Attention Sink Mechanisms Are Not Universal Across Different Transformer Architectures

Anonymous ACL submission

Abstract

Transformers commonly exhibit an attention sink: disproportionately high attention to the first position. We study this behavior in GPT-2–style models with learned query biases and absolute positional embeddings. Combining analysis with targeted interventions, we find that the sink arises from the interaction among (i) a learned query bias, (ii) the first-layer transformation of the positional encoding and (iii) structure in the key projection. Together with observations of sinks in models without query biases or absolute positional embeddings (e.g., RoPE or ALiBi), this indicates that attention sinks do not arise from a single universal mechanism but instead depend on architecture. These findings inform mitigation of attention sink, and motivate broader investigation of sink mechanisms across architectures and training regimes.

1 Introduction

2 Related Work

[YRM: I didn't look at this part yet, waiting for you to finish it fist :)]

[YRM: As mentioned, this needs to be improved - By citing more, making this more concise (as concise as possible while including really vital info, and adding more in appendix if needed, like we did in our attention knockout paper)]Our investigation into the attention sink's origins builds upon four key areas of Transformer research: the methods for encoding positional information, the phenomenon of attention sink, and the recent discovery of massive activations functioning as implicit biases.

2.1 Positional Encoding in Transformers

By design, the self-attention mechanism has no inherent sense of token order. To address this, Transformers must be augmented with positional information. The original Transformer used fixed sinusoidal embeddings (Vaswani et al., 2017). Many models in the GPT family, including the GPT-2 model we investigate, use learned absolute positional embeddings - a vector for each position that is added to the token embedding at input. More recent architectures have introduced alternative methods, such as the LLaMA architecture (Touvron et al., 2023) which utilizes Rotary Positional Embeddings (RoPE) (Su et al., 2021), and Attention with Linear Biases (ALiBi) (Press et al., 2021), which is a key feature in models like BLOOM (BigScience Workshop, 2023).

2.2 The Attention Sink Phenomenon

Recent empirical work has identified a curious and robust phenomenon in auto-regressive language models termed the "attention sink" (Xiao et al., 2023). This refers to the tendency of models to allocate a significant portion of their attention to token(s) even when they are not semantically important, typically the first one(s) in the sequence. As Gu et al. (2025) demonstrate, this phenomenon is not an anomaly but emerges consistently during pre-training. Some studies have found the attention sink could have a negative impact on the achievable accuracy of LLMs (Yu et al., 2024). A similar phenomenon has been observed in LMMs (Large Multimodal Models) (Kang et al., 2025) and ViTs (Visual transformers) Feng and Sun (2025). Barbero et al. (2025) argue that attention sinks are a way for LLMs to avoid over-mixing and representational collapse. The main mechanism for attention sink that has been identified so far is the softmax normalization (Xiao et al. (2023), Gu et al. (2025), Zuhri et al. (2025)).

2.3 Massive Activations

The phenomenon of *massive activations*—where a tiny subset of coordinates exhibit orders-of-magnitude larger activations—has been identified and characterized across a variety of LLMs ((Sun et al., 2024), (Cancedda, 2024), (Lin et al., 2024)).

Crucially, these massive activations often appear at specific token positions, most notably the very first token of a sequence. (Sun et al., 2024) hypothesize that these activations function as bias terms that are learned by the model. A similar phenomenon has been observed in LMMs (Large Multimodal Models) (Kang et al., 2025).

3 Preliminaries

3.1 Attention mechanism

Let $X^{(i)} = [x_1^{(i)}, \ldots, x_n^{(i)}]$ denote the input to the attention layer i, where each $x_t^{(i)} \in \mathbb{R}^d$ is the hidden representation for position t (the input has been normalized by LayerNorm). We denote the projection matrices and optional biases by $W_q^{(i)}, W_k^{(i)}, W_v^{(i)} \in \mathbb{R}^{d\times d}$ and $b_Q^{(i)}, b_K^{(i)}, b_V^{(i)} \in \mathbb{R}^d$. Queries, keys, and values are computed as affine transformations of the input: $q_t^{(i)} = W_q^{(i)} x_t^{(i)} + b_Q^{(i)}$, $k_t^{(i)} = W_k^{(i)} x_t^{(i)} + b_K^{(i)}$, and $v_t^{(i)} = W_v^{(i)} x_t^{(i)} + b_V^{(i)}$. The biases $b_K^{(i)}$ and $b_V^{(i)}$ are learned parameters in some architectures (Vaswani et al., 2017) and omitted in others (Touvron et al., 2023). [YRM: Verify and add more specific citations for bias usage patterns]

For autoregressive generation, given query $q_t^{(i)}$ and keys $\{k_j^{(i)}\}_{j\le t}$, the attention weights are $\alpha_{tj} = \mathrm{softmax}_j\big((q_t^{(i)})^\top k_j^{(i)}/\sqrt{d}\big)$ [YRM: I changed the transpose here to be on the Q, please make sure that OK (this is better for later). Make sure this changes fit all other places/notations)]where the softmax is over valid positions $j \le t$. When layer indices are clear from context, or when layer indices can be arbitrary, we omit the superscript (e.g., W_k instead of $W_k^{(i)}$). Multi-head attention reshapes the queries and keys, runs h such heads in parallel and concatenates their outputs. We conducted the experiments on W_k and b_Q before reshaping for multiple heads. [Resolved: Add text about heads and how we are doing experiments for them troughout the paper]

3.2 Positional encoding

Attention layers are invariant to input permutations, lacking inherent awareness of token order. To address this limitation, Transformers incorporate positional information through various encoding schemes [YRM: Cite]. We focus on learned absolute positional encodings: a set of trainable vectors $\{p_i\}_{i=1}^L \subset \mathbb{R}^d$, where p_i corresponds to position i and L is the maximum sequence length. These are added to token embeddings at the input: $x_i^{(0)} = e_i + p_i$, where e_i is the token embedding for position i.

3.2.1 Effective positional encoding (EPE)

We define the *effective positional encoding* for position i as $\mathrm{EPE}_i = \mathrm{MLP}^{(1)}(p_i) + p_i$, where $\mathrm{MLP}^{(1)}$ denotes the first layer's feed-forward network applied to the raw positional encoding p_i, and the residual connection preserves the original positional signal. We term this "effective" because our analysis reveals that EPE_i captures the first-order effect of the positional information contributed by adding absolute position i after the first layer's processing (Experiments demonstrating this can be found in [YRM: Do an experiment to show this, put it in the appendix, and cref to it here])

4 Methodology and Results

First, we state the result of our analysis - a description of the mechanism underlying the attention sink in models with learnable query biases and absolute positional encodings. Then, through experimental analyses and causal interventions we provide evidence for our hypothesis.

4.1 Result: Mechanism behind the attention sink

Consider layer i. Before softmax (and scaling), the attention score from source position t to target position j is $s_{t\to j}^{(i)} = (q_t^{(i)})^\top k_j^{(i)}$, with $q_t^{(i)} = W_q^{(i)} x_t^{(i)} + b_Q^{(i)}$ and $k_j^{(i)} = W_k^{(i)} x_j^{(i)} + b_K^{(i)}$. Expanding gives

$$s_{t\to j}^{(i)} = (W_q^{(i)} x_t^{(i)})^\top (W_k^{(i)} x_j^{(i)}) + (W_q^{(i)} x_t^{(i)})^\top b_K^{(i)} + (b_Q^{(i)})^\top (W_k^{(i)} x_j^{(i)}) + (b_Q^{(i)})^\top b_K^{(i)}. \quad (157)$$

The third term, $\Delta_j^{(i)} \triangleq (b_Q^{(i)})^\top W_k^{(i)} x_j^{(i)}$, is a token-specific, source-agnostic shift: it raises or lowers the score for *all* sources t toward the same target j. This term represents the projection of token j's representation onto the direction $(b_Q^{(i)})^\top W_k^{(i)}$. We find that this bias term for the first token, $\Delta_1^{(i)}$, is conspicuously large in most deep layers, creating a strong prior to attend to position 1. We also find that the underlying reason for the large $\Delta_1^{(i)}$ is the effective positional encoding EPE_1. We find that EPE_1 has very large absolute values on a small set of coordinates (this is a known phenomenon called *massive activations* [YRM: Cite]) which are exactly those coordinates where $(b_Q^{(i)})^\top W_k^{(i)}$ has

the largest magnitude in almost all layers. This co-adaptation enables EPE_1 to dramatically amplify $\Delta_1^{(i)}$, yielding an attention sink at the first position. [Backlog: Can we add a diagram to illustrate this? This sounds time consuming but really really helpful since this passage turned out to be somewhat dense. Let's wait to see how much space we have before doing this]

4.2 Empirical Validation

We validate our proposed mechanism through three complementary analyses on GPT-2, followed by causal interventions that confirm the necessity of each component described in section 4.1. In section 4.2.1 we show that $\Delta_1^{(i)}$ is conspicuously large relative to other positions across multiple layers. Having established this, we investigate its underlying cause and show in section 4.2.2 that $W_k^{(i)}\text{EPE}_1$ exhibits strong alignment with vector $b_Q^{(i)}$ in deep layers. In section 4.2.3 we establish that EPE_1 exhibits massive activations precisely at coordinates where the bias projection $b_Q^{(i)}W_k^{(i)}$ has high magnitude. Finally, in section 4.2.4 we use causal interventions to verify that disrupting any component abolishes the sink while transplanting components transfers it to new positions.

