unchanged, hence the loss is unchanged. Therefore, we will assume without loss

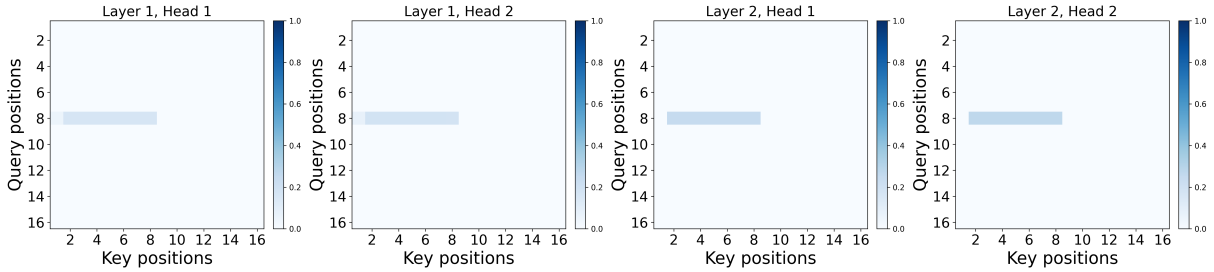

Figure 4: **ReLU attention: 2-layer 2-head model.** Attention patterns on a single test input (trigger at position 8). No sink formation occurs in any head; attention on BOS remains near zero throughout.

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 $\varepsilon_0, \delta_0 \in \mathbb{R}_{>0}$ and suppose by contradiction that 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 $(\mathbf{x}, j) \sim \mathcal{P}$ 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 position 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 $\gamma \in \mathbb{R}_{>0}$ (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 $\delta \in \mathbb{R}_{>0}$ 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 7.

Since the event in (2) has positive probability at least δ, by Lemma 7 (applied to content coordinates) there exists $\varepsilon' \in \mathbb{R}_{>0}$ (independent of t) such that for every content coordinate $m \in \{4, \ldots, n\}$ there exist tokens $x^{(m)}, y^{(m)}$ with the following properties: (i) $x_k^{(m)} = y_k^{(m)}$ for all $k \ne m$, and $\big|x_m^{(m)} - y_m^{(m)}\big| \ge \varepsilon'$; and (ii) there exist sequences with either $x^{(m)}$ or $y^{(m)}$ at position $h^\star$ and with trigger position j satisfying $i^\star < j$, such that

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

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

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

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

$$\big\|\mathbf{V}\mathbf{z}\big\|_2 \le \tfrac{4}{\gamma}\eta_t. \tag{4}$$

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 $m \in \{4, \ldots, n\}$ and let $\mathbf{x}_t := x^{(m)}$ and $\mathbf{y}_t := y^{(m)}$ be the two tokens from Step 2 satisfying $|\mathbf{x}_t^{(m)} - \mathbf{y}_t^{(m)}| \ge \varepsilon'$. Instantiate two sequences by setting the trigger at $j = 3$, taking $\mathbf{x}^{(2)} \in \{\mathbf{x}_t, \mathbf{y}_t\}$, and fixing the trigger token $\mathbf{x}^{(3)}$ to any arbitrary value $\mathbf{t}$ such that the sequence is in the support of $\mathcal{D}$. At position $i = 3$ the target is

$$\mathbf{y}^{(3)} = \frac{1}{2}(\mathbf{x}^{(2)} + \mathbf{t}). \tag{5}$$

For any $\mathbf{z} \in \{\mathbf{x}_t, \mathbf{y}_t\}$, let $\beta_t(\mathbf{z})$ be the attention weight $\alpha_{3,3}$ computed on the sequence where $\mathbf{x}^{(2)} = \mathbf{z}$ and $\mathbf{x}^{(3)} = \mathbf{t}$. Define the fixed value vector

$$\mathbf{v}_t \;:=\; \mathbf{V}_t\,\mathbf{t}. \tag{6}$$

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

$$\hat{\mathbf{y}}^{(3)}(\mathbf{z}) = \underbrace{\alpha_{3,1}\,\mathbf{V}e_1 + \alpha_{3,2}\,\mathbf{V}\mathbf{z}}_{=:\,\mathbf{r}_t(\mathbf{z})} + \beta_t(\mathbf{z})\,\mathbf{v}_t, \tag{7}$$

$$\|\mathbf{r}_t(\mathbf{z})\|_2 \le C_0\,\eta, \tag{8}$$

with $C_0 := 1 + \frac{4}{\gamma}$ independent of t. Consider coordinate 3 (the non-trigger non-BOS indicator).

Since $(\mathbf{y}^{(3)})_3 = \frac{1}{2}((\mathbf{x}^{(2)})_3 + (\mathbf{t})_3) = \frac{1}{2}(1+0) = 0.5$ and $0 < \beta_t(\mathbf{z}) \le 1$, from (7) and the uniform loss bound we obtain

$$\begin{aligned}
&\left|\beta_t(\mathbf{z})\,(\mathbf{v}_t)_3 - 0.5\right| \\
&\quad \le \left|\hat{\mathbf{y}}_3^{(3)}(\mathbf{z}) - 0.5\right| + \left|(\mathbf{r}_t(\mathbf{z}))_3\right| \\
&\quad \le \eta + C_0\eta \\
&\quad = C_1\eta, \qquad (9)
\end{aligned}$$

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

$$(\mathbf{v}_t)_3 \;\ge\; \frac{0.5 - C_1\eta}{\beta_t(\mathbf{z})} \;\ge\; 0.5 - C_1\eta \;>\; 0, \quad (10)$$

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 coordinate $m_0 \in \{4, 5\}$ such that $\|P_t e_{m_0}\|_2 \ge 1/\sqrt{2}$; fix m to be that coordinate. Now, applying P_t to (7) kills the $\mathbf{v}_t$ component:

$$P_t\hat{\mathbf{y}}^{(3)}(\mathbf{z}) = P_t\mathbf{r}_t(\mathbf{z}), \qquad (11)$$

$$\|P_t\hat{\mathbf{y}}^{(3)}(\mathbf{z})\|_2 \le \|\mathbf{r}_t(\mathbf{z})\|_2 \le C_0\eta. \qquad (12)$$

