# It's Not RoPE that Creates Sinks: The Role of Self-Concentration and Value-Non-Mixing in Attention

**Raito Kiya[1] Satoki Ohashi[1] Kosuke Sato[1] Go Kamoda[2,3] Ryosuke Takahashi[1,4,5] Yuji Yamamoto[2,3] Daiki Shiono[1] Keisuke Sakaguchi[1,5] Goro Kobayashi[6,1]**

[1]Tohoku University [2]SOKENDAI [3]NINJAL
[4]MBZUAI [5]RIKEN [6]Preferred Networks, Inc.

raito.kiya@dc.tohoku.ac.jp

## Abstract

Large Language Models (LLMs) often exhibit "Attention Sink" (AS) and the accompanying "Massive Activations" (MAs) at the initial position of a sequence. These phenomena frequently co-occur, and MAs can pose challenges for low-bit quantization. In this study, we analyze the factors underlying AS and MAs that emerge at the initial position regardless of the token occupying it. Our experiments suggest that self-concentration of attention, resulting from the causal mask, and the subsequent Value-non-mixing in attention outputs contribute to AS and MAs. These findings provide new empirical evidence on the internal dynamics of LLMs, offering insights that may inform future quantization strategies and advance our understanding of the internal mechanisms of attention layers.[1]

## 1 Introduction

Large Language Models (LLMs) exhibit a phenomenon known as the "Attention Sink" (AS), where attention concentrates on specific tokens, particularly the first token of a sequence (Xiao et al., 2024; Gu et al., 2025). These same tokens also often exhibit a few extremely large hidden-state activation values in intermediate layers, termed "Massive Activations" (MAs) (Sun et al., 2024). Large-magnitude activations of this kind are known to pose a major challenge for low-precision quantization (Dettmers et al., 2022).

Previous studies have offered various accounts of these phenomena, including their role in enabling no-op behavior under Softmax normalization constraints (Bondarenko et al., 2023), the interpretation of the first token as a key bias (Gu et al., 2025), and the influence of causal-mask asymmetry (Wu et al., 2025).[2] Despite these different perspectives, how the properties specific to the initial position contribute to the emergence of AS and MAs remains unclear.

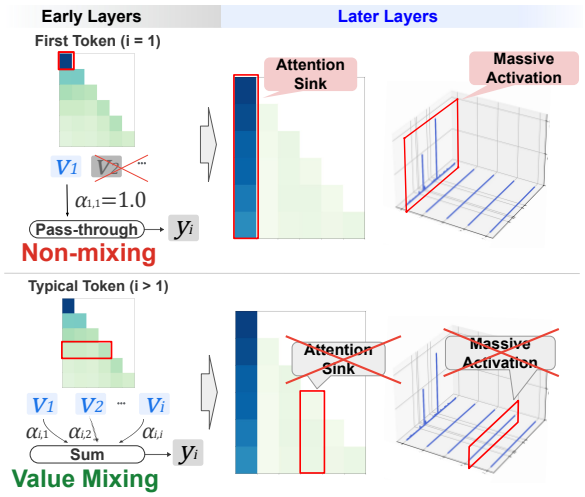



Figure 1: Main finding of this study. While attention outputs for standard tokens ($i > 1$) are computed based on weighted averages (i.e., "mixing") of Value vectors, the output for the initial token is based solely on its own "unmixed" Value vector. This "Value-non-mixing" state in early layers is a contributing factor in AS and MAs at the initial position of a sequence, regardless of the token occupying it.

To address this question, we decompose the properties of the initial position into three components: the typical presence of the BOS token, the first positional encoding, and forced self-concentration under causal masking. We then disentangle their respective influences on AS and MAs at the initial position of a sequence. Our experiments provide evidence that self-concentration contributes to the emergence of these phenomena. Under self-concentration, the attention output carries a single Value vector rather than a mixture of distinct ones, a state we call "Value-non-mixing." We argue that this state contributes to AS and MAs at the initial position regardless of the token occupying it

[1] github.com/kiya-raito/value-non-mixing
[2] Refer to Appendix A for more related works.



*Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, pages 31160–31184
October 24-29, 2026 ©2026 Association for Computational Linguistics

(Figure 1). Our findings provide new empirical evidence on how AS and MAs form in LLMs, which may inform future quantization strategies and advance our understanding of the internal mechanisms of attention layers.

## 2 Background: Attention Mechanism and Attention Sink Metric

An attention mechanism computes the output $\boldsymbol{y}_i \in \mathbb{R}^d$ for the $i$-th token from an input sequence $\{\boldsymbol{x}_j\}_{j=1}^T \in \mathbb{R}^{T\times d}$ as follows:

$$\{\boldsymbol{q}_i, \boldsymbol{k}_i, \boldsymbol{v}_i\} = \boldsymbol{x}_i\{\boldsymbol{W}_Q\boldsymbol{R}_{\Theta,i}^\top, \boldsymbol{W}_K\boldsymbol{R}_{\Theta,i}^\top, \boldsymbol{W}_V\}$$

$$\alpha_{i,j} = \frac{\exp\left(\boldsymbol{q}_i\boldsymbol{k}_j^\top/\sqrt{d'}\right)}{\sum_{m=1}^{i}\exp\left(\boldsymbol{q}_i\boldsymbol{k}_m^\top/\sqrt{d'}\right)}$$

$$\boldsymbol{y}_i = \sum_{j=1}^{i}\alpha_{i,j}\boldsymbol{v}_j\boldsymbol{W}_O, \tag{1}$$

where $\boldsymbol{W}_{Q,K,V} \in \mathbb{R}^{d\times d'}$ and $\boldsymbol{W}_O \in \mathbb{R}^{d'\times d}$ denote projection matrices for Query/Key/Value/Output, and $\boldsymbol{R}_{\Theta,i}$ denotes the RoPE (Su et al., 2024) rotation matrix.

AS refers to the phenomenon where the token at position $j$ (often the first) receives disproportionately large attention weights $\alpha_{i,j}$ across many heads (Xiao et al., 2024). Following Gu et al. (2025), we quantify the strength of AS at position $j$ using the $\text{Sink}_j^\epsilon$ metric:

$$\bar{\alpha}_j^{l,h} = \frac{1}{T-j+1}\sum_{i=j}^{T}\alpha_{i,j}^{l,h}$$

$$\text{Sink}_j^\epsilon = \frac{1}{L\cdot H}\sum_{l=1}^{L}\sum_{h=1}^{H}\mathbb{I}(\bar{\alpha}_j^{l,h} > \epsilon), \tag{2}$$

where $L$ is the number of layers, $H$ is the number of heads at each layer, $\epsilon$ is a threshold[3], and $\mathbb{I}(\cdot)$ is the indicator function.

## 3 Preliminaries: Influence of BOS Token

Before our main analysis, we examine whether AS and MAs observed at the initial position of a sequence are attributable to the BOS embedding or to the positional characteristics of the initial position. To disentangle these two factors, we swap the BOS token with the token at a non-initial position ($t = 16$) in each sequence. Our experiments use WikiText (Merity et al., 2017)[4] and, among the models listed in Table 4, the three whose tokenizer prepends a BOS token, since this swap is undefined for the other two.

| Model | Position | Vanilla | BOS at 16th |
|---|---|---|---|
| Llama-2-7b-hf | **1** | 0.9321 | 0.9178 |
| | **16** | 0.0001 | **0.7597** |
| Llama-3.2-3B | **1** | 0.9879 | 0.8856 |
| | **16** | 0.0000 | **0.4312** |
| Mistral-7B-v0.3 | **1** | 0.9833 | 0.1675 |
| | **16** | 0.0000 | **0.6376** |

Table 1: $\text{Sink}_j^\epsilon$ when the BOS token is moved to the 16th position. Unless otherwise noted, all subsequent tables in this analysis use $T = 64$ over $N = 100$ WikiText sequences.

| Model | Vanilla | RoPE Interv |
|---|---|---|
| Llama-2-7b-hf | 0.9305 | 0.9273 |
| Llama-3.2-3B | 0.9218 | 0.9133 |
| Mistral-7B-v0.3 | 0.0343 | 0.1525 |
| Qwen2-7B | 0.8173 | 0.7931 |
| pythia-1b | 0.7337 | 0.7547 |

Table 2: $\text{Sink}_1^\epsilon$ with and without RoPE intervention (both w/o BOS).

Table 1 reports the mean $\text{Sink}_j^\epsilon$ under this setting. The results show that AS still occurs at the BOS token even when the token is relocated to the middle of each sequence. Moreover, although its strength varies across models, AS consistently emerges at the initial position of the sequence even when the initial token is not the BOS token. We observe a similar trend for MAs, as shown in Figure 2. Taken together, these results indicate that both the BOS token and the initial position contribute to AS and MAs, with the latter operating independently of token type. These findings are consistent with prior reports that MAs occur at the initial position of a sequence (Sun et al., 2024) and that the BOS token triggers AS and MAs (Oh et al., 2025). Our results complement these works by providing a comprehensive quantitative evaluation of these phenomena using $\text{Sink}_j^\epsilon$ as a unified metric across multiple samples and models.

## 4 Mechanistic Analysis of the Initial-Position Effect on AS and MAs

In Section 3, we showed that AS and MAs are associated not only with the BOS token but also with

[3]Following Gu et al. (2025), we use $\epsilon = 0.3$. See Appendix E.5 for results confirming that our findings hold for $\epsilon \in \{0.2, 0.3, 0.4, 0.5\}$.

[4]We use the `wikitext-2-raw-v1` subset from https://huggingface.co/datasets/Salesforce/wikitext.

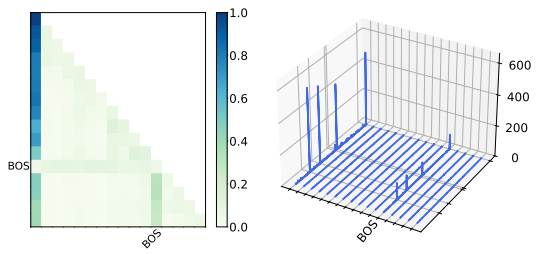



Figure 2: Attention heatmap (left) and activation magnitudes (right) from Layer 14 of Llama-3.2-3B with the BOS token swapped with the token at the 12th position of a 16-token sequence (shorter than the setting in Table 1 for readability). Both AS and MAs emerge at the initial position, independent of the BOS token location.

| Model | Position | Vanilla | Interv |
|---|---|---|---|
| Llama-2-7b-hf | 1 | 0.9305 | 0.9164 |
| | 16 | 0.0073 | **0.8397** |
| Llama-3.2-3B | 1 | 0.9218 | 0.8785 |
| | 16 | 0.0012 | **0.7424** |
| Mistral-7B-v0.3 | 1 | 0.0343 | 0.0198 |
| | 16 | 0.0373 | **0.1509** |
| Qwen2-7B | 1 | 0.8173 | 0.7278 |
| | 16 | 0.0000 | **0.5085** |
| pythia-1b | 1 | 0.7337 | 0.6702 |
| | 16 | 0.0000 | **0.4703** |

Table 3: Comparison of $\text{Sink}_j^\epsilon$ when intervention forces self-concentration of attention at position 16 (both w/o BOS).

properties specific to the sequence-initial position. In this section, we focus on the initial-position effect by examining two factors: positional encoding (Section 4.1) and the self-concentration of attention (Section 4.2). We then use these observations to further investigate the underlying mechanism that gives rise to AS and MAs (Section 4.3).

### 4.1 Influence of Positional Encoding

We investigate whether the positional information encoded by Rotary Position Embedding (RoPE; Su et al., 2024) contributes to the emergence of AS at the initial position. [5] To this end, we intervene on the RoPE index of the Key vector for the initial token. Specifically, following Equation (1), we replace the Key-side rotation matrix $\boldsymbol{R}_{\Theta,1}$ with $\boldsymbol{R}_{\Theta,r}$, where $r$ is sampled uniformly from $\{2,\ldots,T\}$, which excludes the initial token's original RoPE index. This yields the following intervened Key vector:

$$\boldsymbol{k}_1^{\text{interv}} = \boldsymbol{x}_1 \boldsymbol{W}_K \boldsymbol{R}_{\Theta,r}^\top \quad . \tag{3}$$

If RoPE were the primary driver of AS, this intervention would be expected to reduce $\text{Sink}_1^\epsilon$ substantially. To isolate the effect of the initial position from the BOS token, we conduct this experiment using sequences without a BOS token.

As shown in Table 2, the RoPE intervention does not substantially reduce $\text{Sink}_1^\epsilon$ in any of the five models. This result suggests that the initial token's Key-side RoPE encoding has only a limited effect on AS under this intervention.

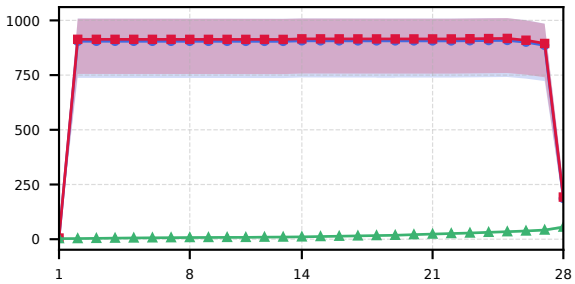



Figure 3: Layer-wise averaged hidden-state norms for Llama-3.2-3B under the self-concentration intervention (w/o BOS token). Blue circles, red squares, and green triangles denote the first token, the intervention token, and the mean over all other input tokens, respectively. The horizontal axis shows the layer index. Shaded bands show the sample min–max range.