4.2.1 Bias Term Magnitude Analysis

We first verify that the bias term $\Delta_j^{(i)} = (b_Q^{(i)})^\top W_k^{(i)} x_j^{(i)}$ is indeed anomalously large for the first position. We plot histograms of $\Delta_j^{(i)}$ across all positions j in multiple layers and find that $\Delta_1^{(i)}$ consistently forms a distinct outlier. Figure 2 shows this pattern for layer 10, where $\Delta_1^{(i)}$ is substantially larger than all other positions (For all the layers, see Appendix A). [Resolved: Two things: One, we need to say that this is only an example, but that this is true for many layers. Either by giving more examples in the appendix, or running a more extensive experiment, running this on more layers and say something like "it's the highest bias by a factor of more than 2 std in some % of layers" this is much more convincing than some possibly cherry picked layer, and luckily we don't need to cherry pick.][Resolved: Secondly - is this true even after normalization? if so, add a footnote here to say something about this. People might see these numbers and think that they are too high to be logits, we should say something about this. (not top priority if you don't have the time, worst case we'll fix it if someone asks about this.)]

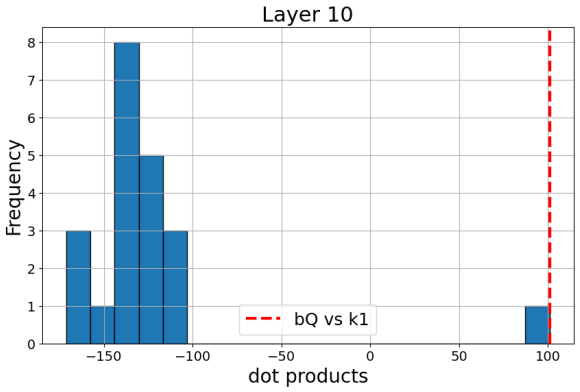


Figure 1: Distribution of bias terms $\Delta_j^{(10)}$ across positions. The first-position term $\Delta_1^{(10)}$ (red) centers at ≈ 100, while all other positions (blue) center at ≈ -140, demonstrating a learned preference for the first token.

4.2.2 EPE-Bias Projection Alignment

Having established the magnitude of $\Delta_1^{(i)}$, we investigate its underlying cause. Since $x_1^{(i)}$ contains both token and positional information, it remains to disentangle which of the two is responsible for the large $\Delta_1^{(i)}$. To this end, we examine the alignment between $W_k^{(i)}\text{EPE}$ and the query bias $b_Q^{(i)}$. We compute cosine similarity between these vectors across layers, comparing the first position against all others. Figure 1 demonstrates that $W_k^{(10)}\text{EPE}_1$ exhibits strong positive alignment with $b_Q^{(10)}$, while other positions cluster near zero (For all the layers, see appendix B). [Resolved: again regarding extending this to either have more examples in appendix or have a "meta" metric for more layers] [Backlog: If we can also do that for $x_1^{(i)} - \text{EPE}_1$ and show that this isn't aligned that would be great for this paragraph (we can put it in the appendix and just write it casually. This is not top priority at all.)]

4.2.3 Coordinate-Level Structural Analysis

Massive coordinates of EPE_1 should coincide with coordinates favored by the bias projection. Let $\gamma^{(i)} = (W_k^{(i)})^\top b_Q^{(i)} \in \mathbb{R}^d$; its entry $\gamma^{(i)}[d]$ measures how strongly input coordinate d contributes to the source-agnostic shift $\Delta_j^{(i)}$. We expect large $|\text{EPE}_1[d]|$ exactly where $|\gamma^{(i)}[d]|$ is large.

We select top-$|\text{EPE}_1|$ coordinates (see Appendix C for deta) [Resolved: how do you identify this? I think its best to say we do this by hand

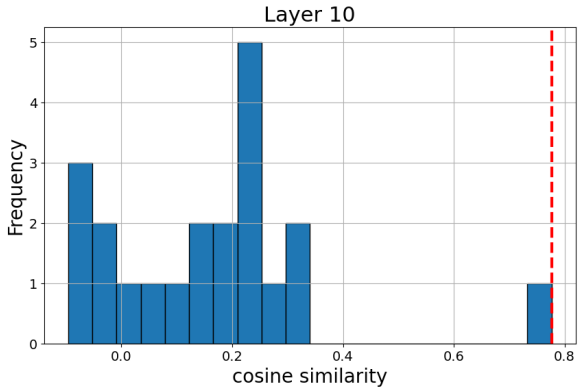


Figure 2: Cosine similarity between query bias $b_Q^{(10)}$ and $W_k^{(10)}$EPE. $W_k^{(10)}\text{EPE}_1$ (red) shows strong positive alignment (≈ 0.7), while other positions (blue) cluster near zero.

Layer	Baseline (rand)	d=138	d=447
layer 7	1.12±2.701	12.453	18.17
layer 9	1.23±3.224	17.846	28.01
layer 11	1.403±4.000	27.547	27.691

Table 1: $\gamma^{(i)} = (W_k^{(i)})^\top b_Q^{(i)}$ at coordinates where EPE_1 has massive activations (dims 138, 447) versus the mean of all of the columns. Massive-EPE_1 coordinates have substantially higher values across layers, confirming that EPE_1 is large exactly where the bias projection is large.

a add a link to the appendix and add a histogram there of EPE_1 showing that it's really clear which values are big]. For each such coordinate d, we compare $|\gamma^{(i)}[d]|$ against the mean of the all of the columns. [Resolved: Edit this to instead measure $\gamma^{(i)}[d]$ instead of the cos similarity to fit our new notation] Table 1 shows that the two massive coordinates (d=138, 447) are substantially higher than baseline across layers 7, 9, and 11, confirming that EPE_1 is large exactly where the bias projection is large (The rest of the layers are in D)

4.2.4 Causal Interventions

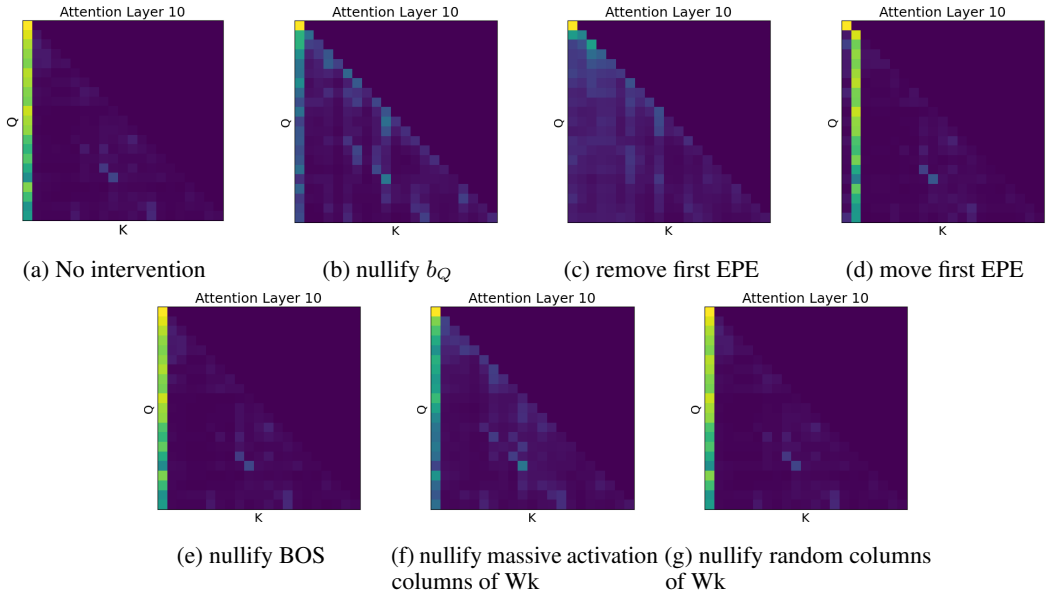
To establish causality beyond correlation, we perform targeted interventions on each mechanism component during forward passes to test necessity (removing a component) and sufficiency (transplanting it) of each component. For all the layers see Appendix E.

[Resolved: Now to the annoying part - we need to somehow say that this is robust and happens in many layers, so say we add more layers in the appendix and do this, it's ok if it's not as robust this experiment but people expect this info to be present. Please go after this over all the paper and think where else we need to add more experiments to the appendix.]