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

$$\begin{aligned}
&\left\|P_t\hat{\mathbf{y}}^{(3)}(\mathbf{x}_t) - P_t\hat{\mathbf{y}}^{(3)}(\mathbf{y}_t)\right\|_2 \\
&\quad \le\; \|P_t\mathbf{r}_t(\mathbf{x}_t)\|_2 + \|P_t\mathbf{r}_t(\mathbf{y}_t)\|_2 \\
&\quad \le\; 2C_0\eta. \qquad (13)
\end{aligned}$$

On the other hand, we have $\mathbf{y}^{(3)}(\mathbf{z}) = \frac{1}{2}(\mathbf{z} + \mathbf{t})$, so $P_t\mathbf{y}^{(3)}(\mathbf{z}) = \frac{1}{2}P_t\mathbf{z} + \frac{1}{2}P_t\mathbf{t}$. Since the $\mathbf{t}$ term is constant in $\mathbf{z}$, it cancels in the difference:

$$\begin{aligned}
&\left\|P_t\mathbf{y}^{(3)}(\mathbf{x}_t) - P_t\mathbf{y}^{(3)}(\mathbf{y}_t)\right\|_2 \\
&\quad = \frac{1}{2}\|P_t(\mathbf{x}_t - \mathbf{y}_t)\|_2 \\
&\quad = \frac{1}{2}\|P_t((\mathbf{x}_{t,m} - \mathbf{y}_{t,m})e_m)\|_2 \\
&\quad = \frac{1}{2}|\mathbf{x}_{t,m} - \mathbf{y}_{t,m}|\,\|P_t e_m\|_2 \\
&\quad \ge \frac{1}{2}\varepsilon'\,\|P_t e_m\|_2 \\
&\quad \ge \frac{1}{2\sqrt{2}}\varepsilon'. \qquad (14)
\end{aligned}$$

where the third equality uses 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,

$$\begin{aligned}
&\left\|P_t\mathbf{y}^{(3)}(\mathbf{x}_t) - P_t\mathbf{y}^{(3)}(\mathbf{y}_t)\right\|_2 \\
&\quad \le \left\|P_t\hat{\mathbf{y}}^{(3)}(\mathbf{x}_t) - P_t\hat{\mathbf{y}}^{(3)}(\mathbf{y}_t)\right\|_2 + 2\eta \\
&\quad \le (2C_0 + 2)\eta, \qquad (15)
\end{aligned}$$

which contradicts (14) for all sufficiently small η, because $\varepsilon'\|P_t e_m\|_2 > 0$ is independent of t.

□

Theorem 5 (Post-Trigger Necessity)**.** *For any $\varepsilon, \delta \in \mathbb{R}_{>0}, L \in \mathbb{N}_{\ge 4}$, $n \in \mathbb{N}_{\ge 5}$, and a bounded probability density function $\mathcal{P}$, there exists a constant $\eta \in \mathbb{R}_{>0}$ such that the following holds. Consider any single-layer softmax attention model f with loss $\mathcal{L}(f) \le \eta$ on sequences with length L and dimension n where non-trigger coordinates are drawn from $\mathcal{P}$. Then with probability at least $1 - \delta$ over the choice of $\mathbf{x}$ with trigger position j, for all post-trigger positions $j < i \le L$, we have $\alpha_{i,1} \ge 1 - \varepsilon$.*

Proof. **Step 1: The trigger receives arbitrarily small attention post-trigger.** Fix any trigger token $\mathbf{t}$ and any non-trigger token $\mathbf{z}$, and consider the length-3 prefix $(\texttt{BOS}, \mathbf{t}, \mathbf{z})$ (so the trigger position is $j = 2$ and position 3 is post-trigger). Let $\widetilde{\alpha}_{3,1}, \widetilde{\alpha}_{3,2}, \widetilde{\alpha}_{3,3}$ be the attention weights at position 3.

We first bound the self term $\widetilde{\alpha}_{3,3}\mathbf{V}\mathbf{z}$ using Lemma 4. Embed the pair $(\texttt{BOS}, \mathbf{z})$ as the first two tokens of any valid sequence from $\mathcal{D}$ whose trigger position satisfies $j \ge 3$ (so position 2 is pre-trigger and non-trigger). Applying Lemma 4 at $i = 2$ gives $\|\alpha_{2,2}\mathbf{V}\mathbf{z}\|_2 \le 2\eta$ for that sequence, and by Lemma 2 (adding the extra key $\mathbf{t}$ can only decrease the probability assigned to $\mathbf{z}$) we have $\widetilde{\alpha}_{3,3} \le \alpha_{2,2}$, hence $\|\widetilde{\alpha}_{3,3}\mathbf{V}\mathbf{z}\|_2 \le 2\eta$. Also, Lemma 1 gives $\|\widetilde{\alpha}_{3,1}\mathbf{V}e_1\|_2 \le \eta$. Since the target at position 3 is $\mathbf{0}$, we have $\|\hat{\mathbf{y}}^{(3)}\|_2 \le \eta$, and therefore

$$\begin{aligned}
\|\widetilde{\alpha}_{3,2}\mathbf{V}\mathbf{t}\|_2 \le \|\hat{\mathbf{y}}^{(3)}\|_2 + \|\widetilde{\alpha}_{3,1}\mathbf{V}e_1\|_2 \\
+ \|\widetilde{\alpha}_{3,3}\mathbf{V}\mathbf{z}\|_2 \le 4\eta.
\end{aligned}$$

Finally, Lemma 6 (applied to any valid sequence with trigger at position 2) gives $\|\mathbf{V}\mathbf{t}\|_2 \ge 1 - 2\eta$, so

$$\widetilde{\alpha}_{3,2} \le \frac{4\eta}{1 - 2\eta}. \qquad (16)$$

Now fix any valid sequence $\mathbf{x}$ with trigger position j and any $i > j$. By Lemma 3(2), $\alpha_{i,j} \le \widehat{\alpha}_{3,2}$ where $\widehat{\alpha}_{3,2}$ is the attention weight on the second token in the prefix $(\texttt{BOS}, \mathbf{x}^{(j)}, \mathbf{x}^{(i)})$, and applying (16) to that prefix yields

$$\alpha_{i,j} \le \frac{4\eta}{1 - 2\eta} \qquad \text{for all } i > j. \qquad (17)$$