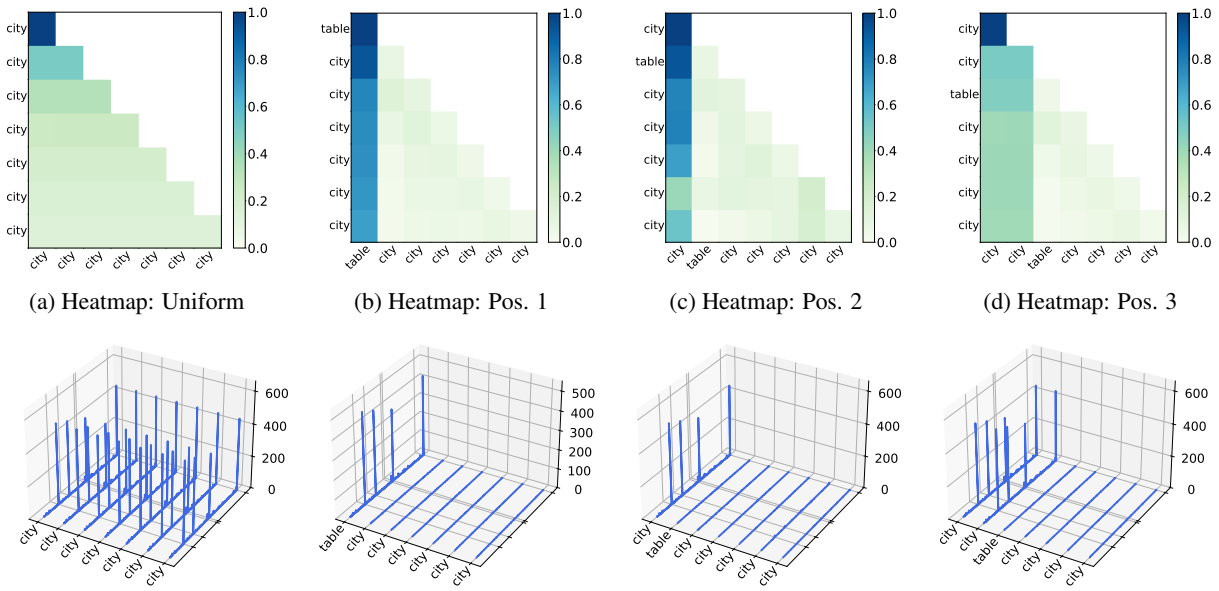
### 4.2 Influence of Self-Concentration of Attention

Given the limited impact of positional encoding, we then focus on the property that the initial token is forced to direct all attention to itself ($\alpha_{1,1} = 1.0$). We refer to this property as "self-concentration." To investigate whether this property contributes to AS and MAs, we intervene on the attention weights across all layers at a non-initial position ($t = 16$) to force self-concentration ($\alpha_{t,t} = 1.0, \alpha_{t,j\neq t} = 0.0$). We conduct this experiment using sequences without a BOS token as well.

Table 3 reports $\text{Sink}_t^\epsilon$ before and after forcing self-concentration at $t = 16$ across five models, and Figure 3 presents the corresponding layer-wise hidden-state norms for Llama-3.2-3B. These results show that forcing self-concentration induces both AS and MAs at non-initial positions in all five models,[6] although both effects are weaker for Mistral-7B-v0.3. This suggests that self-concentration is

[5] We focus on RoPE as it is adopted by most recent open-weight LLMs. Our models also differ in their RoPE configuration, such as the fraction of head dimensions to which the rotation is applied (Appendix B). Other positional-encoding schemes are outside our scope (see the Limitations section).

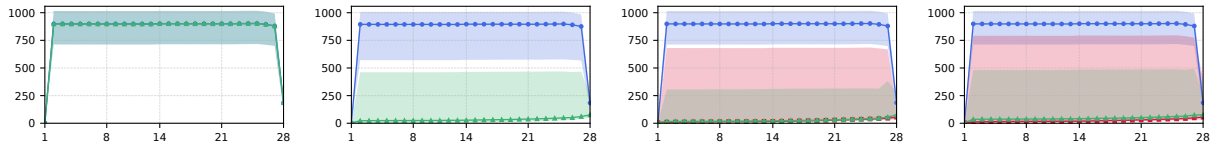
[6] Appendix E.4 shows similar trends at $t = 8$ and $t = 32$.



(a) Heatmap: Uniform (b) Heatmap: Pos. 1 (c) Heatmap: Pos. 2 (d) Heatmap: Pos. 3

(e) Activation: Uniform (f) Activation: Pos. 1 (g) Activation: Pos. 2 (h) Activation: Pos. 3

Figure 4: Attention heatmap (top row) and hidden-state activations (bottom row) at Layer 14 of Llama-3.2-3B for repeated-token sequences with varying distinct token positions.



(a) Norms: Uniform (b) Norms: Pos. 1 (c) Norms: Pos. 2 (d) Norms: Pos. 3

Figure 5: Layer-wise averaged hidden-state norms in Llama-3.2-3B for repeated-token sequences ($T = 64$, without a BOS token and without any attention intervention), averaged over $N = 100$ sequences per panel. Panels correspond to the four input patterns: a uniform repeated-token sequence (Uniform) and sequences with a single distinct token inserted at the first, second, or third position (Pos. 1–3). Blue circles and green triangles denote the first token and the mean over all remaining input tokens, respectively. Red squares denote the distinct token, and therefore appear only in the Pos. 2 and Pos. 3 panels: the Uniform panel contains no distinct token, and in Pos. 1 the distinct token is the first token itself. The horizontal axis shows the layer index in every panel. Shaded bands show the sample min–max range.

one factor that can trigger the emergence of AS and MAs.

### 4.3 Value-non-mixing as a Driver

**Hypothesis.** We hypothesize that the emergence of AS and MAs under self-concentration is driven by the resulting "Value-non-mixing" state, rather than by self-concentration itself. As shown in Equation (1), an attention mechanism computes each output $\boldsymbol{y}_i$ as a weighted average of the Value vectors corresponding to context tokens. Under self-concentration, however, the output consists solely of the token's own Value vector, without being mixed with those of other tokens (Figure 1). We formalize this Value-non-mixing state at position $i$ as

$$\boldsymbol{v}_j \mid j \leq i, ; \alpha_{i,j} > 0 = \boldsymbol{v}_i, \tag{4}$$

meaning that all Value vectors receiving nonzero attention from the query at position $i$ are identical to $\boldsymbol{v}_i$. Because the attention weights sum to one, this implies $\boldsymbol{y}_i = \boldsymbol{v}_i \boldsymbol{W}_O$.

**Experimental Design.** To test whether Value-non-mixing can induce AS and MAs, we reproduce this state without forcing self-concentration by using sequences composed of repeated identical tokens without a BOS token. Because all tokens share the same Value vector, the attention output remains a pass-through of that single Value vector even without self-concentration. By inserting

Hidden-state norms for the other four models are shown in Figure 6.

a distinct token into the sequence, we can control the range of positions over which the Value-non-mixing state is maintained. We evaluate four conditions: a uniform repeated-token sequence and sequences with a distinct token inserted at the first, second, or third position. For each of the five models, we generate $N = 100$ sequences per condition, each of length $T = 64$, randomly sampling the repeated token and, where applicable, the distinct token for each sequence. Although repeated-token inputs deviate from natural language, this manipulation complements the self-concentration intervention in Section 4.2, which induces the same state on natural text.

**Results.** Figure 4 shows the results for Llama-3.2-3B. When the sequence consists entirely of identical tokens, MAs are observed across all positions (Figure 4e), even though the attention weights are distributed across positions (Figure 4a). For sequences containing a distinct token, Figures 4b–4d and Figures 4f–4h show that AS and MAs occur at the initial token or at the identical-token positions preceding the distinct token, but subside at subsequent positions. This positional pattern is also visible in the layer-wise hidden-state norms shown in Figure 5. These observations for Llama-3.2-3B indicate that AS and MAs emerge at positions that remain in the Value-non-mixing state, supporting our hypothesis. See Appendices D and E.2 for results across all models, quantitative analyses, and additional experiments under varied settings.

For three of the five models examined, including Llama-3.2-3B, inserting a distinct token consistently produces AS and MAs across the sampled token combinations. For the remaining two, Mistral-7B-v0.3 and pythia-1b, the emergence of AS and MAs is more input-dependent, and the effect is also weaker on average (Appendix D). What drives this input dependence—for example, whether it relates to the similarity between the Value vectors of the repeated and inserted tokens—is left for future work.

Our findings provide further insight into the mechanisms underlying AS and MAs. Recent architectures proposed by major LLM developers, including Qwen3-Next (Qiu et al., 2025) and gpt-oss (OpenAI et al., 2025), highlight the growing importance of reducing these phenomena. These findings may also be relevant to the design of quantization-robust architectures.

## 5 Conclusion

In this study, we investigated the mechanism underlying AS and MAs in LLMs, with a particular focus on the sequence-initial position. At the sequence-initial position, causal masking inherently enforces self-concentration of attention ($\alpha_{1,1} = 1.0$). Our experiments show that forcing such self-concentration at a non-initial position induces AS and MAs across all five models. We further investigated the "Value-non-mixing" state resulting from self-concentration as a potential mechanism through which self-concentration induces AS and MAs. Repeated-token experiments provide additional support for this hypothesis in three of the five models, while the emergence of AS and MAs is more input-dependent and weaker on average in the remaining two. Together, these findings support a mechanistic link between self-concentration, Value-non-mixing, and the emergence of AS and MAs. By clarifying this mechanistic link, our findings provide empirical evidence that may inform future quantization strategies and advance our understanding of the internal mechanisms of attention layers. Future work should directly quantify Value-vector similarity and identify the contributing head-level pathways.

## Limitations

Our study has several limitations.

First, the five models we evaluate span four families (Llama, Mistral, Qwen, and Pythia; Appendix B), but whether our findings extend to other architectures (e.g., Mixture-of-Experts models (Shazeer et al., 2017)), languages, and domains beyond the English datasets evaluated here remains open. Among the five models, Mistral-7B-v0.3 responds more weakly to our interventions than the other models, and pythia-1b shows the weakest and most input-dependent effect in the repeated-token experiments (Section 4.3). Identifying the architectural differences behind this variance is left for future work. All five models are RoPE-based, though they vary structurally within RoPE (e.g., partial RoPE in pythia-1b), and Gu et al. (2025) report that positional embeddings did not affect the emergence of AS. How positional-encoding schemes influence the AS and MAs mechanism is outside our scope and an important future direction.

Second, we constrain our scope to the sequence-initial position to isolate its structural drivers. Our intervention induces AS at non-initial posi-

tions (Section 4.2 and Appendix E.4), but whether naturally occurring intermediate AS (Sun et al., 2024) can be explained within the same Value-non-mixing framework remains an open question.

Third, our interventions cannot exclude the contribution of other emergent factors. The RoPE intervention manipulates only the key-side positional index of the initial token, so isolating query-side effects and query–key interactions, and rigorously excluding RoPE's entire contribution, are left for future work. Likewise, restricting the intervention to early layers reproduces most of the all-layer effect (Appendix E.3), but pinpointing the head-level causal pathway is challenging and architecture-dependent, and we leave head-selective analyses for future work.

Fourth, our definition of Value-non-mixing (Equation (4)) characterizes a discrete state; extending it to a continuous, predictive "Value Mixing Index" requires modeling the geometry of Value vectors and validating that it predicts both $\mathrm{Sink}_j^\epsilon$ and MAs magnitude across models and natural inputs. The threshold-based $\mathrm{Sink}_j^\epsilon$ is also sensitive to sequence length, because attention weights for a query at position $i$ average $1/i$ over its accessible keys and thus more readily exceed a fixed $\epsilon$ in short sequences (e.g., repeat length 16 at $\epsilon = 0.2$ in Table 7), which limits fixed-threshold comparisons and motivates length-aware metrics.

Finally, although prior work distinguishes MAs from the activation outlier features that complicate low-bit quantization (Sun et al., 2024; Dettmers et al., 2022), we do not evaluate whether MAs or Value-non-mixing affects quantization error, perplexity, or downstream accuracy; establishing such a connection requires dedicated experiments.

## Acknowledgements

This work was supported by Google Research Grant; JST BOOST JPMJBY24D2, JPMJBS2412, JPMJBS2421; JST SPRING JPMJSP2104; JSPS KAKENHI 25K03175; and the Nakajima Foundation. We used generative AI tools to assist with coding and manuscript preparation. We thank the members of the Tohoku NLP Group for their cooperation in this research.

## References

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael M. Bronstein. 2025. Why do LLMs attend to the first token? In *Second Conference on Language Modeling (COLM)*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2023. Quantizable Transformers: Removing outliers by helping attention heads do nothing. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Enrique Queipo de Llano, Alvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael M. Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. 2026. Attention sinks and compression valleys in LLMs are two sides of the same coin. In *The Fourteenth International Conference on Learning Representations (ICLR)*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit matrix multiplication for Transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight:

Analyzing Transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations (ICLR)*.

Jaehoon Oh, Seungjun Shin, and Dokwan Oh. 2025. House of cards: Massive weights in LLMs. *Preprint*, arXiv:2410.01866.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *Preprint*, arXiv:2508.10925.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Nikolaus Salvatore, Hao Wang, and Qiong Zhang. 2025. Lost in the middle: An emergent property from information retrieval demands in LLMs. *Preprint*, arXiv:2510.10276.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. In *First Conference on Language Modeling (COLM)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. 2025. On the emergence of position bias in Transformers. In *Forty-second International Conference on Machine Learning (ICML)*.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, junxian guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2025. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations (ICLR)*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Re, Clark Barrett, Zhangyang Wang, and Beidi Chen. 2023. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*.