- **Intervention 1 — Nullify b_Q (query bias is necessary).** Set b_Q to zero; the sink substantially diminishes (fig. 3b), showing that b_Q is necessary for the large first-token contribution (for all layers, see Appendix E.2).
- **Intervention 2 — Replace EPE_1 (specificity of the positional signal).** Swap EPE_1 with another position's EPE; the first-position sink disappears (fig. 3c), indicating that EPE_1 is critical to induce a sink (for all layers, see Appendix E.3).
- **Intervention 3 — Moving EPE_1 induces a sink at the new token (sufficiency).** We transplant EPE_1 from position 1 to position 2 (and give position 1 a different EPE). A strong sink forms at position 2 (fig. 3d), demonstrating that EPE_1 is sufficient to elicit a sink at the new location. (for all layers, see Appendix E.4).
- **Intervention 4 — BOS token does not drive the sink.** We zero the BOS token embedding before adding positional signals. The sink persists (fig. 3e), ruling out the embedding of the BOS token as a main driver of the sink. (for all layers, see Appendix E.5).
- **Intervention 5 — Zero W_k at bias-projection coordinates (structural pathway is necessary).** Zero W_k columns at massive-EPE_1 coordinates compared to zeroing W_k columns at random coordinates; only the prior case substantially reduces the sink (fig. 3f, fig. 3g), confirming that these specific coordinates are core drivers for translating EPE_1 into the attention bias. (for all layers, see Appendix E.6 and E.7).

5 Conclusions

Attention sinks are not universal. In architectures with learned query biases and absolute positional encodings, our analyses and interventions indicate that the sink is implemented through an interaction between (i) a learned query bias, (ii) the first-layer transformation of positional information that yields a high-magnitude effective positional encoding at the first position (EPE_1), and (iii) structure in the key projection aligned with the large-magnitude coordinates of EPE_1. Together with observations of sinks in models without these components (e.g



(a) No intervention (b) nullify b_Q (c) remove first EPE (d) move first EPE

(e) nullify BOS (f) nullify massive activation columns of Wk (g) nullify random columns of Wk

Figure 3: Comparison of attention maps under different interventions. (a) no intervention; (b) intervention 1: nullify b_Q; (c) intervention 2: remove the learned EPE at position 1 and add a different EPE (the second); (d) intervention 3: transplant the learned EPE to another position (the second). (e) intervention 4: nullify BOS token embedding. intervention 5: (f) nullify massive activation columns of Wk. (g) nullify random columns of Wk. [YRM: this is absolutely great, but *must* take up less space. Is there a way to make it smaller without hurting too much? (just making the pic smaller will make the x and y text too small, so we need something more delicate)]

with alternative positional schemes (e.g., RoPE or ALiBi; [YRM: cite]) or without a learned query bias [YRM: cite]), these results argue against a single universal mechanism for attention sink

Implications The lack of a single universal mechanism for attention sink hints that attention sink are an optimization-friendly attractor rather than a single architectural quirk: when multiple representational routes are available, training can discover a circuit that implements a sink. Consequently, naive, component-wise regularization is unlikely to eliminate the phenomenon—penalizing one element (e.g., shrinking the query bias) can be absorbed by alternative pathways. We offer these findings as a basis for developing and evaluating mitigations.

6 Limitations

6.1 Scope across architectures and scales

Our analyses focus on a GPT-2–style model with learned query biases and absolute positional encodings. The broader Transformer ecosystem includes architectures that omit such biases or use alternative positional schemes (e.g., RoPE, ALiBi). We do not establish whether the same circuit forms in those settings, nor whether the EPE–W_k–b_Q interaction generalizes unchanged. In addition, GPT-2 is small by contemporary standards; with scale, the mechanism could strengthen, fragment into multiple pathways, or be replaced by different circuits.

6.2 Learning dynamics

We provide a post-hoc, static analysis of a trained checkpoint. We do not track when the circuit emerges during pre-training, which gradients give rise to it, or whether intermediate snapshots exhibit qualitatively different pathways. Train-time causality—e.g., whether specific regularizers prevent the circuit from forming—remains outside our scope.

6.3 Mechanism vs. function

Our contribution is mechanistic: we explain *how* an attention sink can be implemented in the studied architecture. We do not claim a definitive *functional* rationale for *why* such a sink is beneficial or harmful across tasks. Establishing the downstream utility or cost of the sink, and the conditions under which it is selected by optimization, is left for future work.

References

Federico Barbero, 'Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael M. Bronstein, Petar Velivckovi 'c, and Razvan Pascanu. 2025. Why do llms attend to the first token? *ArXiv*, abs/2504.02732.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100. ArXiv:2211.05100v4.

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

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

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. *arXiv preprint arXiv:2410.10781*.

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. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*. ArXiv preprint arXiv:2406.01721v3.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3533–3547.

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

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

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, Brian Fuller, and 48 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288. ArXiv:2307.09288v2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS (NIPS) 2017*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Layer	**Baseline**	d=138	d=447
layer 1	4.47±22.226	12.116	11.064
layer 2	2.8±6.62	8.065	24.468
layer 3	1.717±5.826	10.178	23.047
layer 4	1.657±5.02	18.198	17.149
layer 5	1.561±4.618	3.072	23.854
layer 6	0.86±1.59	5.644	6.142
layer 8	1.404±3.546	19.806	28.01
layer 10	1.313±3.616	23.131	28.42
layer 12	1.145±2.65	4.5	13.59

Table 2: $\gamma^{(i)} = (W_k^{(i)})^\top b_Q^{(i)}$ at coordinates where EPE_1 has massive activations (dims 138, 447) versus the mean of all of the columns. Massive-EPE_1 coordinates have substantially higher values across layers, confirming that EPE_1 is large exactly where the bias projection is large.

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. *ArXiv*, abs/2406.15765.

Zayd Muhammad Kawakibi Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. 2025. Softpick: No attention sink, no massive activations with rectified softmax. *ArXiv*, abs/2504.20966.