Step 2 (contradiction via shifting the trigger). Fix $\varepsilon_0, \delta_0 \in \mathbb{R}_{>0}$ and suppose, for contradiction,

that the theorem is false. Then there exists a sequence of one-layer softmax models $\{f_t\}_{t\geq 1}$ with $\eta_t := \mathcal{L}(f_t) \to 0$ such that for every t,

$$\mathbb{P}_{(\mathbf{x},j)\sim\mathcal{D}}\Big(\exists\, i > j :\ \alpha^{(t)}_{i,1}(\mathbf{x}) \leq 1-\varepsilon_0\Big) \ \geq\ \delta_0. \tag{18}$$

By Step 1 (i.e., (17)), for all $\mathbf{x}$ in $\mathrm{support}(\mathcal{D})$ and all $i > j$,

$$\alpha^{(t)}_{i,j}(\mathbf{x}) \leq \frac{4\eta_t}{1-2\eta_t}.$$

Fix t large enough so that $\frac{4\eta_t}{1-2\eta_t} \leq \varepsilon_0/2$.

Let E_t be the event in (18). For each $(\mathbf{x}, j) \in E_t$ there exists $i(\mathbf{x}) > j$ with $\alpha^{(t)}_{i(\mathbf{x}),1}(\mathbf{x}) \leq 1-\varepsilon_0$. For that $i(\mathbf{x})$ we also have $\alpha^{(t)}_{i(\mathbf{x}),j}(\mathbf{x}) \leq \varepsilon_0/2$, hence

$$\begin{aligned}
&\sum_{k\leq i(\mathbf{x}),\ k\notin\{1,j\}} \alpha^{(t)}_{i(\mathbf{x}),k}(\mathbf{x}) \\
&\quad = 1 - \alpha^{(t)}_{i(\mathbf{x}),1}(\mathbf{x}) - \alpha^{(t)}_{i(\mathbf{x}),j}(\mathbf{x}) \\
&\quad \geq \varepsilon_0/2.
\end{aligned} \tag{19}$$

Now define the shift map Shift that moves the trigger token to the end: $\mathbf{x}' = \mathrm{Shift}(\mathbf{x}, j)$, where $\mathbf{x}'^{(k)} = \mathbf{x}^{(k)}$ for $1 \leq k < j$ (positions before the trigger are unchanged), $\mathbf{x}'^{(k)} = \mathbf{x}^{(k+1)}$ for $j \leq k \leq L-1$ (positions after the trigger shift left by one), and $\mathbf{x}'^{(L)} = \mathbf{x}^{(j)}$ (trigger moves to the end). Then $\mathbf{x}' \in \mathrm{support}(\mathcal{D})$ with trigger position L, and moreover, by the definition of the task (section 3.2), we have that the probability density of $\mathbf{x}'$ is the same as that of $\mathbf{x}$:

$$\mathcal{P}(\mathbf{x}) = \mathcal{P}(\mathbf{x}'). \tag{20}$$

Fix $(\mathbf{x}, j) \in E_t$. Applying Lemma 2 we get that removing the key j can only *increase* the attention weight of each remaining key (at position $i(\mathbf{x})$ in $\mathbf{x}$, the "candidate" key set is $\{1,\dots,i(\mathbf{x})\}$; at position $i(\mathbf{x})-1$ in $\mathbf{x}'$, the "candidate" key set is $\{1,\dots,i(\mathbf{x})\}\setminus\{j\}$, the same set with the trigger key removed.) Therefore,

$$\begin{aligned}
&\sum_{r\leq i(\mathbf{x})-1,\ r\neq 1} \alpha'^{(t)}_{i(\mathbf{x})-1,r}(\mathbf{x}') \\
&\quad \geq \sum_{k\leq i(\mathbf{x}),\ k\notin\{1,j\}} \alpha^{(t)}_{i(\mathbf{x}),k}(\mathbf{x}) \\
&\quad \geq\ \varepsilon_0/2,
\end{aligned}$$

where the last inequality is (19). Equivalently,

$$\alpha'^{(t)}_{i(\mathbf{x})-1,1}(\mathbf{x}') \ \leq\ 1-\varepsilon_0/2. \tag{21}$$

Since this holds for *every* $(\mathbf{x}, j) \in E_t$, by the Pigeonhole Principle, there exist fixed indices $j^* \in \{1,\dots,L\}$ and $r^* < L$ and a constant $c_1 \in \mathbb{R}_{>0}$ such that for infinitely many t,

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{x},j)\sim\mathcal{D}}\Big(\alpha'^{(t)}_{r^*,1}(\mathbf{x}') \leq 1-\varepsilon_0/2 \text{ and } j = j^*\Big) \\
&\qquad \geq\ c_1\delta_0,
\end{aligned} \tag{22}$$

where $\mathbf{x}' = \mathrm{Shift}(\mathbf{x}, j)$.

Finally, consider the bijection $\mathbf{x} \mapsto \mathrm{Shift}(\mathbf{x}, j^*)$ from the set of sequences with trigger at j^* to the set of sequences with trigger at L. By eq. (20), this map preserves probability density. Thus, the event in (22) has the exact same probability as the corresponding event for sequences with trigger at L:

$$\begin{aligned}
&\mathbb{P}_{(\mathbf{z},j)\sim\mathcal{D}}\Big(\alpha'^{(t)}_{r^*,1}(\mathbf{z}) \leq 1-\varepsilon_0/2 \text{ and } j = L\Big) \\
&\quad \geq\ c_1\delta_0.
\end{aligned}$$

Conditioning on $j = L$, this implies that for infinitely many t,

$$\mathbb{P}\Big(\alpha'^{(t)}_{r^*,1}(\mathbf{z}) \leq 1-\varepsilon_0/2 \ \Big|\ j = L\Big) \geq \frac{c_1\delta_0}{\mathbb{P}(j=L)}.$$

Since $r^* < L$, this contradicts Theorem 4, as needed.