## A Related Work

AS and MAs are phenomena of practical importance from an engineering perspective. Specific tokens function as aggregation points for attention, enabling the reuse and compression of KV caches (Xiao et al., 2024, 2025; Zhang et al., 2023). Furthermore, the presence of activations with large magnitudes has been widely recognized as a major challenge for effectively quantizing LLMs (Dettmers et al., 2022; Sun et al., 2024). Note that MAs are distinct from the outlier features studied in the quantization literature (Sun et al., 2024, §2.3).

Existing literature analyzing AS and MAs has documented their layer-wise formation profiles and functional significance. Empirical observations indicate that these phenomena typically emerge in the initial layers, stabilize across the middle layers, and weaken near the final layers (Sun et al., 2024; Gu et al., 2025). To explain this vertical persistence, de Llano et al. (2026) proposed the Mix-Compress-Refine Theory, arguing that AS in the middle layers prevents the over-mixing of contextual information. Additionally, Gu et al. (2025) provided causal evidence that AS can be replaced by explicit key biases. Barbero et al. (2025) instead argue that AS prevents over-mixing, with the first token anchoring the residual stream against representational collapse (the Anti-Overmixing Theory).

From a mechanistic standpoint, several studies have investigated why specific tokens attract disproportionate attention. Kobayashi et al. (2020) showed, via a norm-based analysis of attention outputs, that BERT's attention heads assign large weights to uninformative delimiter tokens (e.g., [SEP]) whose value vectors are nonetheless small, so that the attention output leaves the residual stream almost unchanged. Bondarenko et al. (2023) relate this behavior to activation outliers observed more broadly in trained transformers (BERT, OPT, and ViT), framing such heads as having learned a no-op that leaves the residual stream unchanged or only partially updated. They report an analogous pattern for uninformative tokens beyond BERT (e.g., background patches in ViT). They further argue that a strict no-update requires near-exact zeros in the attention matrix, and that producing them drives the softmax inputs to ever larger values during training, which in turn creates outliers elsewhere in the network. Architectural factors have also been related to position bias more broadly. Wu et al. (2025) analyze multi-layer attention with a graph-theoretic framework and show that causal masking biases attention toward earlier positions, since tokens in deeper layers attend to increasingly contextualized representations of earlier tokens. They treat AS as one manifestation of the position bias covered by this framework. A complementary account attributes part of this bias to the training data. Salvatore et al. (2025) train GPT-2 and Llama variants from scratch on tasks that simulate long-term and short-term memory demands, and report that the primacy effect is induced by the uniform long-term demand and is further influenced by the autoregressive architecture and the formation of AS.

Our study differs from these accounts in its research question and direction of causality. Prior work primarily investigates the effects, functional roles, or alternative formulations of existing AS (e.g., anti-overmixing, key bias), whereas we investigate how AS and MAs emerge at the sequence-initial position. The notion of "mixing" also differs in meaning. For Barbero et al. (2025), mixing refers to the propagation of contextual information across positions and layers, whereas our "Value-non-mixing" concerns the composition of Value vectors within a single attention output (Equation (4)). The two accounts are therefore complementary. Our contribution is to control the Value-non-mixing state artificially, using attention interventions and repeated-token inputs into which a distinct token is inserted, and to show that AS and MAs emerge largely within the range of positions over which this state holds in the models we evaluate.

| Model | HF Identifier |
|---|---|
| Llama-2-7b-hf | meta-llama/Llama-2-7b-hf |
| Llama-3.2-3B | meta-llama/Llama-3.2-3B |
| Mistral-7B-v0.3 | mistralai/Mistral-7B-v0.3 |
| Qwen2-7B | Qwen/Qwen2-7B |
| pythia-1b | EleutherAI/pythia-1b |

Table 4: Open-weight LMs used in this study, with their Hugging Face identifiers as reported by the respective sources. Parameter counts are indicated in the model names.

## B Experimental Setup

To evaluate the generality of our findings regarding the mechanism of AS and MAs, we conduct our main experiments on the WikiText dataset (Merity et al., 2017)[7]. For the robustness analyses in Appendix E, we additionally use the GSM8K (Cobbe et al., 2021) and SlimPajama[8] datasets. We evaluate our hypothesis across a diverse set of prominent open-weight LLMs, including Llama-3.2-3B (Grattafiori et al., 2024), Llama-2-7b-hf (Touvron et al., 2023), Mistral-7B-v0.3 (Jiang et al., 2023), pythia-1b (Biderman et al., 2023), and Qwen2-7B (Yang et al., 2024). The details of the models and their Hugging Face identifiers are summarized in Table 4.

This selection covers several architectural variations in the attention and normalization components:

1. **Attention and Normalization Formats:** Spanning standard Multi-Head Attention (Llama-2-7b-hf, pythia-1b) and Grouped-Query Attention (Llama-3.2-3B, Mistral-7B-v0.3, Qwen2-7B), as well as contrasting standard Layer Normalization (pythia-1b) with RMSNorm (Llama families, Mistral-7B-v0.3, Qwen2-7B).

2. **Positional Encoding and Projection Biases:** pythia-1b applies RoPE to only 25% of each head's dimensions, whereas the other models

[7]We use the `wikitext-2-raw-v1` subset from https://huggingface.co/datasets/Salesforce/wikitext.

[8]We use `DKYoon/SlimPajama-6B` from https://huggingface.co/datasets/DKYoon/SlimPajama-6B.

apply it to the full head dimension. Qwen2-7B and pythia-1b use biases in their QKV projections, whereas the Llama and Mistral models do not.

We used an NVIDIA RTX 6000 Ada Generation GPU for our experiments, which took approximately 4 hours for the main experiments, with additional compute for the position-sweep experiments in Appendix E.4. All experiments were implemented with `PyTorch` 2.9.0 and `transformers` 4.57.1.

## C Transition of $\mathrm{Sink}_j^\epsilon$ with relocated BOS

Figure 7 shows the layer-wise transition of $\mathrm{Sink}_j^\epsilon$ under the setting of Section 3, at the sequence-initial position ($j = 1$) and at the position to which the BOS token is relocated ($j = 16$). As in Table 1, we report the three models whose tokenizer prepends a BOS token. Below we summarize the layers at which AS and MAs emerge for the relocated BOS token in each model:

- **Llama-3.2-3B**: AS and MAs are observed for the relocated BOS token from Layer 1 on.
- **Mistral-7B-v0.3**: AS emerges from Layer 1, whereas MAs emerge from Layer 2. AS therefore precedes MAs here, reversing the order we observe for Llama-2-7b-hf below.
- **Llama-2-7b-hf**: MAs emerge from Layer 2 and AS from Layer 3, so both appear in later layers than in the other two models.

## D Results for Repeated Token Sequences across Models

Figures 8–18 below and Figure 4 in the main text show a clear trend supporting the "Value-non-mixing" hypothesis in Llama-3.2-3B, Llama-2-7b-hf, and Qwen2-7B, and the same trend in a weaker form in Mistral-7B-v0.3. In pythia-1b, by contrast, $\mathrm{Sink}_1^\epsilon$ stays near zero for the first token, yet the attention pattern is not featureless. The distinct token and the following token, whose representations mix into a distinct value, receive visibly less attention than the repeated tokens, in the direction "Value-non-mixing" predicts. As in Figure 4, these example figures use short sequences ($T = 7$) with a fixed pair of tokens so that the per-token axis labels stay readable. Sample 1 repeats "home" with "cat" as the distinct token, and Sample 2 repeats "city" with "table". Figures 10, 13, 16 and 19 show the corresponding layer-wise hidden-state norm trajectories for these models.

Table 5 summarizes $\mathrm{Sink}_1^\epsilon$ at the first token position across the four repeated-token patterns. Uniform sequences (Pattern 0) yield $\mathrm{Sink}_1^{0.3}$ of 0.000 for all five models, whereas inserting a single distinct token (Pattern 1–3) raises it to 0.68–0.80 for Llama-2-7b-hf, 0.69–0.80 for Llama-3.2-3B, and 0.42–0.55 for Qwen2-7B. Mistral-7B-v0.3 (0.07–0.16) shows the same direction more weakly. For pythia-1b the effect is far smaller: $\mathrm{Sink}_1^{0.3}$ is 0.006 for Pattern 1 and rounds to 0.000 for Patterns 2–3. The model still moves in the direction "Value-non-mixing" predicts. A looser threshold of $\epsilon = 0.2$ separates Pattern 1 (0.030) from Pattern 0 (0.000), and Patterns 1 and 2 remain above Pattern 3, as in the other four models. This effect is nonetheless much smaller than in Llama-2-7b-hf or Llama-3.2-3B, and $\mathrm{Sink}_1^\epsilon$ stays at or below 0.03 at every threshold we test.

## E Additional Ablations under Varied Configurations

To assess the robustness of our findings, we additionally evaluate the self-concentration intervention of Section 4.2 under varied experimental configurations: different natural-language datasets and input sequence lengths, different repeat lengths for the repeated-token setting introduced in Appendix D, restricting the intervention to subsets of layers, applying the intervention at additional intervened positions, and varying the threshold $\epsilon$ used to define $\mathrm{Sink}_j^\epsilon$.

### E.1 Varying Dataset and Input Sequence Length

Table 6 reports $\mathrm{Sink}_{16}^\epsilon$ at the intervened token position ($t = 16$) under the self-concentration intervention ($\alpha_{16,16} = 1.0$, all layers, w/o BOS) for all five models across WikiText, GSM8K, and SlimPajama, and across input sequence lengths of 32, 64, 128, and (for WikiText) 256 tokens. At $\epsilon = 0.3$, the intervention yields a nonzero $\mathrm{Sink}_{16}^\epsilon$ for every model across all datasets and lengths, although the magnitude varies substantially across models and settings, particularly for Mistral-7B-v0.3.

As a complementary baseline (no-intervention) check, Figure 20 shows layer-wise hidden-state norms for Llama-3.2-3B at an input length of 64 tokens across the three datasets, confirming that the

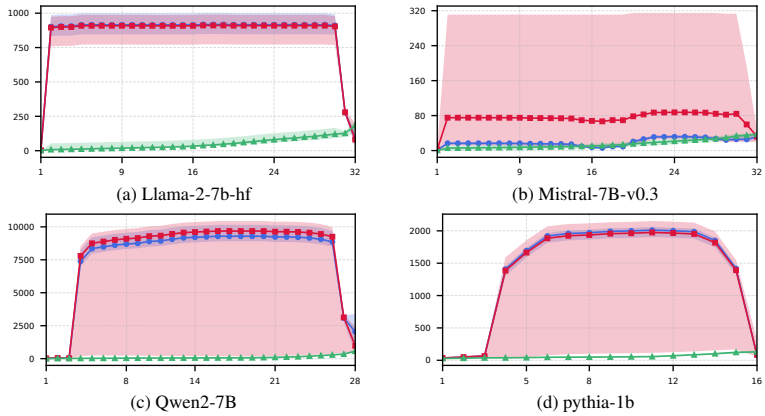



Figure 6: Layer-wise averaged hidden-state norms under the self-concentration intervention (w/o BOS token) for the four models other than Llama-3.2-3B. Blue circles, red squares, and green triangles denote the first token, the intervention token, and the mean over all other input tokens, respectively. The horizontal axis shows the layer index in every panel. Shaded bands show the sample min–max range.

| **Model** | **Pattern** | $\epsilon = 0.2$ | $\epsilon = 0.3$ | $\epsilon = 0.4$ | $\epsilon = 0.5$ |
|---|---|---|---|---|---|
| Llama-2-7b-hf | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | Distinct (Pattern 1) | 0.8454 | 0.7754 | 0.6898 | 0.6146 |
| | Distinct (Pattern 2) | 0.8629 | 0.7955 | 0.7248 | 0.6529 |
| | Distinct (Pattern 3) | 0.7934 | 0.6829 | 0.5166 | 0.0843 |
| Llama-3.2-3B | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | Distinct (Pattern 1) | 0.8197 | 0.7740 | 0.7177 | 0.6457 |
| | Distinct (Pattern 2) | 0.8464 | 0.8042 | 0.7439 | 0.6665 |
| | Distinct (Pattern 3) | 0.8039 | 0.6876 | 0.4526 | 0.0182 |
| Mistral-7B-v0.3 | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | Distinct (Pattern 1) | 0.2344 | 0.1548 | 0.1104 | 0.0811 |
| | Distinct (Pattern 2) | 0.2413 | 0.1568 | 0.1083 | 0.0784 |
| | Distinct (Pattern 3) | 0.1435 | 0.0737 | 0.0311 | 0.0004 |
| Qwen2-7B | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | Distinct (Pattern 1) | 0.6475 | 0.5482 | 0.4688 | 0.3941 |
| | Distinct (Pattern 2) | 0.6450 | 0.5404 | 0.4586 | 0.3803 |
| | Distinct (Pattern 3) | 0.5829 | 0.4179 | 0.2264 | 0.0171 |
| pythia-1b | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | Distinct (Pattern 1) | 0.0302 | 0.0055 | 0.0035 | 0.0024 |
| | Distinct (Pattern 2) | 0.0111 | 0.0000 | 0.0000 | 0.0000 |
| | Distinct (Pattern 3) | 0.0001 | 0.0000 | 0.0000 | 0.0000 |

Table 5: Comparison of $\text{Sink}_1^\epsilon$ at the first token position between the uniform sequence (Pattern 0) and sequences with a single distinct token (Pattern 1–3) for repeated-token sequences with $T = 64$.

large first-token norms reported in the main text are not specific to WikiText.