A

B EPE Bias

C

D

E

E.1

E.2

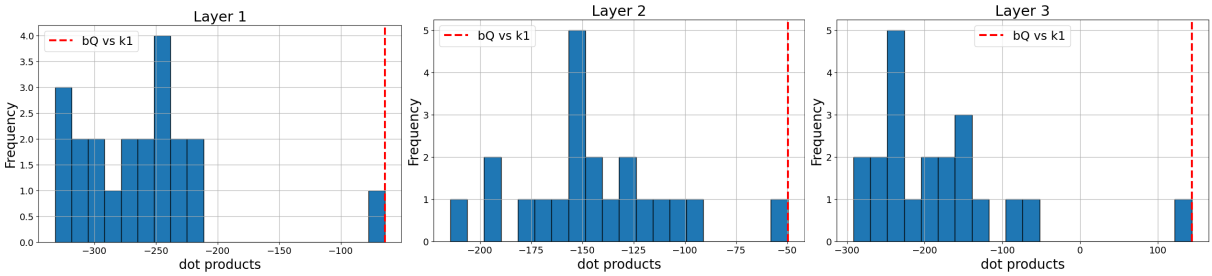
E.3

E.4

E.5

E.6

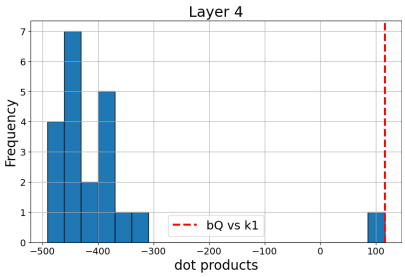
E.7



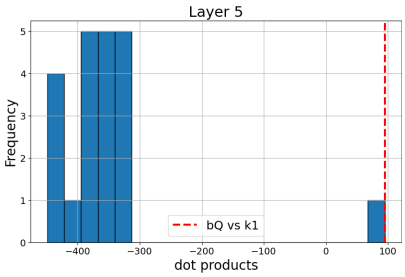
(a) layer 1

(b) layer 2

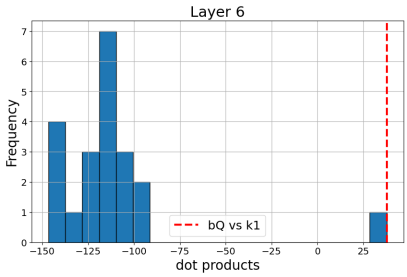
(c) layer 3



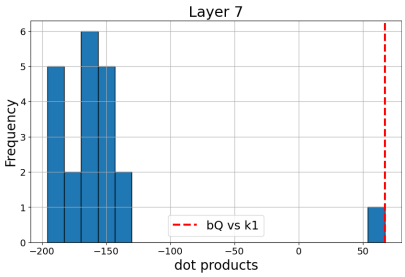
(d) layer 4



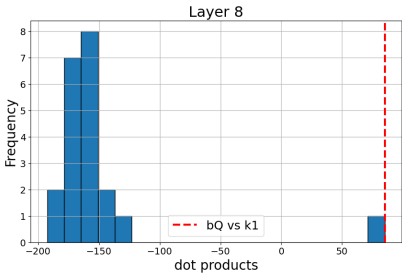
(e) layer 5



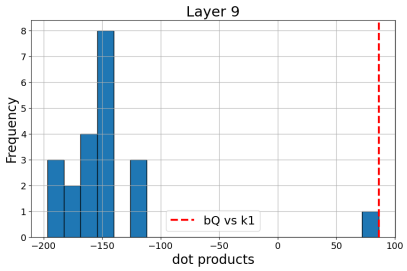
(f) layer 6



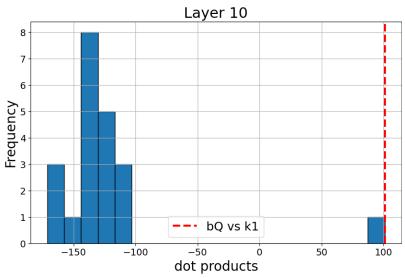
(g) layer 7



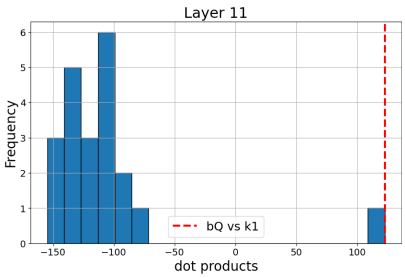
(h) layer 8



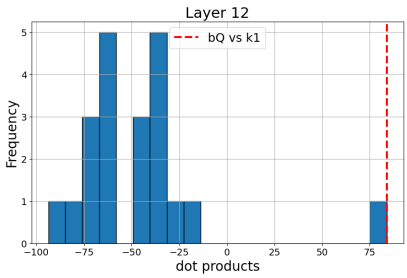
(i) layer 9



(j) layer 10

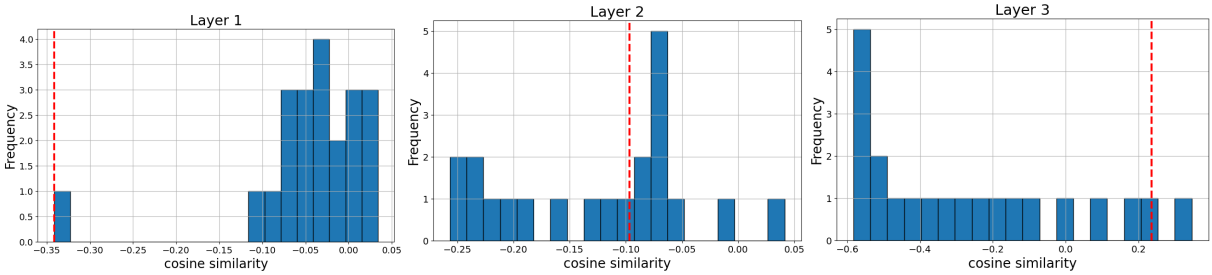


(k) layer 11



(l) layer 12

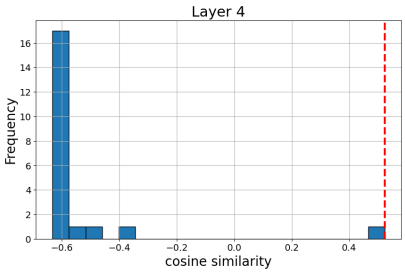
Figure 4: Distribution of bias terms Δ_j across positions for all of the layers



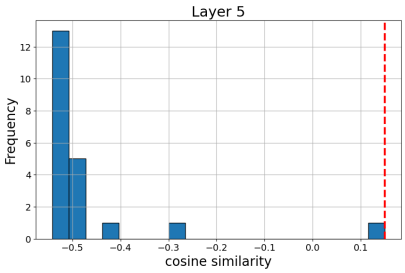
(a) layer 1

(b) layer 2

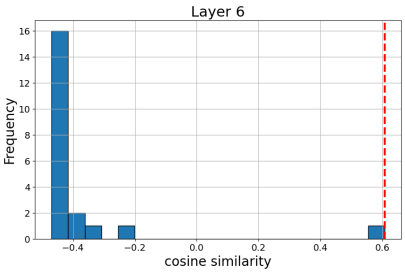
(c) layer 3



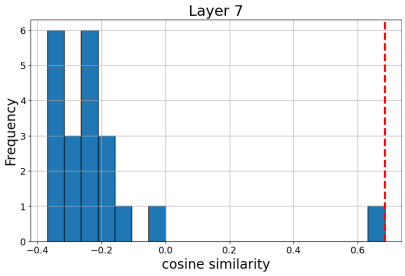
(d) layer 4



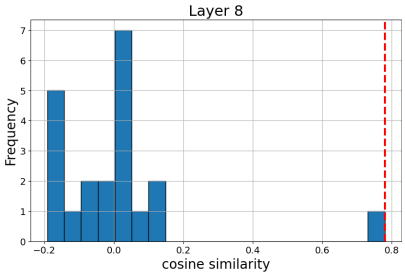
(e) layer 5



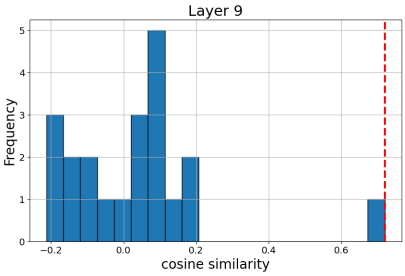
(f) layer 6



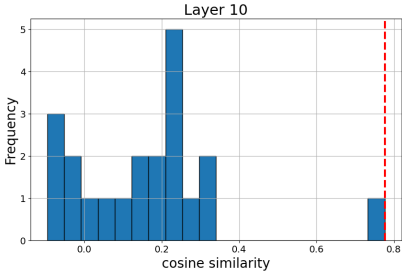
(g) layer 7



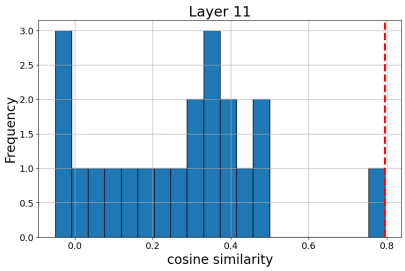
(h) layer 8



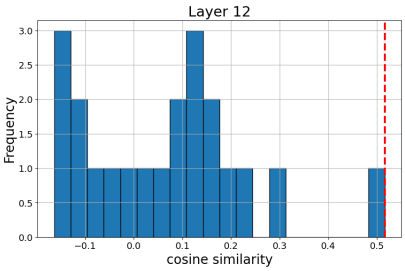
(i) layer 9



(j) layer 10



(k) layer 11



(l) layer 12

Figure 5: Cosine similarity between query bias $b_Q^{(10)}$ and $W_k^{(10)}$EPE.

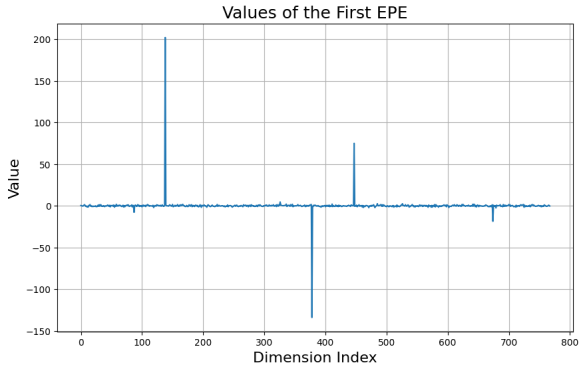
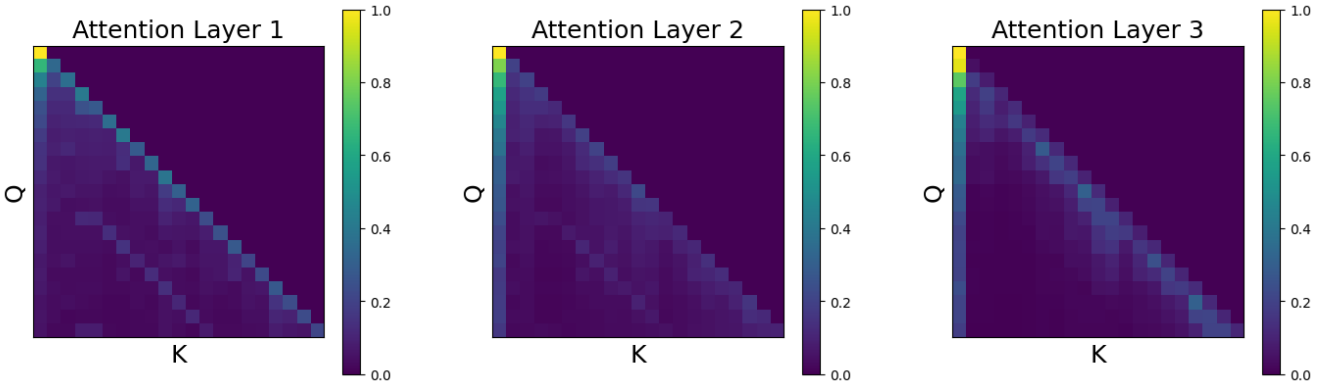