□

E Proof of theorem 2

E.1 Proof Sketch

Proof sketch. We unroll the multi-layer network and apply similar reasoning as in theorem 1: if no layer exhibits sink behavior, the effective attention weights on content tokens remain large, forcing the value map to crush them to zero, which again contradicts the sensitivity required at the trigger position. □

E.2 Full Proof

Step 1: Setup and contradiction assumption. Fix $\varepsilon_0, \delta_0 \in \mathbb{R}_{>0}$. 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,

$$\begin{aligned}
&\mathbb{P}\Big(\forall d \in \{1,\dots,D\},\ \forall\, 1 < i < j :\\
&\qquad \alpha^{(d)}_{i,1} \leq 1-\varepsilon_0 \Big| j \geq 3\Big) \ \geq\ \delta_0.
\end{aligned} \tag{23}$$

Let E_t denote the event inside the probability in (23) intersected with the event $j \geq 3$. For each t, let $\mathbf{V}_t$ be the combined value map from Lemma 8, and write $\beta_{i,k}^{(t)}(\cdot)$ for the corresponding coefficients.

Step 2: No sink implies small value projections. On the event E_t, position 2 is pre-trigger (since $j \geq 3$) and for every layer d,

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

Therefore, by Lemma 9 conditioned on E_t we have that

$$\beta_{2,2}^{(t)}(\mathbf{x}) \;\geq\; \varepsilon_0^D \tag{24}$$

Moreover, Lemma 11 applied to f_t yields

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

Combining with (24) gives

$$\|\mathbf{V}_t\mathbf{x}^{(2)}\|_2 \;\leq\; \frac{2}{\varepsilon_0^D}\,\eta_t \qquad \text{on } E_t. \tag{25}$$

Define the measurable set

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

Since $E_t \subseteq \{\mathbf{x}^{(2)} \in S_t\}$ by (25), (23) implies

$$\mathbb{P}\big(\mathbf{x}^{(2)} \in S_t \big| j \geq 3\big) \;\geq\; \delta_0. \tag{26}$$

By Lemma 7 (applied to content coordinates) and (26), there exists $\varepsilon' \in \mathbb{R}_{>0}$ (independent of t) such that for every content coordinate $m \in \{4, \ldots, n\}$ there exist tokens $\mathbf{x}_t^{(m)}, \mathbf{y}_t^{(m)} \in S_t$ satisfying

$$\begin{aligned} \mathbf{x}_{t,k}^{(m)} &= \mathbf{y}_{t,k}^{(m)} \text{ for all } k \neq m, \\ \left|\mathbf{x}_{t,m}^{(m)} - \mathbf{y}_{t,m}^{(m)}\right| &\geq \varepsilon'. \end{aligned} \tag{27}$$

Step 3: Transplanting to $j = 3$ and deriving a contradiction. Fix t and abbreviate $\eta := \eta_t$. Pick a content coordinate $m \in \{4, \ldots, n\}$ and let $\mathbf{x}_t := \mathbf{x}_t^{(m)}$ and $\mathbf{y}_t := \mathbf{y}_t^{(m)}$ 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 $j = 3$, taking $\mathbf{x}^{(2)} \in \{\mathbf{x}_t, \mathbf{y}_t\}$, and fixing the trigger token $\mathbf{x}^{(3)}$ to an arbitrary value $\mathbf{t}$ such that the sequence is in the support of $\mathcal{D}$. At position $i = 3$ the target is

$$\mathbf{y}^{(3)} \;=\; \frac{1}{2}(\mathbf{x}^{(2)} + \mathbf{t}). \tag{28}$$

For any $\mathbf{z} \in \{\mathbf{x}_t, \mathbf{y}_t\}$, let $\beta_t(\mathbf{z})$ be the coefficient $\beta_{3,3}^{(t)}(\mathbf{z})$ computed on the sequence where $\mathbf{x}^{(2)} = \mathbf{z}$ and $\mathbf{x}^{(3)} = \mathbf{t}$. Define the fixed value vector

$$\mathbf{v}_t \;:=\; \mathbf{V}_t\,\mathbf{t}. \tag{29}$$

By Lemma 8, for each choice $\mathbf{x}^{(2)} = \mathbf{z}$ we can decompose

$$\hat{\mathbf{y}}^{(3)}(\mathbf{z}) = \underbrace{\beta_{3,1}^{(t)}(\mathbf{z})\,\mathbf{V}_t e_1 + \beta_{3,2}^{(t)}(\mathbf{z})\,\mathbf{V}_t\mathbf{z}}_{=:\,\mathbf{r}_t(\mathbf{z})} + \beta_t(\mathbf{z})\,\mathbf{v}_t. \tag{30}$$

Since $\beta_{3,1}^{(t)}(\mathbf{z}), \beta_{3,2}^{(t)}(\mathbf{z}) \leq 1$, Lemma 10 gives $\|\mathbf{V}_t e_1\|_2 \leq \eta$, and $\mathbf{z} \in S_t$ implies $\|\mathbf{V}_t\mathbf{z}\|_2 \leq \frac{2}{\varepsilon_0^D}\eta$. Therefore

$$\|\mathbf{r}_t(\mathbf{z})\|_2 \;\leq\; C_0\,\eta, \qquad C_0 \;:=\; 1 + \tfrac{2}{\varepsilon_0^D}. \tag{31}$$

Consider coordinate 3 (the non-trigger non-BOS indicator). For the $j = 3$ construction, we have $(\mathbf{y}^{(3)})_3 = 0.5$. Using (30) and the uniform loss bound,

$$\begin{aligned} &\left|\beta_t(\mathbf{z})\,(\mathbf{v}_t)_3 - 0.5\right| \\ &\quad\leq \left|\hat{\mathbf{y}}_3^{(3)}(\mathbf{z}) - 0.5\right| + \left|(\mathbf{r}_t(\mathbf{z}))_3\right| \\ &\quad\leq \eta + C_0\eta \\ &\quad= C_1\eta, \end{aligned}$$

where $C_1 := 1 + C_0$. Hence $(\mathbf{v}_t)_3 \geq 0.5 - C_1\eta > 0$ for all sufficiently large t, so $\mathbf{v}_t \neq \mathbf{0}$.

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 $m_0 \in \{4, 5\}$ such that

$$\|P_t e_{m_0}\|_2 \;\geq\; 1/\sqrt{2}. \tag{32}$$

Fix such an m, and take $\mathbf{x}_t := \mathbf{x}_t^{(m)}$ and $\mathbf{y}_t := \mathbf{y}_t^{(m)}$ from (27).