### E.2 Varying Repeat Length for Repeated-Token Sequences

Tables 7–9 extend the repeated-token analysis of Appendix D to repeat lengths of 16, 32, 64, 128, and 256 tokens, confirming that the gap between the Uniform and Distinct-token patterns persists at every repeat length. The only exception occurs at the shortest repeat length (16) with the lowest threshold ($\epsilon = 0.2$), where even uniform sequences yield high $\text{Sink}_1^\epsilon$ values. At shorter lengths, baseline attention weights are inherently higher, so normal fluctuations easily exceed the fixed absolute threshold (see Limitations).

Figure 21 shows the corresponding layer-wise hidden-state norms for Llama-2-7b-hf at a repeat length of 128 tokens, mirroring the trend already reported for length 64 in Figure 10.

### E.3 Restricting the Self-Concentration Intervention to Layer Subsets

The self-concentration intervention in Section 4.2 enforces $\alpha_{t,t} = 1.0$ across all layers. To identify which layers drive the resulting AS, we restrict the same intervention (at $t = 16$, $T = 64$, without a BOS token) to three-layer subsets: early (layers 1–3), middle (the three layers around the network midpoint), and late (the three layers preceding the

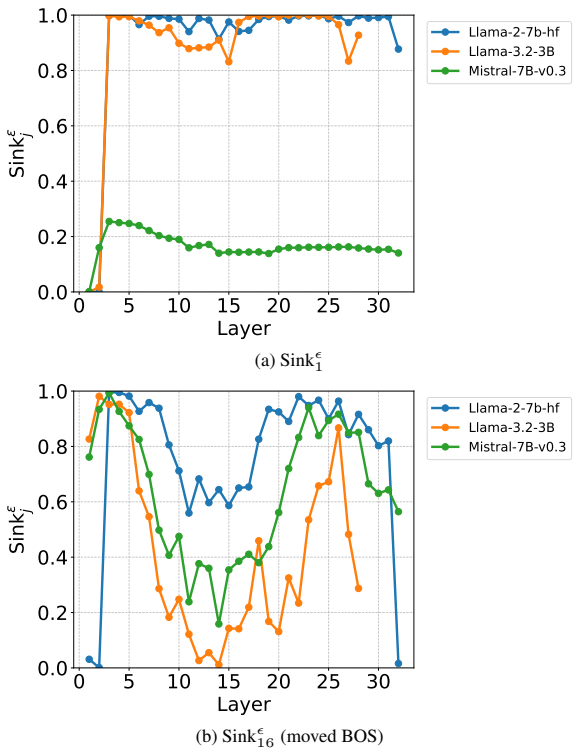



Figure 7: Transition of layer-wise Sink$_j^\epsilon$ when the BOS token is swapped with the token at position 16 (the setting of Table 1).

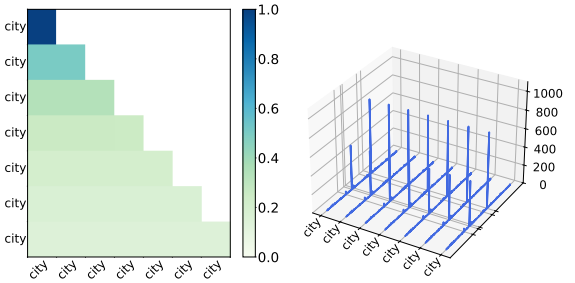
final layer). Table 10 reports Sink$_{16}^\epsilon$ at the intervened position for each setting. Intervening only on the early layers consistently yields Sink$_{16}^{0.3}$ values reaching 90–93% of the all-layer intervention across all five models, whereas intervening on the middle or late layers leaves Sink$_{16}^\epsilon$ near zero or at its vanilla level in all five models. Intervening on layer 1 alone nearly reproduces the all-layer effect in Mistral-7B-v0.3 and pythia-1b, yields a partial increase in Llama-3.2-3B, and has minimal impact on Llama-2-7b-hf and Qwen2-7B, indicating that which early layers are responsible varies across models. These results are consistent with the early-layer changes observed in the hidden-state norm trajectories.

### E.4 Varying the Intervened Position

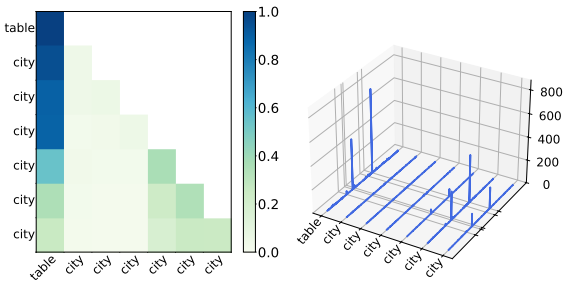
The self-concentration intervention in Section 4.2 forces $\alpha_{t,t} = 1.0$ at a single position, $t = 16$. To test whether the resulting increase in AS extends to other positions, we repeat the same intervention (all layers, w/o BOS, $T = 64$) at $t = 8$ and $t = 32$. Table 11 reports Sink$_t^\epsilon$ at the intervened position for both settings. The intervention increases Sink$_t^\epsilon$ over the vanilla condition for all five models at both $t = 8$ and $t = 32$, showing that the observed increase is not unique to $t = 16$ among the positions tested. The increase is markedly smaller for Mistral-7B-v0.3, rising from 0.0660 to 0.1347 at $t = 8$ and from 0.0000 to 0.0866 at $t = 32$, several times below the intervened values of the other four models (0.4654–0.8542, Table 11).

### E.5 Varying Epsilon Thresholds

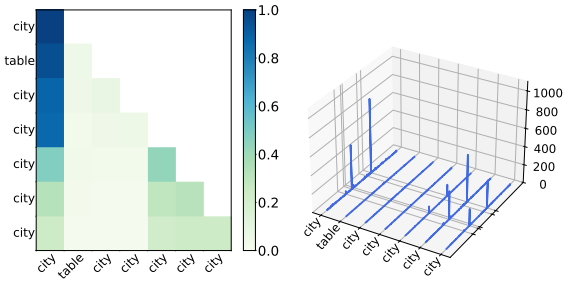
To verify that our main findings are not overly sensitive to the choice of the threshold $\epsilon$, we present the comparison of Sink$_j^\epsilon$ across various thresholds ($\epsilon \in \{0.2, 0.3, 0.4, 0.5\}$) for the RoPE intervention experiment (Table 12) and the self-concentration intervention experiment (Table 13). The overall trends remain consistent regardless of the specific threshold used.



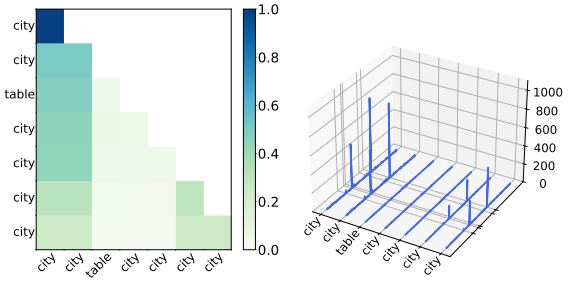

(a) Heatmap: Uniform sequence (b) Activation: Uniform sequence




(c) Heatmap: Distinct token at pos. 1 (d) Activation: Distinct token at pos. 1

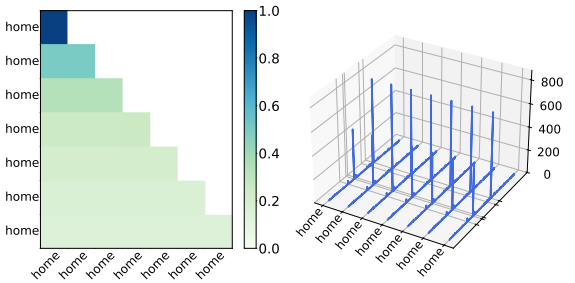



(e) Heatmap: Distinct token at pos. 2 (f) Activation: Distinct token at pos. 2

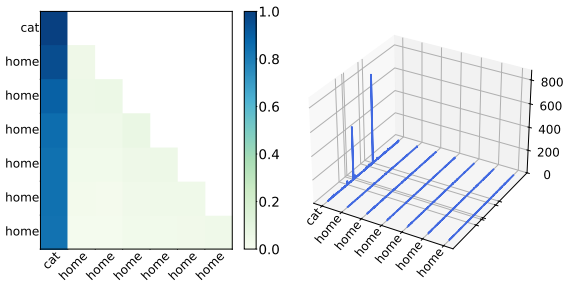



(g) Heatmap: Distinct token at pos. 3 (h) Activation: Distinct token at pos. 3

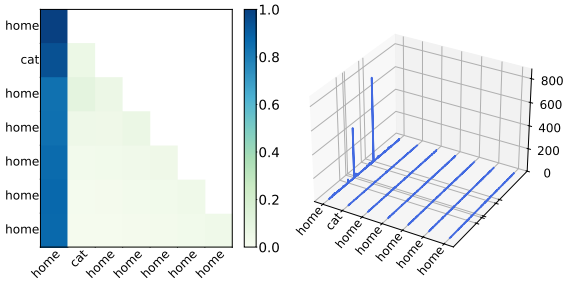
Figure 8: Results for sequences with repeated tokens. Each row compares the attention heatmap (left) and hidden-state activations (right) under the same token distribution pattern for Llama-2-7b-hf (Sample 2).



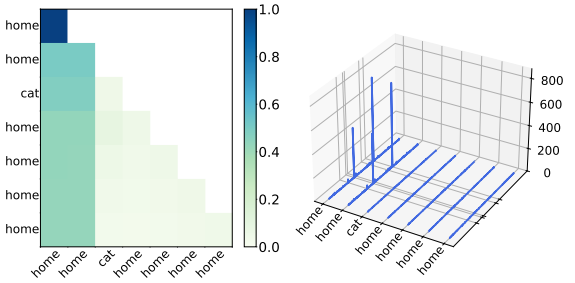

(a) Heatmap: Uniform sequence (b) Activation: Uniform sequence




(c) Heatmap: Distinct token at pos. 1 (d) Activation: Distinct token at pos. 1

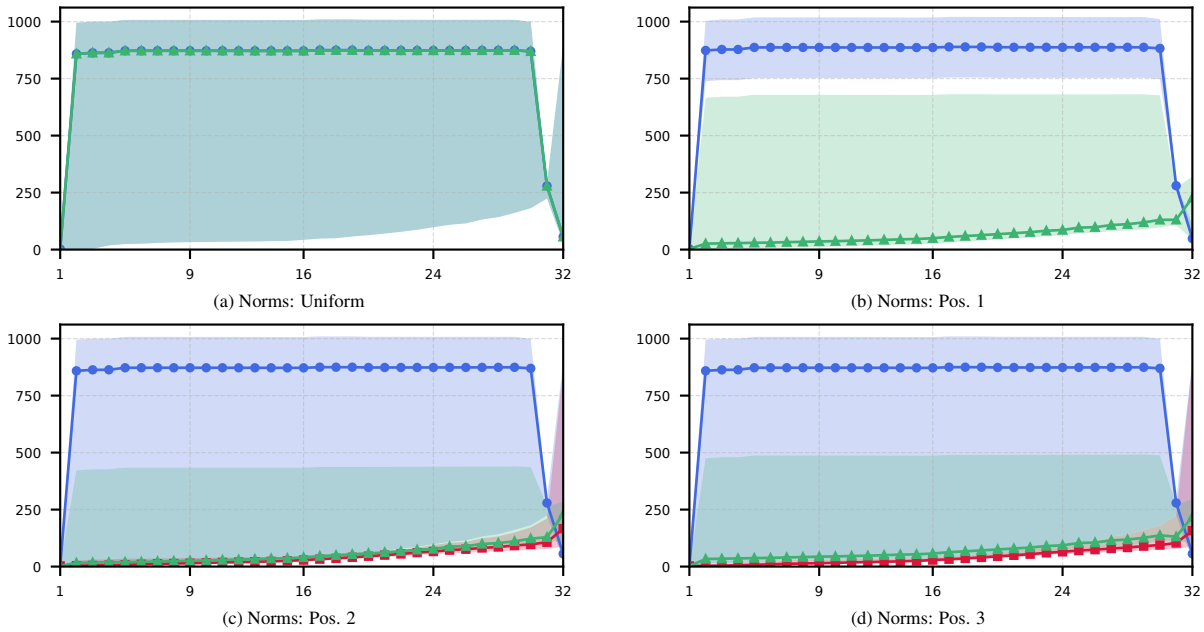



(e) Heatmap: Distinct token at pos. 2 (f) Activation: Distinct token at pos. 2




(g) Heatmap: Distinct token at pos. 3 (h) Activation: Distinct token at pos. 3

Figure 9: Results for sequences with repeated tokens. Each row compares the attention heatmap (left) and hidden-state activations (right) under the same token distribution pattern for Llama-2-7b-hf (Sample 1).