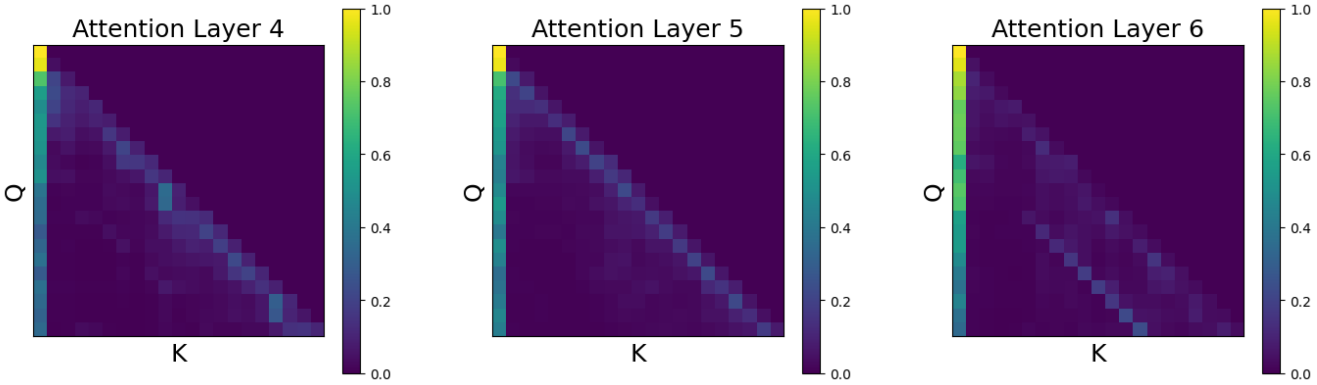
Figure 6: Values of the EPE_1 vector for the first token. Most coordinates remain near zero, but a small number exhibit extremely large positive or negative magnitudes, which we refer to as "massive activations." These amplified dimensions are the ones later aligned with W_k and b_Q in our analysis.



(a) layer 1

(b) layer 2

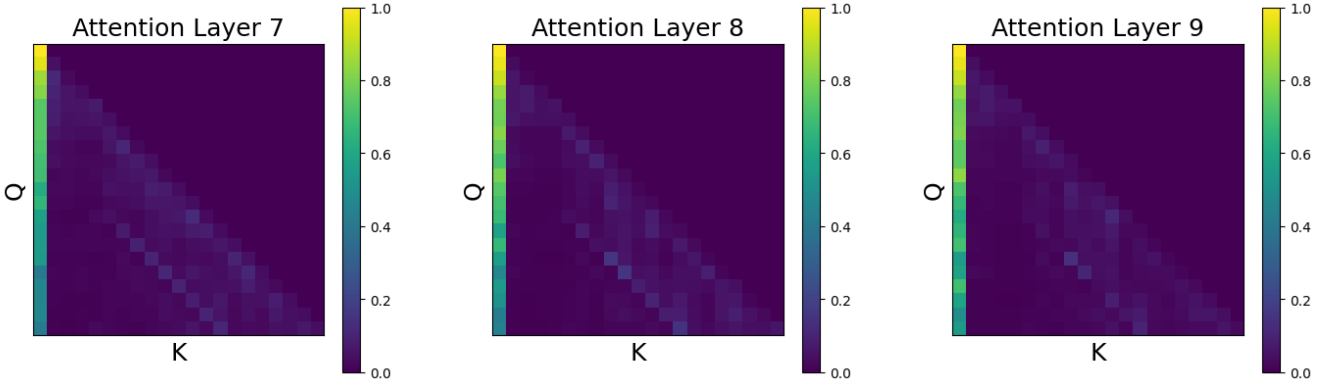
(c) layer 3



(d) layer 4

(e) layer 5

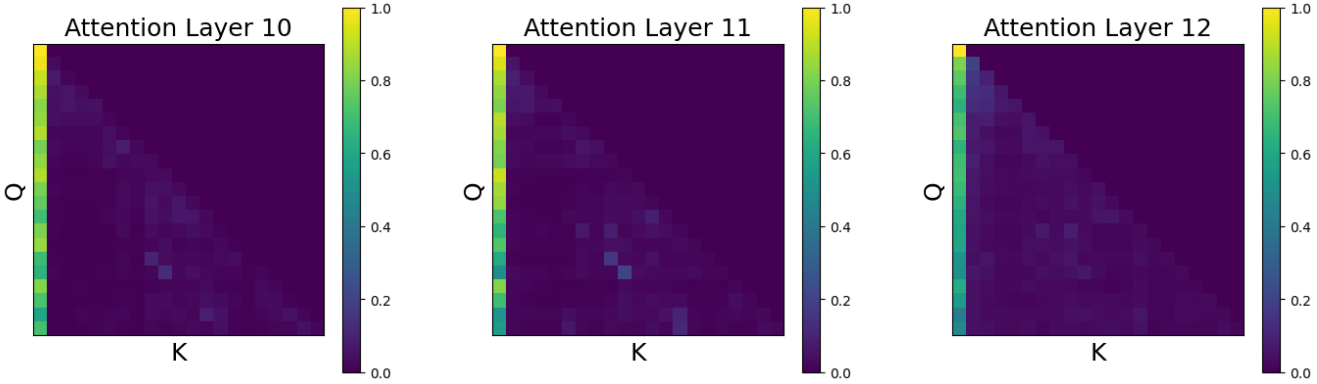
(f) layer 6



(g) layer 7

(h) layer 8

(i) layer 9

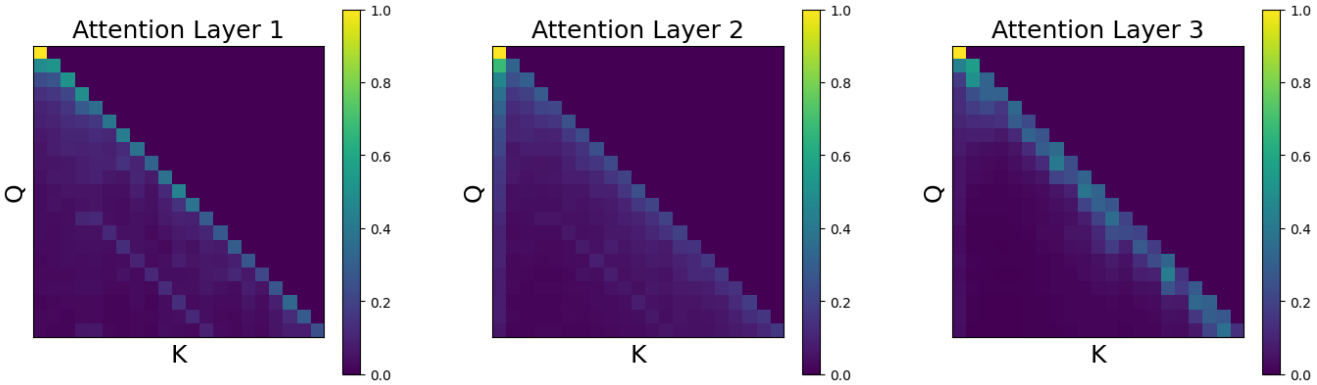


(j) layer 10

(k) layer 11

(l) layer 12

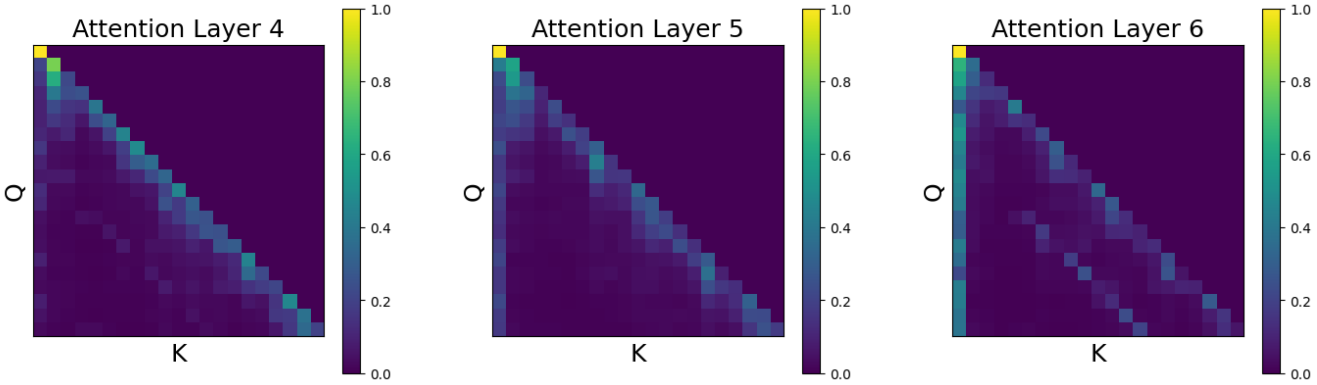
Figure 7: Attention maps for all layers with no intervention. There is a visible attention sink in most layers