Applying P_t to (30) kills the $\mathbf{v}_t$ component, giving $P_t\hat{\mathbf{y}}^{(3)}(\mathbf{z}) = P_t\mathbf{r}_t(\mathbf{z})$. Therefore,

$$\begin{aligned} &\left\|P_t\hat{\mathbf{y}}^{(3)}(\mathbf{x}_t) - P_t\hat{\mathbf{y}}^{(3)}(\mathbf{y}_t)\right\|_2 \\ &\qquad\leq \|P_t\mathbf{r}_t(\mathbf{x}_t)\|_2 + \|P_t\mathbf{r}_t(\mathbf{y}_t)\|_2 \leq 2C_0\eta, \end{aligned} \tag{33}$$

using (31). On the other hand, by (28) we have $P_t\mathbf{y}^{(3)}(\mathbf{z}) = \frac{1}{2}P_t(\mathbf{z} + \mathbf{t})$, so

$$\begin{aligned} &\left\|P_t\mathbf{y}^{(3)}(\mathbf{x}_t) - P_t\mathbf{y}^{(3)}(\mathbf{y}_t)\right\|_2 \\ &\qquad= \frac{1}{2}\|P_t(\mathbf{x}_t - \mathbf{y}_t)\|_2 \\ &\qquad= \frac{1}{2}|\mathbf{x}_{t,m} - \mathbf{y}_{t,m}| \cdot \|P_t e_m\|_2 \\ &\qquad\geq \frac{1}{2\sqrt{2}}\varepsilon', \end{aligned} \tag{34}$$

using (27) and (32).

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 (34) for all sufficiently small η. This contradiction completes the proof.

F Proof of theorem 3

F.1 Proof Sketch

Proof sketch. We provide a simple explicit construction. By choosing query and key weights to align with the trigger indicator coordinate and non-trigger non-BOS indicator coordinate, we ensure that attention scores are equal to some positive constant at the trigger position (where they compute the average) and zero otherwise. Since ReLU does not enforce normalization, the model can output the zero vector by simply having zero attention weights, without needing a sink. □

F.2 Full Proof

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 from section 3.2: coordinate 1 is the BOS indicator; coordinate 2 is the trigger indicator; coordinate 3 is the non-trigger non-BOS indicator, with $\mathbf{x}_3^{(1)} = \mathbf{x}_3^{(j)} = 0$ and $\mathbf{x}_3^{(i)} = 1$ for $i \neq 1, j$. Define

$$\mathbf{W}_Q \;=\; e_2(e_2+e_3)^\top.$$

Computing the attention weights. Using the ReLU attention formula from section 3.4, the unnormalized score from position i to position k is

$$\mathbf{x}^{(i)}\mathbf{W}_Q\mathbf{W}_K^\top(\mathbf{x}^{(k)})^\top \;=\; \mathbf{x}_2^{(i)}\cdot(\mathbf{x}_2^{(k)}+\mathbf{x}_3^{(k)}).$$

Fix a trigger position $j \in \{2,\dots,L\}$. For any non-trigger position $i \neq j$, we have $\mathbf{x}_2^{(i)} = 0$, so all scores are zero and hence $\alpha_{i,k} = \mathrm{ReLU}(0)/n_i = 0$ for all $k \le i$. In particular, $\alpha_{i,1} = 0$.

For the trigger position $i = j$, we have $\mathbf{x}_2^{(j)} = 1$. The score to position k equals $\mathbf{x}_2^{(k)} + \mathbf{x}_3^{(k)}$. This is 1 for non-trigger non-BOS tokens (if such exist) $k \in \{2,\dots,j-1\}$ (where $\mathbf{x}_3^{(k)} = 1$) and for the trigger token $k = j$ (where $\mathbf{x}_2^{(k)} = 1$). It is 0 for $k = 1$ (BOS). After applying ReLU and dividing by $n_j = j-1$, we obtain

$$\begin{aligned}\alpha_{j,k} &= \tfrac{1}{j-1} \quad \text{for } 2 \le k \le j,\\ \alpha_{j,k} &= 0 \qquad \text{otherwise.}\end{aligned}$$

Verifying the output. At non-trigger positions, all attention weights are zero, so $f(\mathbf{x})^{(i)} = \mathbf{0} = \mathbf{y}^{(i)}$. At the trigger position $i = j$, using $\mathbf{W}_O = \mathbf{W}_V = \mathbf{I}$:

$$\begin{aligned}f(\mathbf{x})^{(j)} &= \mathbf{W}_O\sum_{k=1}^{j}\alpha_{j,k}\,\mathbf{W}_V\mathbf{x}^{(k)}\\ &= \frac{1}{j-1}\sum_{k=2}^{j}\mathbf{x}^{(k)} = \overline{\mathbf{x}} = \mathbf{y}^{(j)}.\end{aligned}$$

Thus $\mathcal{L}(f) = 0$ and $\alpha_{i,1} = 0$ for all i, completing the proof.

G Lemmas

Lemma 1. *Let f be a single-layer softmax 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*

$$\begin{aligned}\sigma_S(\mathbf{k}) &= \frac{\exp(\mathbf{q}^\top\mathbf{k})}{\sum_{\mathbf{r}\in S}\exp(\mathbf{q}^\top\mathbf{r})},\\ \sigma_T(\mathbf{k}) &= \frac{\exp(\mathbf{q}^\top\mathbf{k})}{\sum_{\mathbf{r}\in T}\exp(\mathbf{q}^\top\mathbf{r})},\end{aligned}$$

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 $\mathcal{D}$ (section 3.2) and any indices $1 < i$ and $1 < i < h$. 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 position $1 < i \ne j$,*

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

Proof. Fix i and consider the length-2 prefix $(\mathtt{BOS}, \mathbf{x}^{(i)})$. 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 $1 < i < h$ with $i, h \ne j$:*