(a) Norms: Uniform

(b) Norms: Pos. 1

(c) Norms: Pos. 2

(d) Norms: Pos. 3

Figure 10: Layer-wise averaged hidden-state norms in Llama-2-7b-hf for repeated-token sequences under different distinct-token positions (colors and markers as in Figure 5). The horizontal axis shows the layer index in every panel.

| Dataset | $T$ | Model | $\epsilon = 0.2$ | $\epsilon = 0.3$ | $\epsilon = 0.4$ | $\epsilon = 0.5$ |
|---|---|---|---|---|---|---|
| WikiText | 32 | Llama-2-7b-hf | 0.9317 | 0.8932 | 0.6865 | 0.1487 |
| | | Llama-3.2-3B | 0.9329 | 0.8389 | 0.5787 | 0.1651 |
| | | Mistral-7B-v0.3 | 0.3077 | 0.1770 | 0.1109 | 0.0386 |
| | | Qwen2-7B | 0.8199 | 0.6346 | 0.3563 | 0.0735 |
| | | pythia-1b | 0.7342 | 0.6005 | 0.3301 | 0.0544 |
| | 64 | Llama-2-7b-hf | 0.9248 | 0.8397 | 0.5308 | 0.0095 |
| | | Llama-3.2-3B | 0.9121 | 0.7424 | 0.4527 | 0.0120 |
| | | Mistral-7B-v0.3 | 0.1897 | 0.1509 | 0.0868 | 0.0304 |
| | | Qwen2-7B | 0.7533 | 0.5085 | 0.2174 | 0.0182 |
| | | pythia-1b | 0.6975 | 0.4703 | 0.1703 | 0.0054 |
| | 128 | Llama-2-7b-hf | 0.9176 | 0.7870 | 0.4259 | 0.0013 |
| | | Llama-3.2-3B | 0.8888 | 0.6888 | 0.3911 | 0.0024 |
| | | Mistral-7B-v0.3 | 0.1830 | 0.1414 | 0.0762 | 0.0284 |
| | | Qwen2-7B | 0.7122 | 0.4401 | 0.1464 | 0.0113 |
| | | pythia-1b | 0.6729 | 0.3746 | 0.1013 | 0.0005 |
| | 256 | Llama-2-7b-hf | 0.9125 | 0.7543 | 0.3730 | 0.0009 |
| | | Llama-3.2-3B | 0.8767 | 0.6696 | 0.3717 | 0.0017 |
| | | Mistral-7B-v0.3 | 0.1810 | 0.1382 | 0.0726 | 0.0275 |
| | | Qwen2-7B | 0.6952 | 0.4119 | 0.1256 | 0.0102 |
| | | pythia-1b | 0.6570 | 0.3329 | 0.0805 | 0.0005 |
| GSM8K | 32 | Llama-2-7b-hf | 0.9196 | 0.7548 | 0.2460 | 0.0349 |
| | | Llama-3.2-3B | 0.9216 | 0.7736 | 0.4897 | 0.1090 |
| | | Mistral-7B-v0.3 | 0.3357 | 0.2053 | 0.1008 | 0.0051 |
| | | Qwen2-7B | 0.8138 | 0.6294 | 0.3222 | 0.0519 |
| | | pythia-1b | 0.7115 | 0.4763 | 0.1953 | 0.0103 |
| | 64 | Llama-2-7b-hf | 0.8880 | 0.5561 | 0.0577 | 0.0002 |
| | | Llama-3.2-3B | 0.8922 | 0.6769 | 0.3905 | 0.0207 |
| | | Mistral-7B-v0.3 | 0.2479 | 0.1705 | 0.0624 | 0.0009 |
| | | Qwen2-7B | 0.7682 | 0.5200 | 0.1943 | 0.0206 |
| | | pythia-1b | 0.6488 | 0.3407 | 0.0941 | 0.0000 |
| | 128 | Llama-2-7b-hf | 0.8789 | 0.5132 | 0.0438 | 0.0002 |
| | | Llama-3.2-3B | 0.8871 | 0.6656 | 0.3800 | 0.0186 |
| | | Mistral-7B-v0.3 | 0.2462 | 0.1659 | 0.0563 | 0.0006 |
| | | Qwen2-7B | 0.7632 | 0.5053 | 0.1809 | 0.0186 |
| | | pythia-1b | 0.6416 | 0.3180 | 0.0811 | 0.0000 |
| SlimPajama | 32 | Llama-2-7b-hf | 0.9291 | 0.8330 | 0.3042 | 0.0575 |
| | | Llama-3.2-3B | 0.9328 | 0.8576 | 0.6092 | 0.1773 |
| | | Mistral-7B-v0.3 | 0.3432 | 0.2173 | 0.1108 | 0.0249 |
| | | Qwen2-7B | 0.8243 | 0.6595 | 0.3841 | 0.0821 |
| | | pythia-1b | 0.7344 | 0.5374 | 0.2180 | 0.0462 |
| | 64 | Llama-2-7b-hf | 0.9144 | 0.6415 | 0.0995 | 0.0002 |
| | | Llama-3.2-3B | 0.9161 | 0.7680 | 0.4761 | 0.0143 |
| | | Mistral-7B-v0.3 | 0.2329 | 0.1644 | 0.0649 | 0.0153 |
| | | Qwen2-7B | 0.7640 | 0.5403 | 0.2449 | 0.0185 |
| | | pythia-1b | 0.6631 | 0.3365 | 0.0972 | 0.0038 |
| | 128 | Llama-2-7b-hf | 0.8908 | 0.4642 | 0.0178 | 0.0000 |
| | | Llama-3.2-3B | 0.8935 | 0.7009 | 0.4078 | 0.0024 |
| | | Mistral-7B-v0.3 | 0.2126 | 0.1419 | 0.0429 | 0.0134 |
| | | Qwen2-7B | 0.7221 | 0.4605 | 0.1671 | 0.0105 |
| | | pythia-1b | 0.5809 | 0.2037 | 0.0381 | 0.0004 |

Table 6: $\text{Sink}_{16}^{\epsilon}$ at the intervened token position ($t = 16$) under the self-concentration intervention ($\alpha_{16,16} = 1.0$, all layers, w/o BOS) across natural-language datasets and input sequence lengths $T$.

| Repeat Length | Model | Pattern | $\epsilon = 0.2$ | $\epsilon = 0.3$ | $\epsilon = 0.4$ | $\epsilon = 0.5$ |
|---|---|---|---|---|---|---|
| 16 | Llama-2-7b-hf | Uniform (Pattern 0) | 0.9593 | 0.0001 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.9551 | 0.8931 | 0.8340 | 0.7284 |
| | | Distinct (Pattern 2) | 0.9469 | 0.8996 | 0.8479 | 0.7545 |
| | | Distinct (Pattern 3) | 0.9419 | 0.8283 | 0.6307 | 0.3001 |
| | Llama-3.2-3B | Uniform (Pattern 0) | 0.9426 | 0.0035 | 0.0001 | 0.0000 |
| | | Distinct (Pattern 1) | 0.9328 | 0.8489 | 0.7883 | 0.7330 |
| | | Distinct (Pattern 2) | 0.9353 | 0.8634 | 0.8224 | 0.7545 |
| | | Distinct (Pattern 3) | 0.9205 | 0.8111 | 0.6421 | 0.1864 |
| | Mistral-7B-v0.3 | Uniform (Pattern 0) | 0.7456 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.9019 | 0.6684 | 0.3816 | 0.2671 |
| | | Distinct (Pattern 2) | 0.9348 | 0.7320 | 0.4111 | 0.2808 |
| | | Distinct (Pattern 3) | 0.9101 | 0.3667 | 0.1499 | 0.0177 |
| | Qwen2-7B | Uniform (Pattern 0) | 0.9087 | 0.0063 | 0.0025 | 0.0006 |
| | | Distinct (Pattern 1) | 0.8847 | 0.7919 | 0.6725 | 0.5873 |
| | | Distinct (Pattern 2) | 0.8933 | 0.7882 | 0.6714 | 0.5754 |
| | | Distinct (Pattern 3) | 0.8490 | 0.6448 | 0.4252 | 0.1036 |
| | pythia-1b | Uniform (Pattern 0) | 0.7830 | 0.0067 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.8373 | 0.6219 | 0.2639 | 0.1123 |
| | | Distinct (Pattern 2) | 0.8245 | 0.6010 | 0.2709 | 0.0927 |
| | | Distinct (Pattern 3) | 0.8059 | 0.2901 | 0.0246 | 0.0000 |
| 32 | Llama-2-7b-hf | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.8917 | 0.8243 | 0.7371 | 0.6524 |
| | | Distinct (Pattern 2) | 0.8923 | 0.8449 | 0.7622 | 0.6809 |
| | | Distinct (Pattern 3) | 0.8502 | 0.7205 | 0.5495 | 0.1633 |
| | Llama-3.2-3B | Uniform (Pattern 0) | 0.0007 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.8474 | 0.8017 | 0.7498 | 0.6868 |
| | | Distinct (Pattern 2) | 0.8625 | 0.8310 | 0.7784 | 0.7034 |
| | | Distinct (Pattern 3) | 0.8291 | 0.7405 | 0.5247 | 0.0630 |
| | Mistral-7B-v0.3 | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.5238 | 0.2821 | 0.1919 | 0.1446 |
| | | Distinct (Pattern 2) | 0.5833 | 0.2995 | 0.1927 | 0.1447 |
| | | Distinct (Pattern 3) | 0.3187 | 0.1443 | 0.0638 | 0.0017 |
| | Qwen2-7B | Uniform (Pattern 0) | 0.0028 | 0.0004 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.7825 | 0.6552 | 0.5579 | 0.4807 |
| | | Distinct (Pattern 2) | 0.7781 | 0.6526 | 0.5556 | 0.4688 |
| | | Distinct (Pattern 3) | 0.6813 | 0.5158 | 0.3090 | 0.0410 |
| | pythia-1b | Uniform (Pattern 0) | 0.0013 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.4490 | 0.0870 | 0.0164 | 0.0055 |
| | | Distinct (Pattern 2) | 0.4200 | 0.0674 | 0.0045 | 0.0001 |
| | | Distinct (Pattern 3) | 0.1306 | 0.0002 | 0.0000 | 0.0000 |

Table 7: $\text{Sink}_1^{\epsilon}$ for repeated-token sequences across repeat lengths, without intervention (Part 1: Length 16, 32).

| Repeat Length | Model | Pattern | $\epsilon = 0.2$ | $\epsilon = 0.3$ | $\epsilon = 0.4$ | $\epsilon = 0.5$ |
|---|---|---|---|---|---|---|
| 64 | Llama-2-7b-hf | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.8454 | 0.7754 | 0.6898 | 0.6146 |
| | | Distinct (Pattern 2) | 0.8629 | 0.7955 | 0.7248 | 0.6529 |
| | | Distinct (Pattern 3) | 0.7934 | 0.6829 | 0.5166 | 0.0843 |
| | Llama-3.2-3B | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.8197 | 0.7740 | 0.7177 | 0.6457 |
| | | Distinct (Pattern 2) | 0.8464 | 0.8042 | 0.7439 | 0.6665 |
| | | Distinct (Pattern 3) | 0.8039 | 0.6876 | 0.4526 | 0.0182 |
| | Mistral-7B-v0.3 | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.2344 | 0.1548 | 0.1104 | 0.0811 |
| | | Distinct (Pattern 2) | 0.2413 | 0.1568 | 0.1083 | 0.0784 |
| | | Distinct (Pattern 3) | 0.1435 | 0.0737 | 0.0311 | 0.0004 |
| | Qwen2-7B | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.6475 | 0.5482 | 0.4688 | 0.3941 |
| | | Distinct (Pattern 2) | 0.6450 | 0.5404 | 0.4586 | 0.3803 |
| | | Distinct (Pattern 3) | 0.5829 | 0.4179 | 0.2264 | 0.0171 |
| | pythia-1b | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.0302 | 0.0055 | 0.0035 | 0.0024 |
| | | Distinct (Pattern 2) | 0.0111 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 3) | 0.0001 | 0.0000 | 0.0000 | 0.0000 |
| 128 | Llama-2-7b-hf | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.8125 | 0.7484 | 0.6691 | 0.6022 |
| | | Distinct (Pattern 2) | 0.8289 | 0.7752 | 0.7064 | 0.6411 |
| | | Distinct (Pattern 3) | 0.7708 | 0.6646 | 0.5060 | 0.0408 |
| | Llama-3.2-3B | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.7987 | 0.7472 | 0.6803 | 0.6042 |
| | | Distinct (Pattern 2) | 0.8304 | 0.7848 | 0.7171 | 0.6371 |
| | | Distinct (Pattern 3) | 0.7766 | 0.6452 | 0.4095 | 0.0032 |
| | Mistral-7B-v0.3 | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.1295 | 0.0828 | 0.0584 | 0.0410 |
| | | Distinct (Pattern 2) | 0.1294 | 0.0815 | 0.0529 | 0.0326 |
| | | Distinct (Pattern 3) | 0.0721 | 0.0322 | 0.0129 | 0.0001 |
| | Qwen2-7B | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.5524 | 0.4658 | 0.3957 | 0.3326 |
| | | Distinct (Pattern 2) | 0.5207 | 0.4278 | 0.3519 | 0.2928 |
| | | Distinct (Pattern 3) | 0.4756 | 0.3338 | 0.1788 | 0.0073 |
| | pythia-1b | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.0070 | 0.0045 | 0.0023 | 0.0011 |
| | | Distinct (Pattern 2) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 3) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