(a) layer 1

(b) layer 2

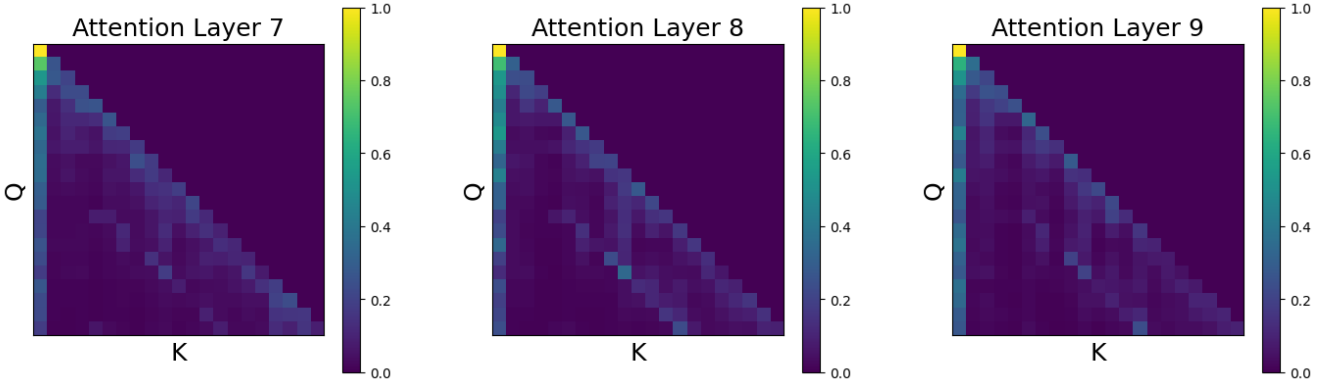
(c) layer 3



(d) layer 4

(e) layer 5

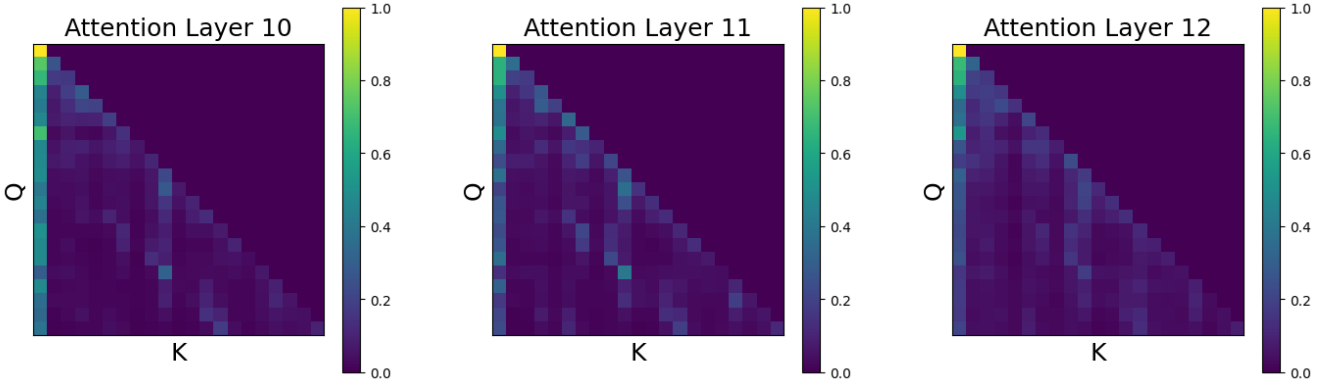
(f) layer 6



(g) layer 7

(h) layer 8

(i) layer 9

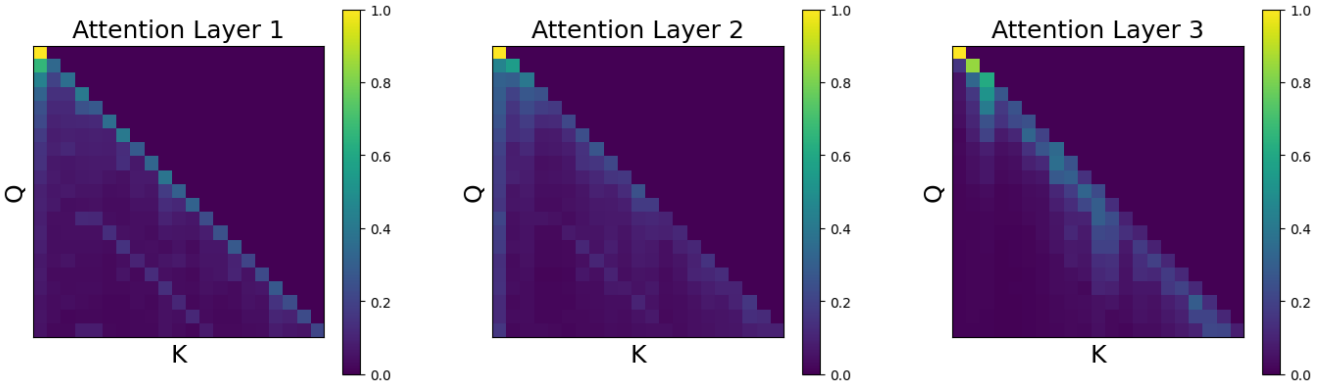


(j) layer 10

(k) layer 11

(l) layer 12

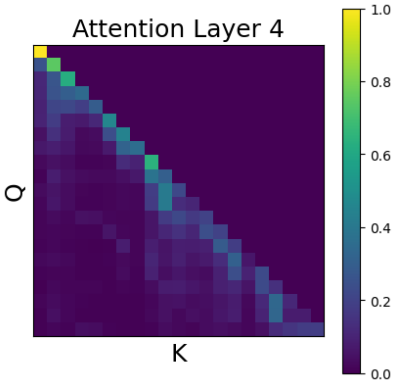
Figure 8: Intervention 1: Attention maps for all layers with nullufying b_Q. The attention sink is significantly resuced across layers