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

Proof. Consider first the length-3 prefix $(\mathtt{BOS}, \mathbf{x}^{(i)}, \mathbf{x}^{(h)})$. 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. *In the setting of lemma 1, assume the attention mechanism is softmax. For every sequence in $\mathrm{support}(\mathcal{D})$ with trigger at position j,*

$$\|\mathbf{V}\mathbf{x}^{(j)}\|_2 \ge 1 - 2\eta.$$

Proof. Consider a sequence where the trigger is at position $j = 2$. The target output at position 2 is $\mathbf{y}^{(2)} = \mathbf{x}^{(2)}$. The model output is

$$\hat{\mathbf{y}}^{(2)} = \alpha_{2,1}\mathbf{V}e_1 + \alpha_{2,2}\mathbf{V}\mathbf{x}^{(2)}.$$

We know $\|\mathbf{y}^{(2)} - \hat{\mathbf{y}}^{(2)}\|_2 \le \eta$ and $\|\mathbf{V}e_1\|_2 \le \eta$ (lemma 1). By triangle inequality, $\|\mathbf{y}^{(2)} - \alpha_{2,2}\mathbf{V}\mathbf{x}^{(2)}\|_2 \le 2\eta$. Since $(\mathbf{y}^{(2)})_2 = 1$ (trigger indicator), we have $|1 - \alpha_{2,2}(\mathbf{V}\mathbf{x}^{(2)})_2| \le 2\eta$. Since $\alpha_{2,2} \le 1$, this implies $(\mathbf{V}\mathbf{x}^{(2)})_2 \ge 1 - 2\eta$, so $\|\mathbf{V}\mathbf{x}^{(2)}\|_2 \ge 1 - 2\eta$. □

Lemma 7. *Let $n \in \mathbb{N}_{\ge 1}$ and $X = (X_1, \dots, X_n) \sim \mu^{\otimes n}$, where μ has a density g bounded by $M := \sup_{x\in\mathbb{R}} g(x) < \infty$. Fix $\delta \in (0, 1]$. Then there exists some $\varepsilon' \in \mathbb{R}_{>0}$ such that if a measurable set $E \subset \mathbb{R}^n$ satisfies $\mathbb{P}(X \in E) \ge \delta$, then for every coordinate $j \in \{1, \dots, n\}$ 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 $E_j(z) := \{t \in \mathbb{R} : (z_1, \dots, z_{j-1}, t, z_{j+1}, \dots) \in E\}$. 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 the 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 $\eta \in \mathbb{R}_{>0}$, so we can choose $t_1, t_2 \in E_j(z)$ with $|t_1 - t_2| \geq \delta/M - \eta$ with $\eta < \delta/2M$. Setting $\varepsilon' = \delta/2M$ and taking x, y to match z on all coordinates $k \neq j$ and have j-th coordinates t_1, t_2 respectively gives the claim. □

Lemma 8. *Let $f = f^{(D)} \circ \cdots \circ f^{(1)}$ be a D-layer causal softmax 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

$$f(\mathbf{x})^{(i)} = \sum_{k=1}^{i} \beta_{i,k}(\mathbf{x}) \, \mathbf{V}\mathbf{x}^{(k)}. \tag{35}$$

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

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

Proof. Let $\mathbf{z}^{(0)} := \mathbf{x}$ and for $d \geq 1$ 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 $k \leq i$. By definition of a single layer,

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

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

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

A direct induction on d gives

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

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 (35). □

Lemma 9. *In the setting of Lemma 8, 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 8, note that position 1 is causal and thus never depends on token 2, directly yielding the product formula. □

Lemma 10. *In the setting of Lemma 8, if the loss $\mathcal{L}(f)$ (see section 3.2) satisfies $\mathcal{L}(f) \leq \eta$ then*

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

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 $\|f(\mathbf{x})^{(1)} - \mathbf{y}^{(1)}\|_2 \leq \mathcal{L}(f) \leq \eta$, the claim follows. □

Lemma 11. *In the setting of Lemma 8, assume softmax attention and that the loss $\mathcal{L}(f)$ (see section 3.2) satisfies $\mathcal{L}(f) \leq \eta$. Then for every $\mathbf{x}$ in $\mathrm{support}(\mathcal{D})$ with trigger position $j \geq 3$ we have that*

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

Proof. Since $j \geq 3$, position 2 is pre-trigger and the target satisfies $\mathbf{y}^{(2)} = \mathbf{0}$. By Lemma 8 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

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

using $\|f(\mathbf{x})^{(2)} - \mathbf{y}^{(2)}\|_2 \leq \eta$, $\beta_{2,1}(\mathbf{x}) \leq 1$, and Lemma 10. □

H Related Work

Theory and analyses of attention sinks. Several recent works study attention sinks directly, aiming to characterize why they arise and what they correlate with. Barbero et al. (2025) argue (theoretically and empirically) that first-token sinks can act as a stabilizing mechanism against over-mixing, and analyze how factors like depth, context length, and packing influence sink strength. Cancedda (2024) connect sink behavior to spectral structure in the vocabulary embedding/unembedding operators, attributing sinking to "dark" (tail-spectrum) components. Ruscio et al. (2025) view sinks as learned "reference-frame anchors" in representation space and show that the resulting anchoring pattern depends strongly on architectural choices, especially the positional encoding. de Llano et al.

(2026) connect attention sinks to "compression valleys" (layers where token representations become unusually low-entropy/compressed), showing both tend to emerge when the BOS token develops extremely large residual-stream activations. Qiu et al. (2026) study attention sinks together with "residual sinks" (persistent large activations in a few residual-stream dimensions) and argue these outliers interact with normalization (softmax/RMSNorm) to rescale the remaining components, supporting stable training. Sok et al. (2026) treat strong BOS-focused heads—especially in later layers—as a marker of functional redundancy and propose a pruning criterion based on sink scores. Hong and Lee (2025) attribute softmax-driven attention entropy collapse (attention concentrating onto a single token) to variance sensitivity of the logits and propose entropy-stable alternatives. Zhang et al. (2025) link sink tokens to large-norm outlier directions in LLM representations and RoPE-focused analyses similarly tie sink behavior to structured frequency artifacts and Q/K "massive values" (Jin et al., 2025; Xiong et al., 2026). These "massive values" were recently revisited in Sun et al. (2026), which argues that massive activations and attention sinks are largely decoupled: spikes can be suppressed via normalization changes while sinks persist. We complement these with a different angle: rather than studying how sinks emerge during training, we ask whether they are structurally *necessary* for certain computations. We prove that any softmax attention model solving a natural trigger-conditional task must develop a sink, regardless of the training procedure or optimization dynamics (theorems 1 and 2).