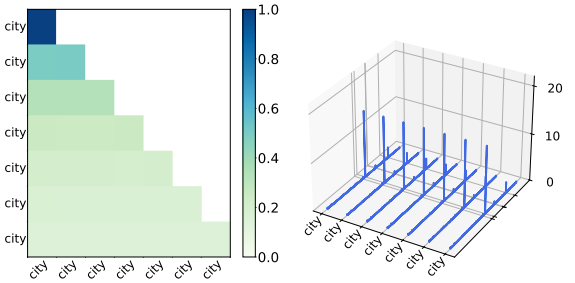
Table 8: $\text{Sink}_1^{\epsilon}$ for repeated-token sequences across repeat lengths, without intervention (Part 2: Length 64, 128).

| Repeat Length | Model | Pattern | $\epsilon = 0.2$ | $\epsilon = 0.3$ | $\epsilon = 0.4$ | $\epsilon = 0.5$ |
|---|---|---|---|---|---|---|
| 256 | Llama-2-7b-hf | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.7957 | 0.7306 | 0.6570 | 0.5928 |
| | | Distinct (Pattern 2) | 0.8138 | 0.7608 | 0.6949 | 0.6342 |
| | | Distinct (Pattern 3) | 0.7488 | 0.6524 | 0.5011 | 0.0130 |
| | Llama-3.2-3B | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.7739 | 0.7121 | 0.6482 | 0.5734 |
| | | Distinct (Pattern 2) | 0.8101 | 0.7543 | 0.6868 | 0.6057 |
| | | Distinct (Pattern 3) | 0.7456 | 0.6120 | 0.3770 | 0.0002 |
| | Mistral-7B-v0.3 | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.0665 | 0.0385 | 0.0255 | 0.0184 |
| | | Distinct (Pattern 2) | 0.0616 | 0.0303 | 0.0209 | 0.0148 |
| | | Distinct (Pattern 3) | 0.0290 | 0.0152 | 0.0061 | 0.0000 |
| | Qwen2-7B | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.4623 | 0.3806 | 0.3147 | 0.2558 |
| | | Distinct (Pattern 2) | 0.4389 | 0.3761 | 0.3215 | 0.2667 |
| | | Distinct (Pattern 3) | 0.4187 | 0.2900 | 0.1526 | 0.0030 |
| | pythia-1b | Uniform (Pattern 0) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 1) | 0.0044 | 0.0019 | 0.0009 | 0.0004 |
| | | Distinct (Pattern 2) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | Distinct (Pattern 3) | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

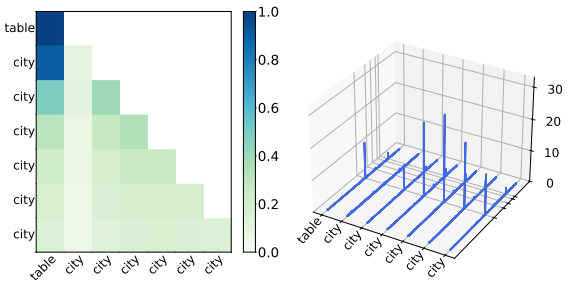
Table 9: $\text{Sink}_1^{\epsilon}$ for repeated-token sequences across repeat lengths, without intervention (Part 3: Length 256).

| Model | Vanilla | All layers | Early | Middle | Late | Layer 1 |
|---|---|---|---|---|---|---|
| Llama-2-7b-hf | 0.0073 | 0.8397 | 0.7724 | 0.0074 | 0.0074 | 0.0241 |
| Llama-3.2-3B | 0.0012 | 0.7424 | 0.6670 | 0.0013 | 0.0012 | 0.1310 |
| Mistral-7B-v0.3 | 0.0373 | 0.1509 | 0.1396 | 0.0374 | 0.0374 | 0.1385 |
| Qwen2-7B | 0.0000 | 0.5085 | 0.4737 | 0.0000 | 0.0000 | 0.0036 |
| pythia-1b | 0.0000 | 0.4703 | 0.4296 | 0.0000 | 0.0000 | 0.4443 |

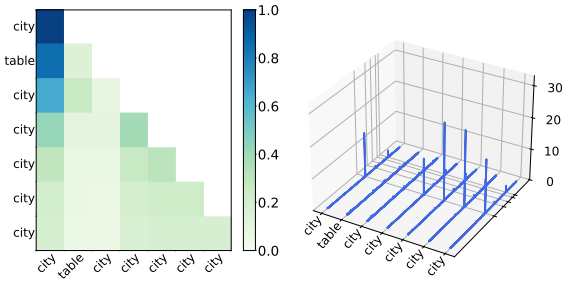
Table 10: $\text{Sink}_{16}^{\epsilon}$ at the intervened token position ($t = 16$) when the self-concentration intervention ($\alpha_{t,t} = 1.0$) is applied to all layers, restricted to the early, middle, or late three layers, or to layer 1 alone, with $T = 64$ (all w/o BOS).



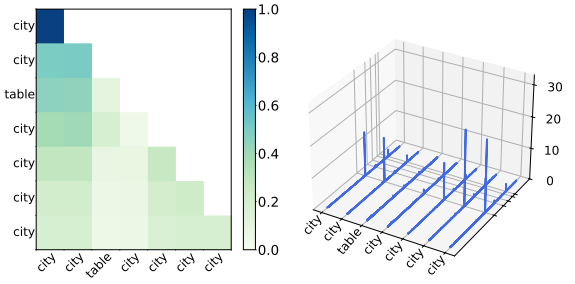

(a) Heatmap: Uniform sequence (b) Activation: Uniform sequence




(c) Heatmap: Distinct token at pos. 1 (d) Activation: Distinct token at pos. 1

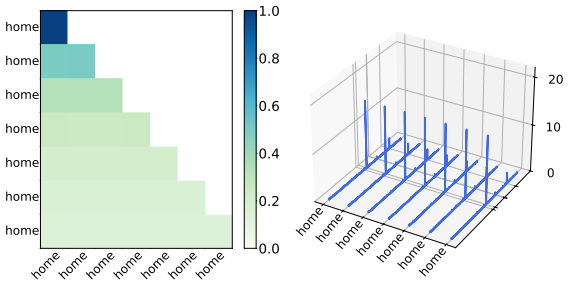



(e) Heatmap: Distinct token at pos. 2 (f) Activation: Distinct token at pos. 2

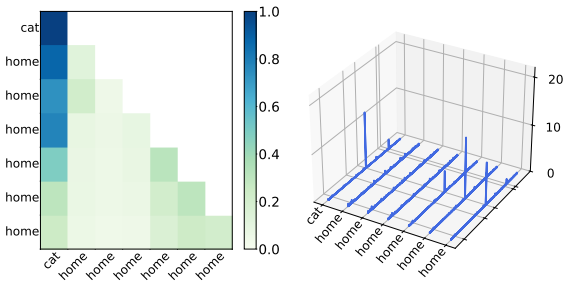



(g) Heatmap: Distinct token at pos. 3 (h) Activation: Distinct token at pos. 3

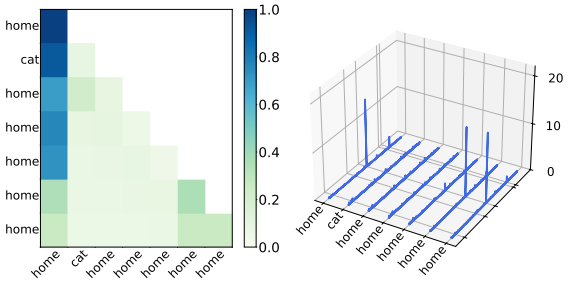
Figure 11: Results for sequences with repeated tokens. Each row compares the attention heatmap (left) and hidden-state activations (right) under the same token distribution pattern for Mistral-7B-v0.3 (Sample 2).



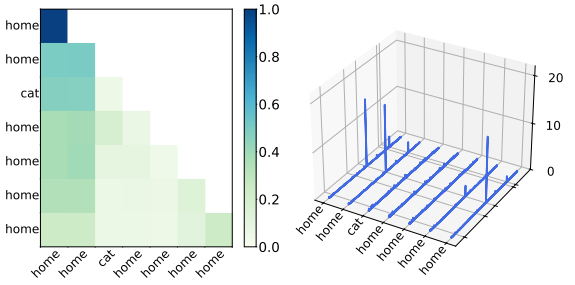

(a) Heatmap: Uniform sequence (b) Activation: Uniform sequence




(c) Heatmap: Distinct token at pos. 1 (d) Activation: Distinct token at pos. 1

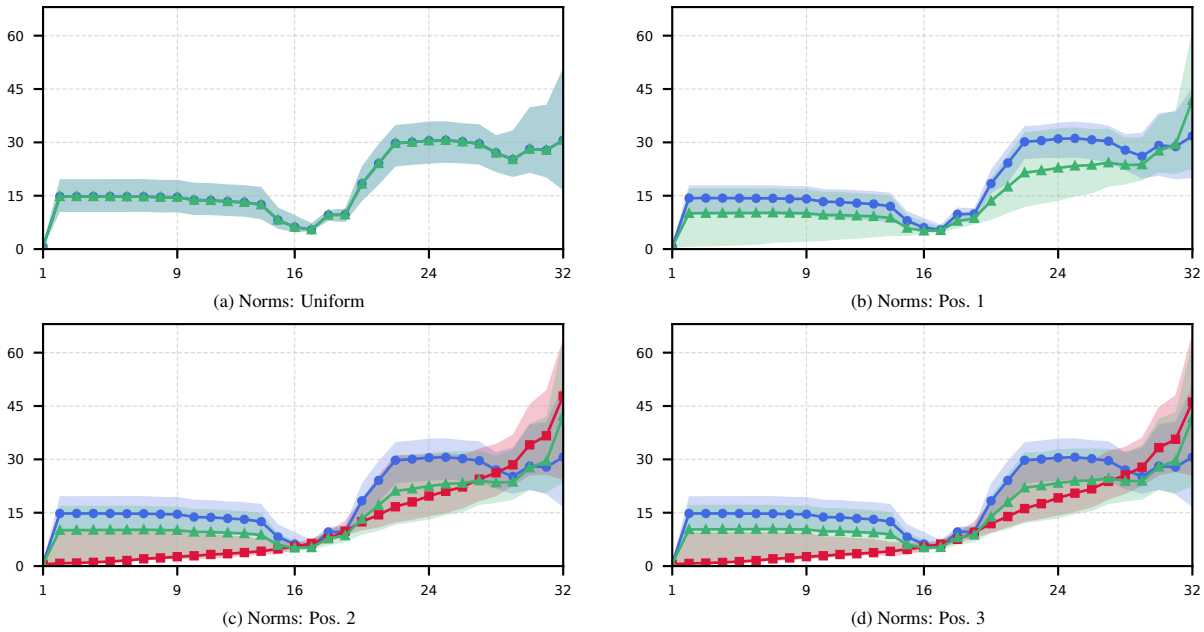



(e) Heatmap: Distinct token at pos. 2 (f) Activation: Distinct token at pos. 2




(g) Heatmap: Distinct token at pos. 3 (h) Activation: Distinct token at pos. 3

Figure 12: Results for sequences with repeated tokens. Each row compares the attention heatmap (left) and hidden-state activations (right) under the same token distribution pattern for Mistral-7B-v0.3 (Sample 1).




(a) Norms: Uniform (b) Norms: Pos. 1

(c) Norms: Pos. 2 (d) Norms: Pos. 3

Figure 13: Layer-wise averaged hidden-state norms in Mistral-7B-v0.3 for repeated-token sequences under different distinct-token positions (colors and markers as in Figure 5). The horizontal axis shows the layer index in every panel.

| Model | Position | Vanilla | Interv |
|---|---|---|---|
| Llama-2-7b-hf | 8 | 0.0000 | **0.8542** |
| | 32 | 0.0072 | **0.8217** |
| Llama-3.2-3B | 8 | 0.0026 | **0.7560** |
| | 32 | 0.0006 | **0.7406** |
| Mistral-7B-v0.3 | 8 | 0.0660 | **0.1347** |
| | 32 | 0.0000 | **0.0866** |
| Qwen2-7B | 8 | 0.0001 | **0.5324** |
| | 32 | 0.0001 | **0.4946** |
| pythia-1b | 8 | 0.0000 | **0.4885** |
| | 32 | 0.0000 | **0.4654** |

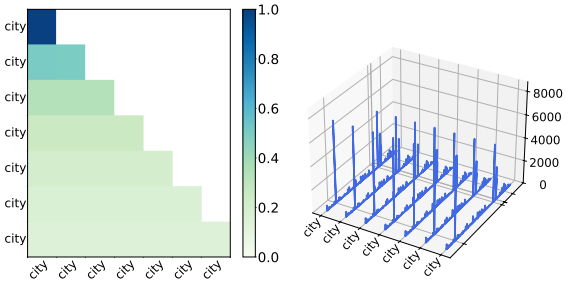
Table 11: $\mathrm{Sink}_t^{\epsilon}$ at the intervened token position for $t = 8$ and $t = 32$, comparing vanilla attention to the self-concentration intervention ($\alpha_{t,t} = 1.0$, all layers, w/o BOS), with $T = 64$.