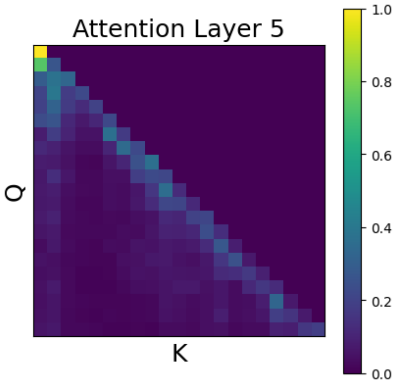
(a) layer 1

(b) layer 2

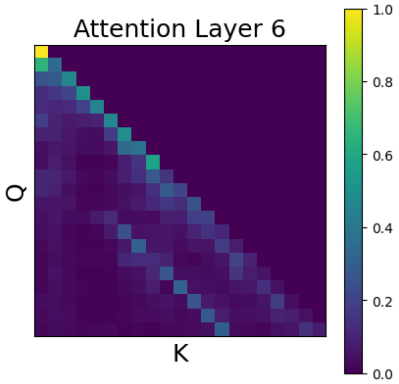
(c) layer 3



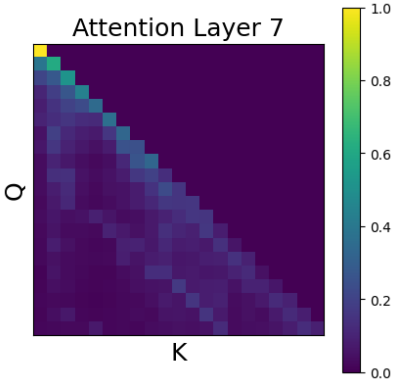
(d) layer 4



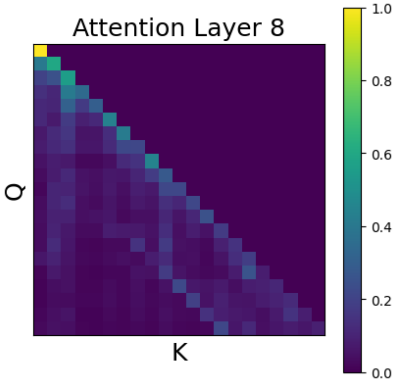
(e) layer 5



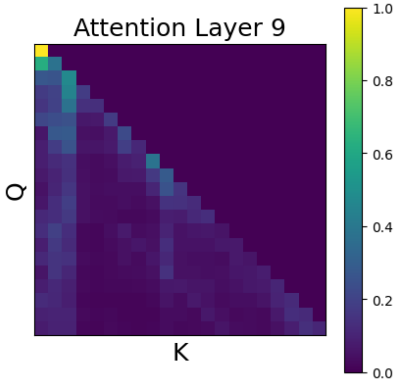
(f) layer 6



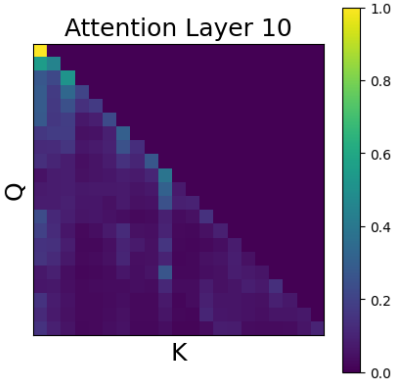
(g) layer 7



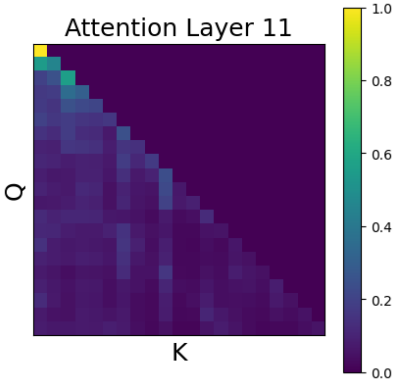
(h) layer 8



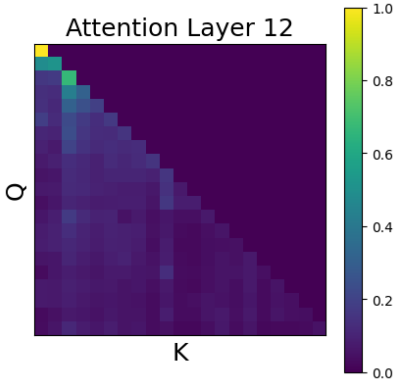
(i) layer 9



(j) layer 10

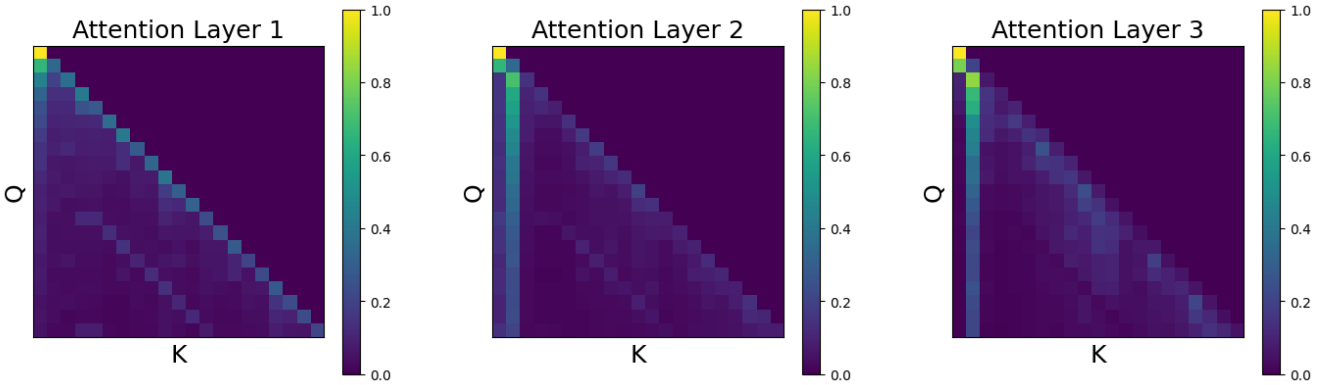


(k) layer 11



(l) layer 12

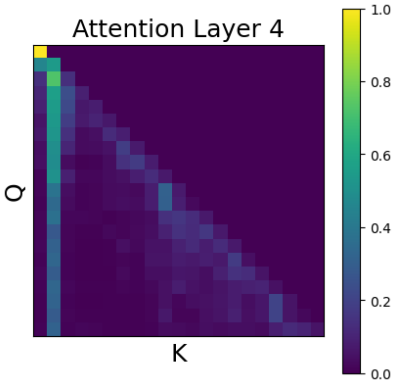
Figure 9: Intervention 2: Attention maps for all layers with swaping EPE_1 with another position's EPE; the first-position sink disappears