Softmax normalization implications. In standard attention, the softmax turns scores into non-negative weights that sum to one. Richter and Wattenhofer (2020) analyze how this simplex constraint can restrict attention behavior and discuss alternatives that relax or replace softmax normalization. Veličković et al. (2025) prove that softmax-based mechanisms can fail to maintain increasingly sharp selection as the problem size grows, leading to degraded behavior under distribution shift when near-argmax behavior is required. We provide a concrete natural task where this constraint is provably the cause of sink formation: a model that must aggregate context on a trigger token and output zero otherwise cannot avoid a sink under softmax normalization (theorem 1), whereas ReLU attention—which lacks the simplex constraint—solves the same task without any sink (theorem 3).

Mitigating sinks. Alongside analyses, multiple papers propose sink-targeted interventions. This includes modified attention normalizations explicitly designed to avoid sinks (Zuhri et al., 2026; Huang et al., 2026), as well as training procedures tailored to long-context regimes, including sliding-window attention that explicitly addresses attention-sink issue(Fu et al., 2025). For inference-time efficiency, Su and Yuan (2025); Hosseini et al. (2026) analyze how KV-cache quantization can disrupt sink behavior and propose predicting and preserving sink tokens during quantization. Mitigation has also been studied for closely related collapse modes of attention: Hong and Lee (2025) analyze softmax-driven entropy collapse (attention concentrating onto a single token) and propose alternatives aimed at stabilizing attention entropy, while Hankemeier and Schilling (2026) study diagonal/temporal self-attention sinks and introduce regularizers to counter them. In a different setting, Lin et al. (2025) show that attention sinks degrade training-free conversion of decoder-only LLMs into text encoders, and reduce this effect by enabling bidirectional attention and masking the first token in attention. In multimodal and AV settings, sink patterns have similarly motivated mitigation strategies aimed at reducing hallucination and stabilizing activations (Zhang et al., 2024; Anand et al., 2026). Lu et al. (2025) analyze attention sinks as a structured artifact in Vision Transformers and leverage this structure to derive efficient approximation schemes. Moreover, in these settings, sinks have been explicitly regularized in the context of harmful fine-tuning (Liu et al., 2026). Sinks have also been studied in alignment and security contexts where Shang et al. (2025) leverage sink behavior as a pathway for backdooring unlearning procedures. Finally, circuit-level interventions have also been explored in regimes where sink-related circuitry correlates with repeated-token failures (Yona et al., 2025). Our necessity results offer a principled lens for evaluating such interventions: for trigger-conditional circuits, the sink is the mechanism enabling the computation, so strategies that operate *within* softmax (penalizing BOS attention, spreading mass, post-hoc reweighting) risk degrading the circuit without addressing the root cause. The contrast with ReLU attention (theorem 3 and section 5) suggests that relaxing the normalization constraint is the more fundamental direction.

Usefulness of sinks. Other work treats sinks as a useful computational primitive rather than an artifact to eliminate. Our work formalizes this intuition: for trigger-conditional behaviors—where a

model must aggregate context on a trigger while outputting zero elsewhere—the sink is not merely a convenient implementation choice but a *provably necessary* consequence of softmax normalization (theorems 1 and 2). Zhang et al. (2025) link sink tokens to representation outliers and argue that simple structural conditions (e.g., low-rank attention structure) can be sufficient to induce sinks that support concrete computations such as averaging and retrieval—a viewpoint that is closely aligned with our trigger-conditional setting. Sinks have been argued to induce or support attention-layer specialization, including MoE-like effects within attention (Fu et al., 2026). Sandoval-Segura et al. (2025) use sink dominance to identify "dormant" heads and validate their redundancy via head ablations. In addition, BOS-sink heads have been treated as a locus of redundancy that can be targeted for model simplification via sink-aware pruning (Sok et al., 2026). In large vision-language models (Luo et al., 2025) show that high-norm ViT sink tokens encode high-level semantic concepts and serve as important visual information pathways into the LLM, and propose methods to better leverage them. Related ideas appear in diffusion LMs as well, where introducing an explicit sink token is used to stabilize sink behavior across steps (Zhang et al., 2026) and where sink locations can be transient across denoising steps, motivating sink-aware pruning that targets unstable sinks (Myrzakhan et al., 2026).