| Model | Condition | Position | $\epsilon = 0.2$ | $\epsilon = 0.3$ | $\epsilon = 0.4$ | $\epsilon = 0.5$ |
|---|---|---|---|---|---|---|
| Llama-2-7b-hf | Vanilla | **1** | 0.9334 | 0.9305 | 0.9187 | 0.8819 |
| | | 15 | 0.0005 | 0.0001 | 0.0000 | 0.0000 |
| | | 16 | 0.0090 | 0.0073 | 0.0028 | 0.0001 |
| | | 17 | 0.0091 | 0.0071 | 0.0023 | 0.0000 |
| | RoPE Interv | **1** | 0.9341 | 0.9273 | 0.9100 | 0.8663 |
| Llama-3.2-3B | Vanilla | **1** | 0.9329 | 0.9218 | 0.8933 | 0.8118 |
| | | 15 | 0.0011 | 0.0008 | 0.0006 | 0.0004 |
| | | 16 | 0.0014 | 0.0012 | 0.0007 | 0.0004 |
| | | 17 | 0.0004 | 0.0003 | 0.0003 | 0.0002 |
| | RoPE Interv | **1** | 0.9317 | 0.9133 | 0.8713 | 0.7749 |
| Mistral-7B-v0.3 | Vanilla | **1** | 0.2258 | 0.0343 | 0.0033 | 0.0001 |
| | | 15 | 0.0195 | 0.0187 | 0.0172 | 0.0143 |
| | | 16 | 0.0386 | 0.0373 | 0.0344 | 0.0285 |
| | | 17 | 0.0194 | 0.0187 | 0.0172 | 0.0146 |
| | RoPE Interv | **1** | 0.3655 | 0.1525 | 0.1129 | 0.0960 |
| Qwen2-7B | Vanilla | **1** | 0.8677 | 0.8173 | 0.7439 | 0.6113 |
| | | 15 | 0.0005 | 0.0002 | 0.0000 | 0.0000 |
| | | 16 | 0.0004 | 0.0000 | 0.0000 | 0.0000 |
| | | 17 | 0.0004 | 0.0001 | 0.0000 | 0.0000 |
| | RoPE Interv | **1** | 0.8671 | 0.7931 | 0.6810 | 0.5290 |
| pythia-1b | Vanilla | **1** | 0.7433 | 0.7337 | 0.6889 | 0.5659 |
| | | 15 | 0.0002 | 0.0000 | 0.0000 | 0.0000 |
| | | 16 | 0.0001 | 0.0000 | 0.0000 | 0.0000 |
| | | 17 | 0.0001 | 0.0000 | 0.0000 | 0.0000 |
| | RoPE Interv | **1** | 0.7955 | 0.7547 | 0.7056 | 0.5749 |

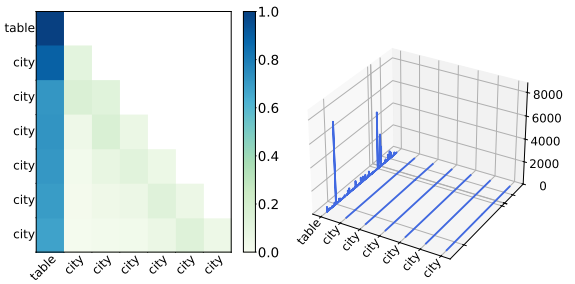
Table 12: Comparison of $\text{Sink}_j^{\epsilon}$ across various thresholds $\epsilon$ for RoPE ablation (both w/o BOS).

| Model | Condition | Position | $\epsilon = 0.2$ | $\epsilon = 0.3$ | $\epsilon = 0.4$ | $\epsilon = 0.5$ |
|---|---|---|---|---|---|---|
| Llama-2-7b-hf | Vanilla | **1 (First)** | 0.9334 | 0.9305 | 0.9187 | 0.8819 |
| | | 15 | 0.0005 | 0.0001 | 0.0000 | 0.0000 |
| | | 16 (Intervened) | 0.0090 | 0.0073 | 0.0028 | 0.0001 |
| | | 17 | 0.0091 | 0.0071 | 0.0023 | 0.0000 |
| | $\alpha_{16,16} = 1.0$ | **1 (First)** | 0.9323 | 0.9164 | 0.8248 | 0.5356 |
| | | 15 | 0.0002 | 0.0000 | 0.0000 | 0.0000 |
| | | 16 (Intervened) | 0.9248 | 0.8397 | 0.5308 | 0.0095 |
| | | 17 | 0.0072 | 0.0006 | 0.0000 | 0.0000 |
| Llama-3.2-3B | Vanilla | **1 (First)** | 0.9329 | 0.9218 | 0.8933 | 0.8118 |
| | | 15 | 0.0011 | 0.0008 | 0.0006 | 0.0004 |
| | | 16 (Intervened) | 0.0014 | 0.0012 | 0.0007 | 0.0004 |
| | | 17 | 0.0004 | 0.0003 | 0.0003 | 0.0002 |
| | $\alpha_{16,16} = 1.0$ | **1 (First)** | 0.9291 | 0.8785 | 0.6972 | 0.4172 |
| | | 15 | 0.0011 | 0.0008 | 0.0006 | 0.0003 |
| | | 16 (Intervened) | 0.9121 | 0.7424 | 0.4527 | 0.0120 |
| | | 17 | 0.0004 | 0.0003 | 0.0003 | 0.0002 |
| Mistral-7B-v0.3 | Vanilla | **1 (First)** | 0.2258 | 0.0343 | 0.0033 | 0.0001 |
| | | 15 | 0.0195 | 0.0187 | 0.0172 | 0.0143 |
| | | 16 (Intervened) | 0.0386 | 0.0373 | 0.0344 | 0.0285 |
| | | 17 | 0.0194 | 0.0187 | 0.0172 | 0.0146 |
| | $\alpha_{16,16} = 1.0$ | **1 (First)** | 0.1662 | 0.0198 | 0.0003 | 0.0000 |
| | | 15 | 0.0195 | 0.0182 | 0.0152 | 0.0117 |
| | | 16 (Intervened) | 0.1897 | 0.1509 | 0.0868 | 0.0304 |
| | | 17 | 0.0194 | 0.0184 | 0.0161 | 0.0129 |
| Qwen2-7B | Vanilla | **1 (First)** | 0.8677 | 0.8173 | 0.7439 | 0.6113 |
| | | 15 | 0.0005 | 0.0002 | 0.0000 | 0.0000 |
| | | 16 (Intervened) | 0.0004 | 0.0000 | 0.0000 | 0.0000 |
| | | 17 | 0.0004 | 0.0001 | 0.0000 | 0.0000 |
| | $\alpha_{16,16} = 1.0$ | **1 (First)** | 0.8485 | 0.7278 | 0.4770 | 0.2108 |
| | | 15 | 0.0004 | 0.0001 | 0.0000 | 0.0000 |
| | | 16 (Intervened) | 0.7533 | 0.5085 | 0.2174 | 0.0182 |
| | | 17 | 0.0005 | 0.0001 | 0.0000 | 0.0000 |
| pythia-1b | Vanilla | **1 (First)** | 0.7433 | 0.7337 | 0.6889 | 0.5659 |
| | | 15 | 0.0002 | 0.0000 | 0.0000 | 0.0000 |
| | | 16 (Intervened) | 0.0001 | 0.0000 | 0.0000 | 0.0000 |
| | | 17 | 0.0001 | 0.0000 | 0.0000 | 0.0000 |
| | $\alpha_{16,16} = 1.0$ | **1 (First)** | 0.7392 | 0.6702 | 0.4391 | 0.1769 |
| | | 15 | 0.0001 | 0.0000 | 0.0000 | 0.0000 |
| | | 16 (Intervened) | 0.6975 | 0.4703 | 0.1703 | 0.0054 |
| | | 17 | 0.0001 | 0.0000 | 0.0000 | 0.0000 |

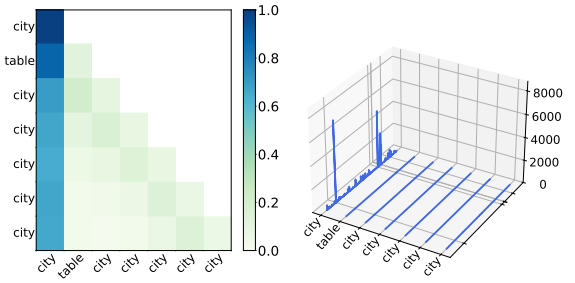
Table 13: Comparison of $\text{Sink}_j^{\epsilon}$ across various thresholds $\epsilon$ when intervention forces self-concentration of attention at position 16 (both w/o BOS).



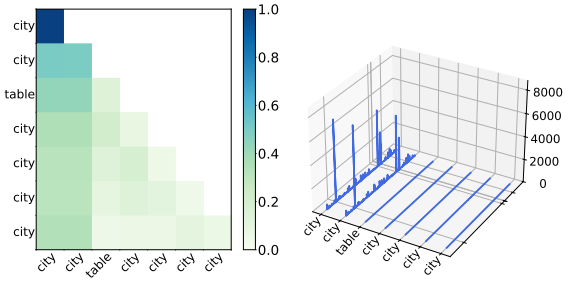

(a) Heatmap: Uniform sequence (b) Activation: Uniform sequence




(c) Heatmap: Distinct token at pos. 1 (d) Activation: Distinct token at pos. 1

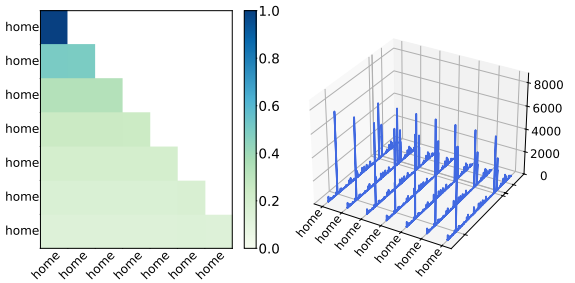



(e) Heatmap: Distinct token at pos. 2 (f) Activation: Distinct token at pos. 2

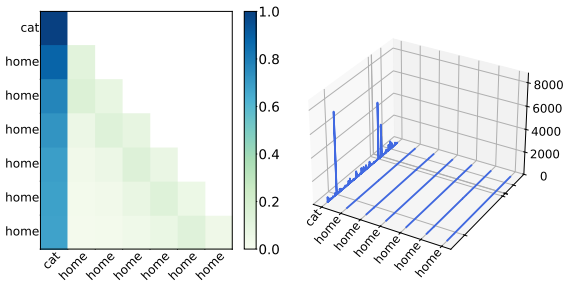



(g) Heatmap: Distinct token at pos. 3 (h) Activation: Distinct token at pos. 3

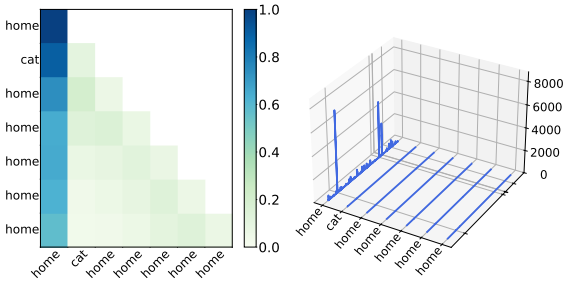
Figure 14: Results for sequences with repeated tokens. Each row compares the attention heatmap (left) and hidden-state activations (right) under the same token distribution pattern for Qwen2-7B (Sample 2).



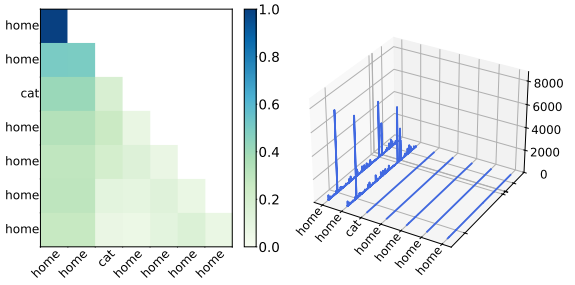

(a) Heatmap: Uniform sequence (b) Activation: Uniform sequence




(c) Heatmap: Distinct token at pos. 1 (d) Activation: Distinct token at pos. 1

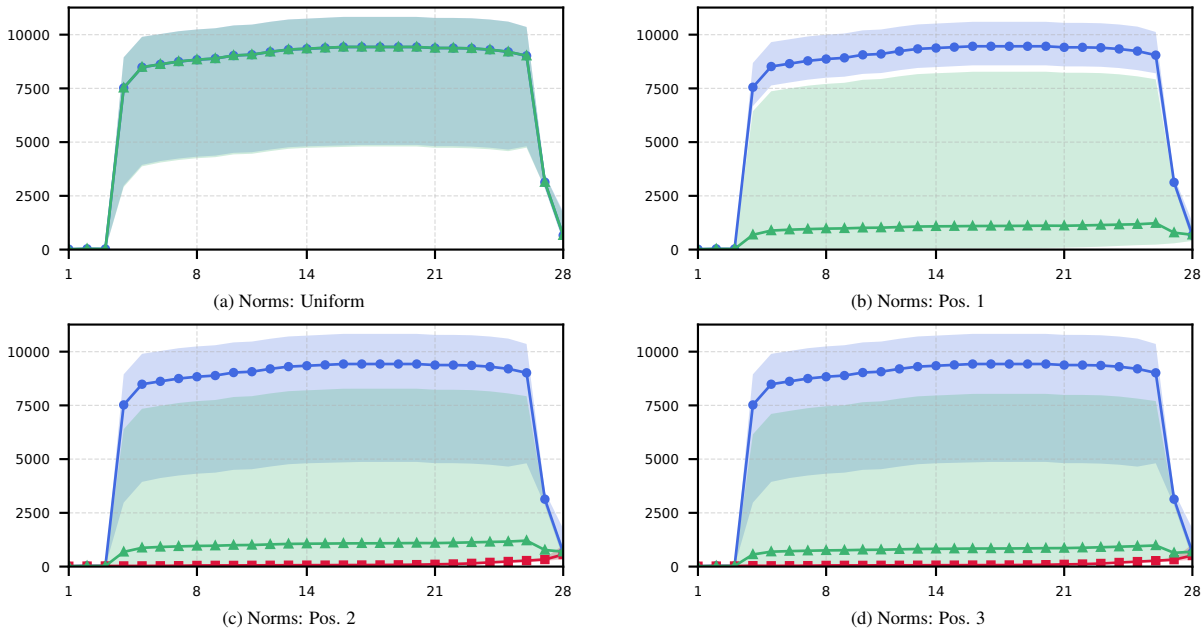



(e) Heatmap: Distinct token at pos. 2 (f) Activation: Distinct token at pos. 2




(g) Heatmap: Distinct token at pos. 3 (h) Activation: Distinct token at pos. 3

Figure 15: Results for sequences with repeated tokens. Each row compares the attention heatmap (left) and hidden-state activations (right) under the same token distribution pattern for Qwen2-7B (Sample 1).