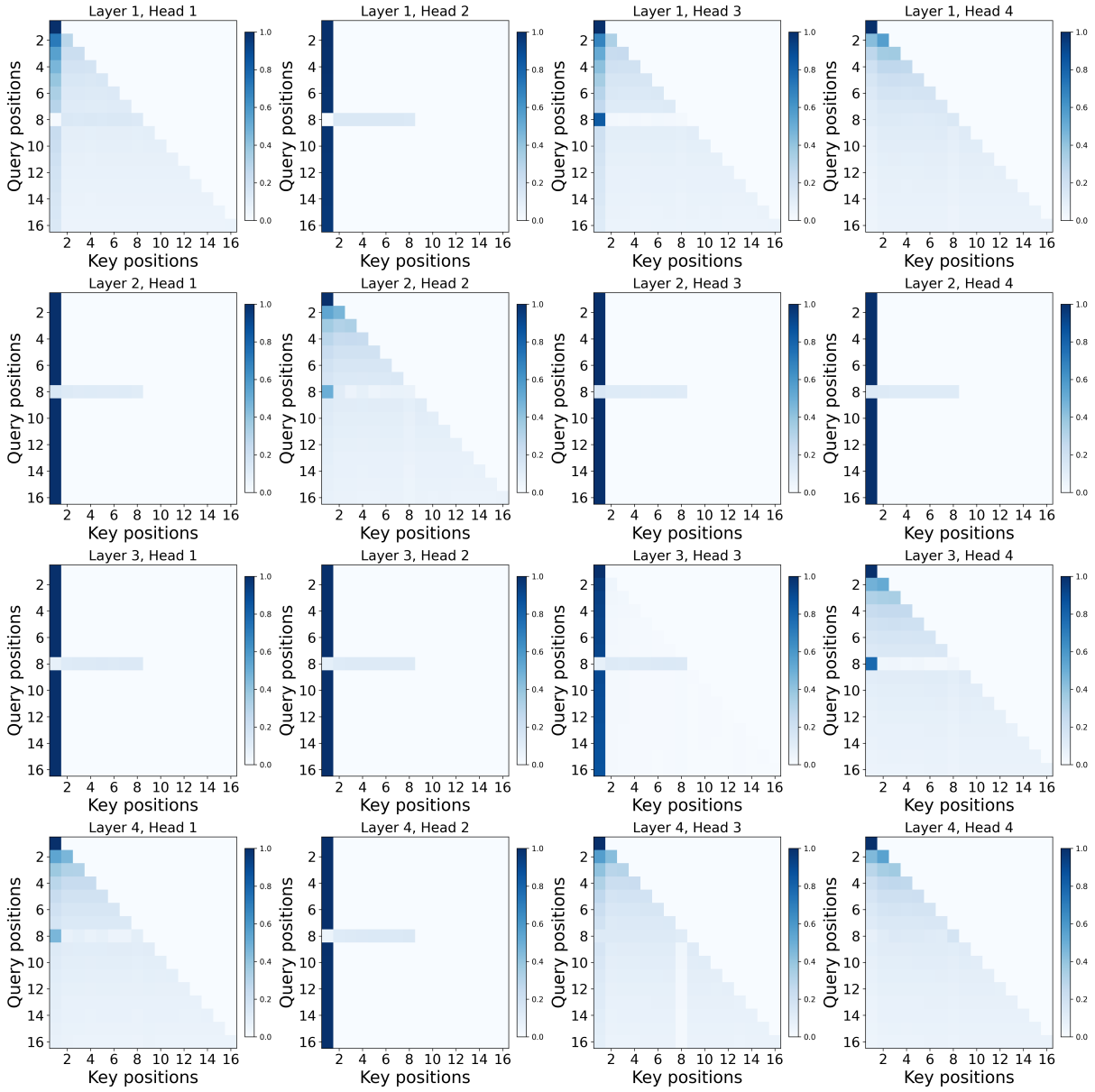

Figure 5: **Softmax attention: 4-layer 4-head model.** Representative attention patterns on a single test input showing strong sink at least in one head across all layers.

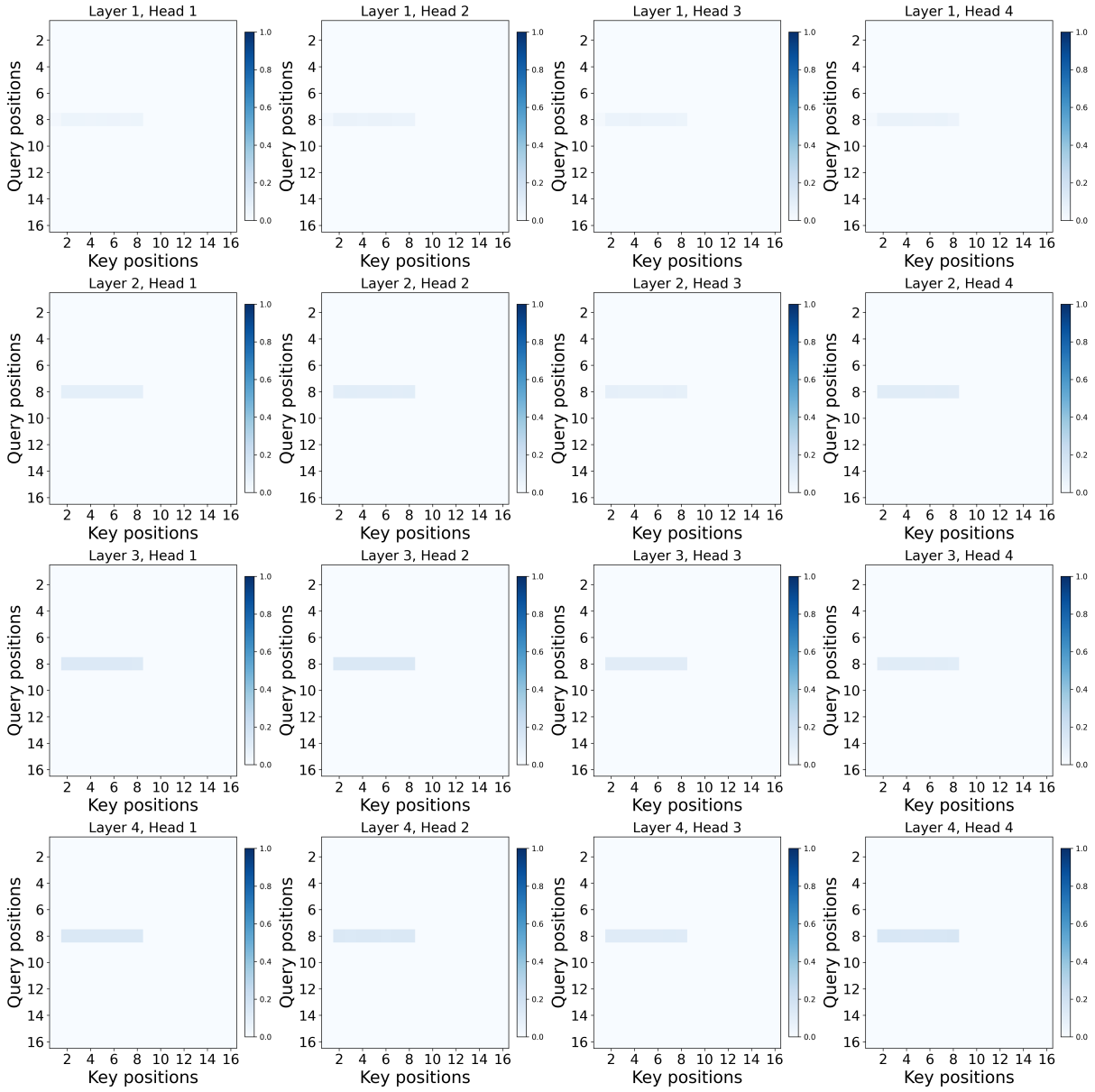

Figure 6: **ReLU attention: 4-layer 4-head model.** Representative attention patterns on a single test input showing absence of sink behavior across all layers.