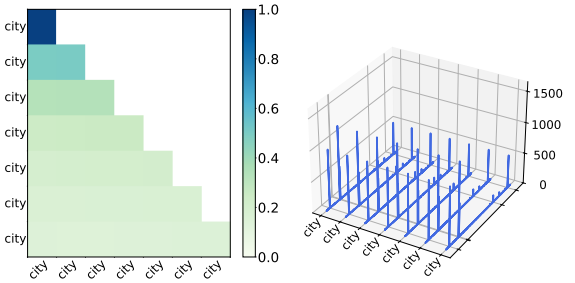

(a) Norms: Uniform

(b) Norms: Pos. 1

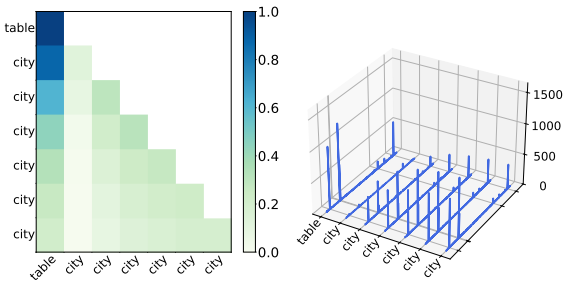
(c) Norms: Pos. 2

(d) Norms: Pos. 3

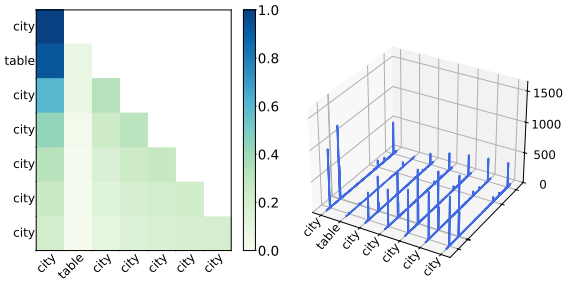
Figure 16: Layer-wise averaged hidden-state norms in Qwen2-7B for repeated-token sequences under different distinct-token positions (colors and markers as in Figure 5). The horizontal axis shows the layer index in every panel.



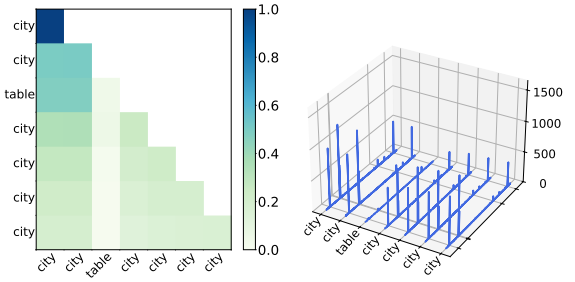

(a) Heatmap: Uniform sequence (b) Activation: Uniform sequence




(c) Heatmap: Distinct token at pos. 1 (d) Activation: Distinct token at pos. 1

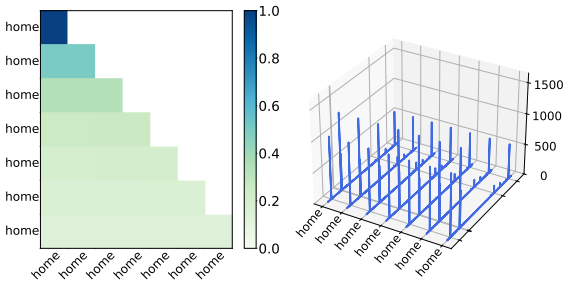



(e) Heatmap: Distinct token at pos. 2 (f) Activation: Distinct token at pos. 2

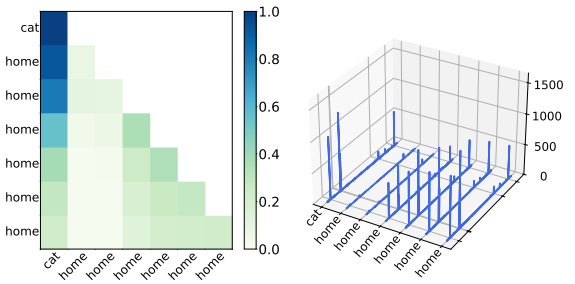



(g) Heatmap: Distinct token at pos. 3 (h) Activation: Distinct token at pos. 3

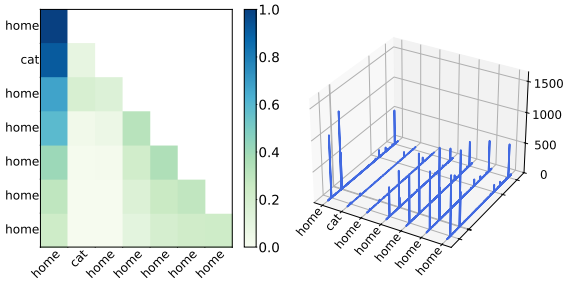
Figure 17: Results for sequences with repeated tokens. Each row compares the attention heatmap (left) and hidden-state activations (right) under the same token distribution pattern for pythia-1b (Sample 2).



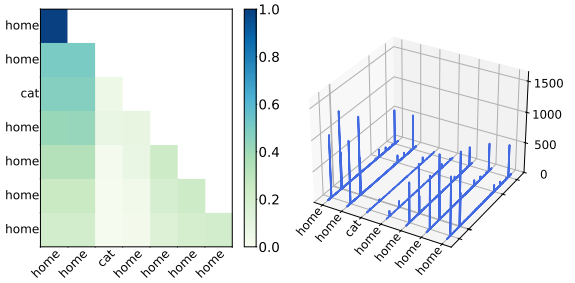

(a) Heatmap: Uniform sequence (b) Activation: Uniform sequence




(c) Heatmap: Distinct token at pos. 1 (d) Activation: Distinct token at pos. 1

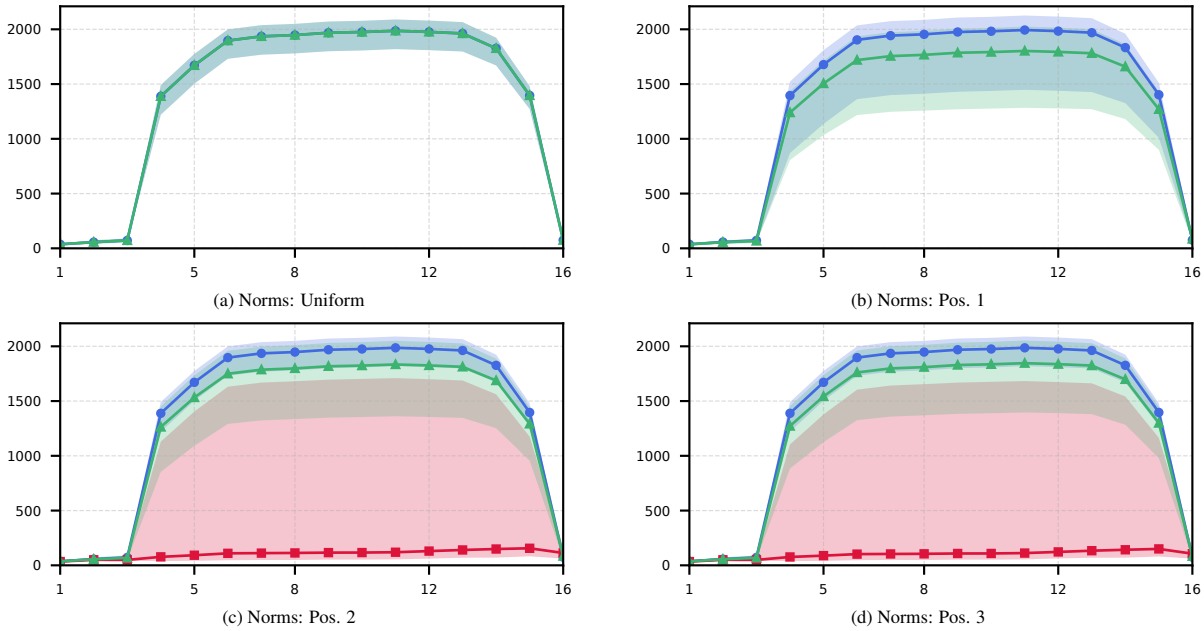



(e) Heatmap: Distinct token at pos. 2 (f) Activation: Distinct token at pos. 2




(g) Heatmap: Distinct token at pos. 3 (h) Activation: Distinct token at pos. 3

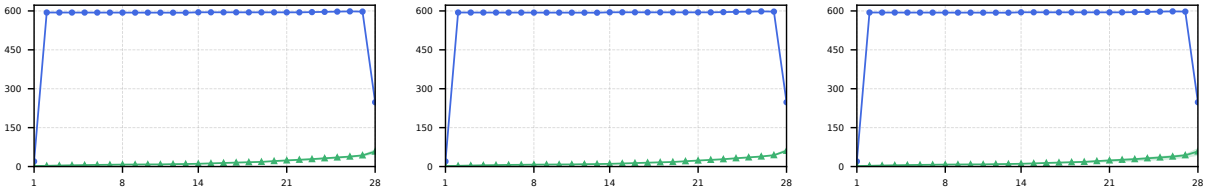
Figure 18: Results for sequences with repeated tokens. Each row compares the attention heatmap (left) and hidden-state activations (right) under the same token distribution pattern for pythia-1b (Sample 1).




(a) Norms: Uniform (b) Norms: Pos. 1

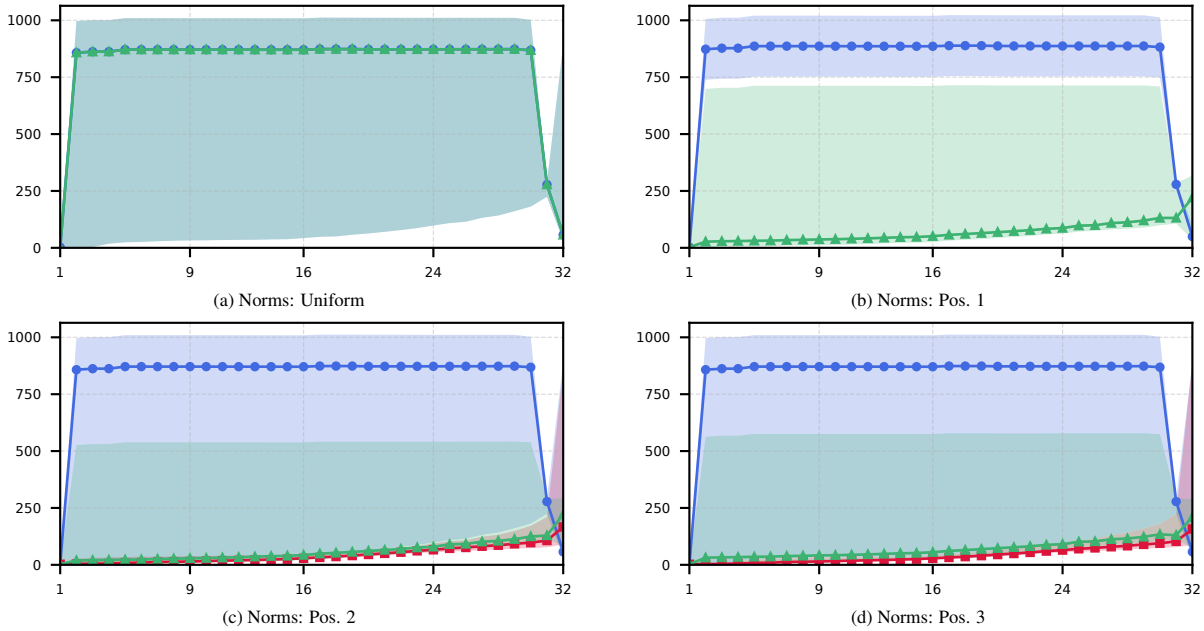
(c) Norms: Pos. 2 (d) Norms: Pos. 3

Figure 19: Layer-wise averaged hidden-state norms in pythia-1b for repeated-token sequences under different distinct-token positions (colors and markers as in Figure 5). The horizontal axis shows the layer index in every panel.




(a) Norms: WikiText (b) Norms: GSM8K (c) Norms: SlimPajama

Figure 20: Layer-wise averaged hidden-state norms in Llama-3.2-3B for an input length of 64 tokens across WikiText, GSM8K, and SlimPajama (baseline, no intervention). Blue circles and green triangles denote the first token and the mean over all other input tokens, respectively. The horizontal axis shows the layer index in every panel. Shaded bands show the sample min–max range.




(a) Norms: Uniform (b) Norms: Pos. 1

(c) Norms: Pos. 2 (d) Norms: Pos. 3

Figure 21: Layer-wise averaged hidden-state norms in Llama-2-7b-hf for repeated-token sequences of length 128 under different distinct-token positions (colors and markers as in Figure 10). The horizontal axis shows the layer index in every panel.