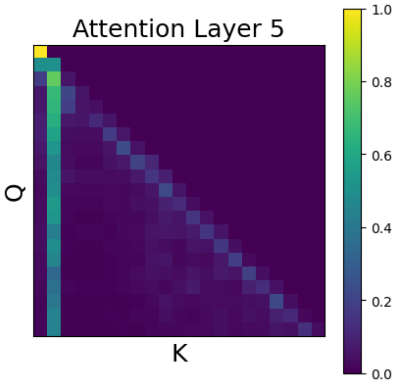
(a) layer 1

(b) layer 2

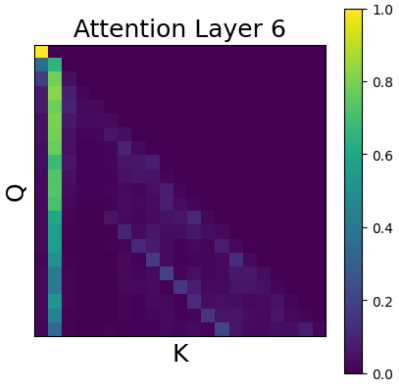
(c) layer 3



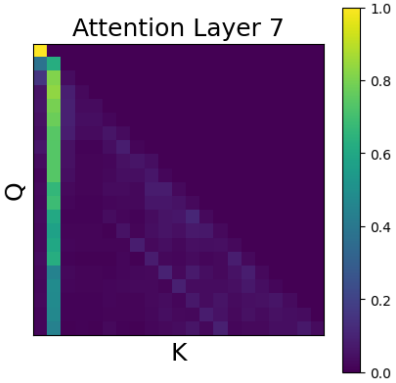
(d) layer 4



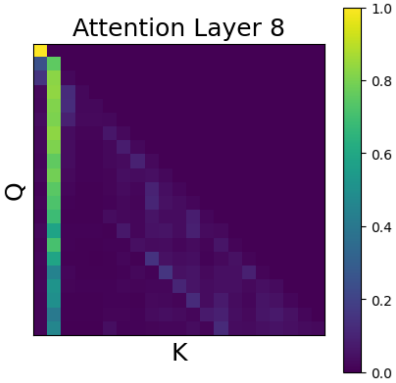
(e) layer 5



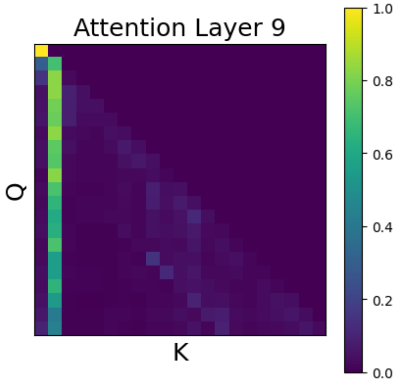
(f) layer 6



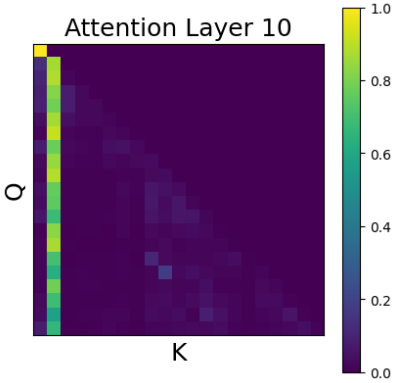
(g) layer 7



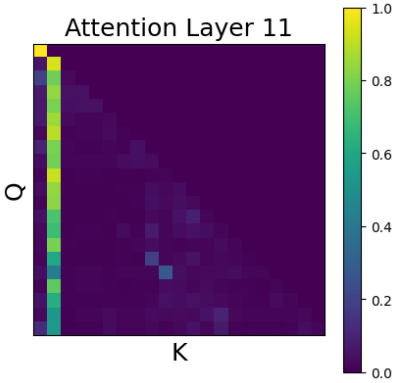
(h) layer 8



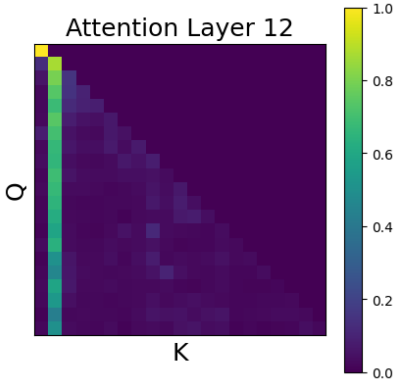
(i) layer 9



(j) layer 10

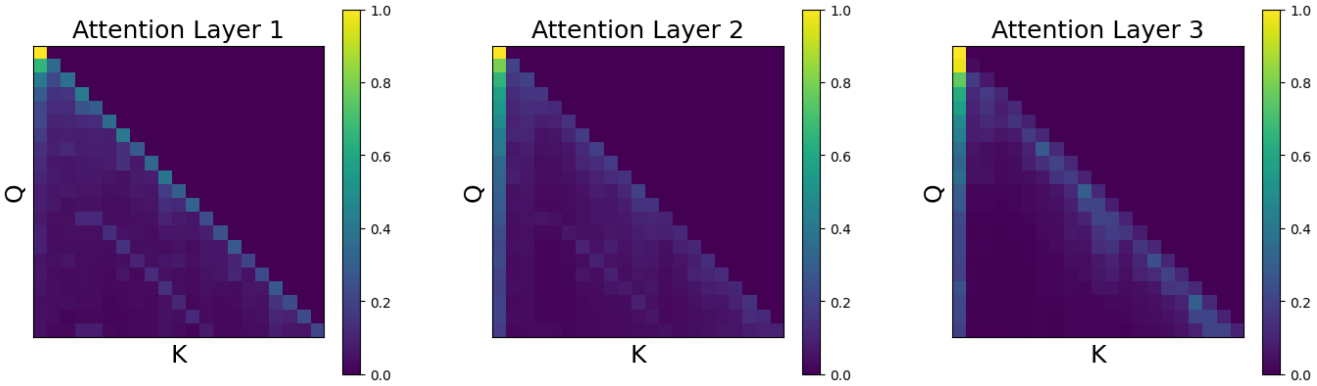


(k) layer 11



(l) layer 12

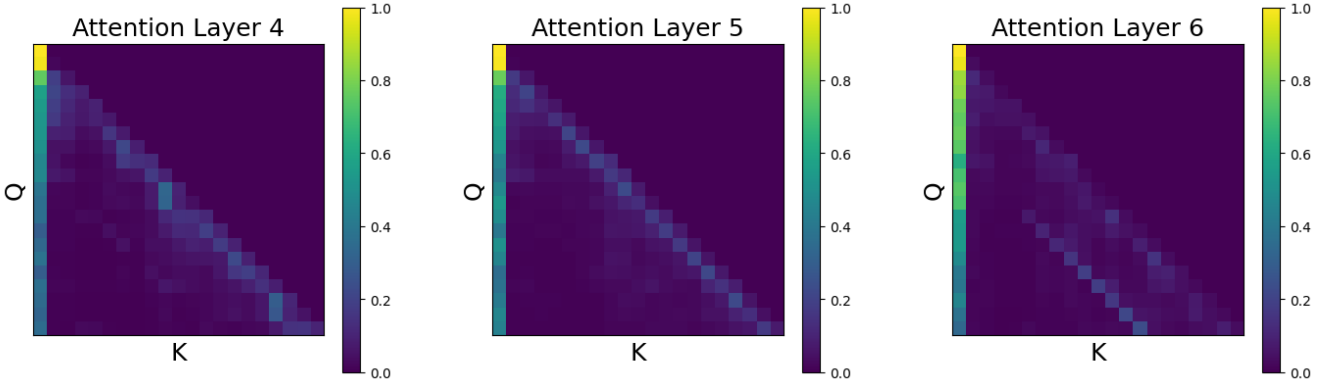
Figure 10: Intervention 3: Attention maps for all layers with transplanting EPE_1 from position 1 to position 2 (and give position 1 a different EPE). A strong sink forms at position 2



(a) layer 1

(b) layer 2

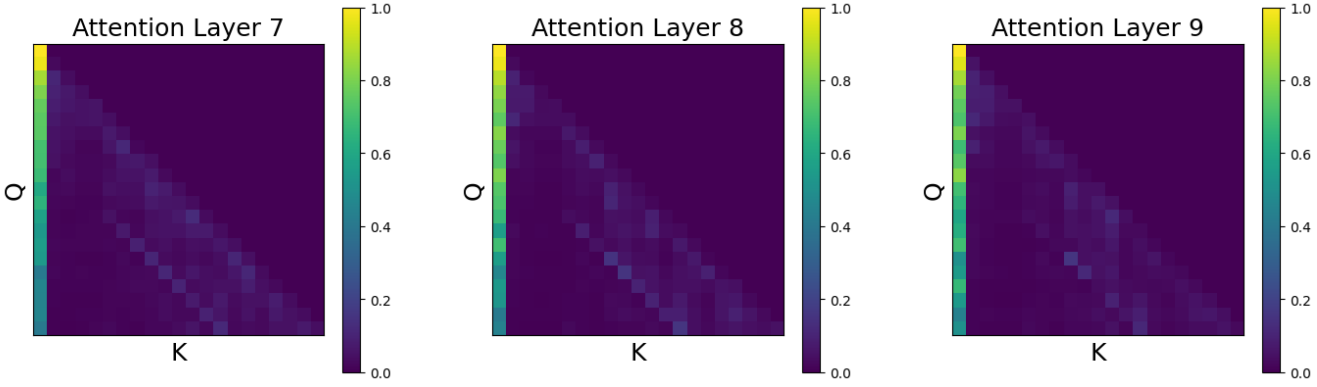
(c) layer 3



(d) layer 4

(e) layer 5

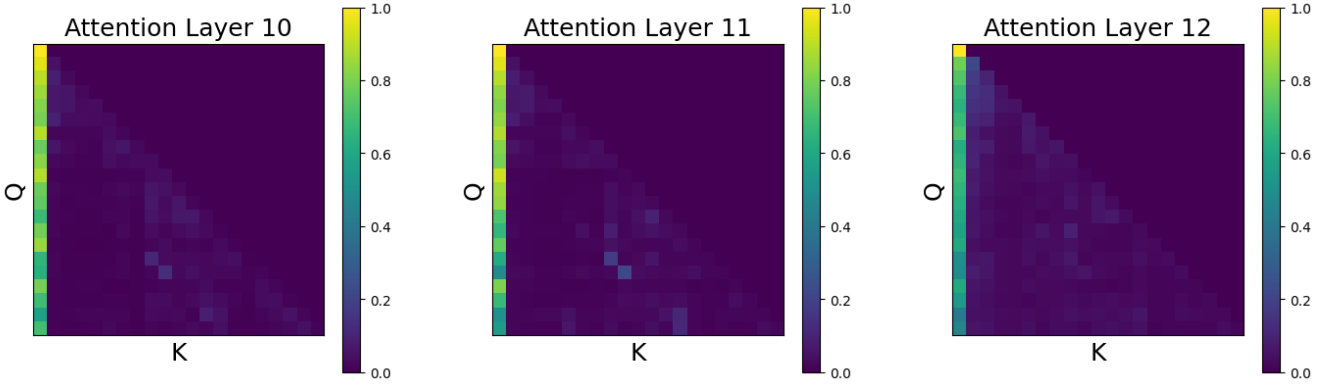
(f) layer 6



(g) layer 7

(h) layer 8

(i) layer 9

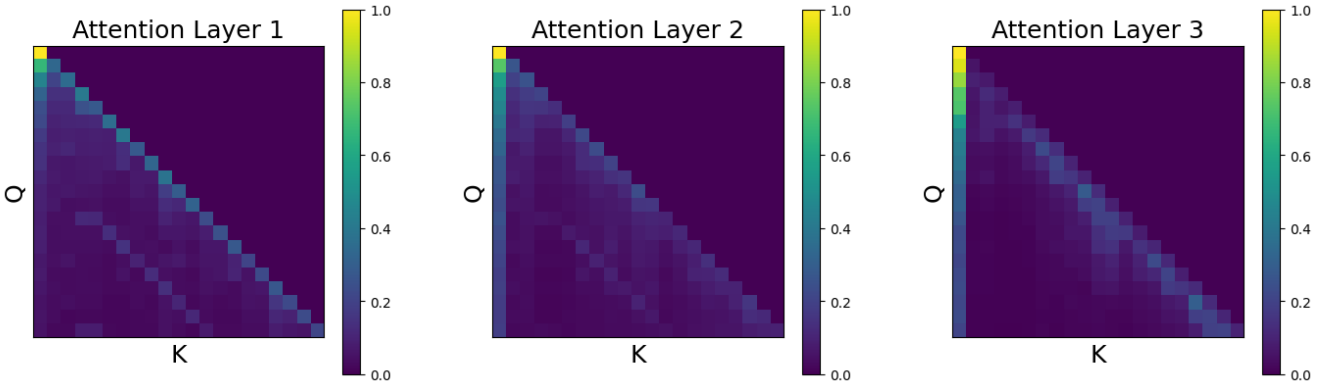


(j) layer 10

(k) layer 11

(l) layer 12

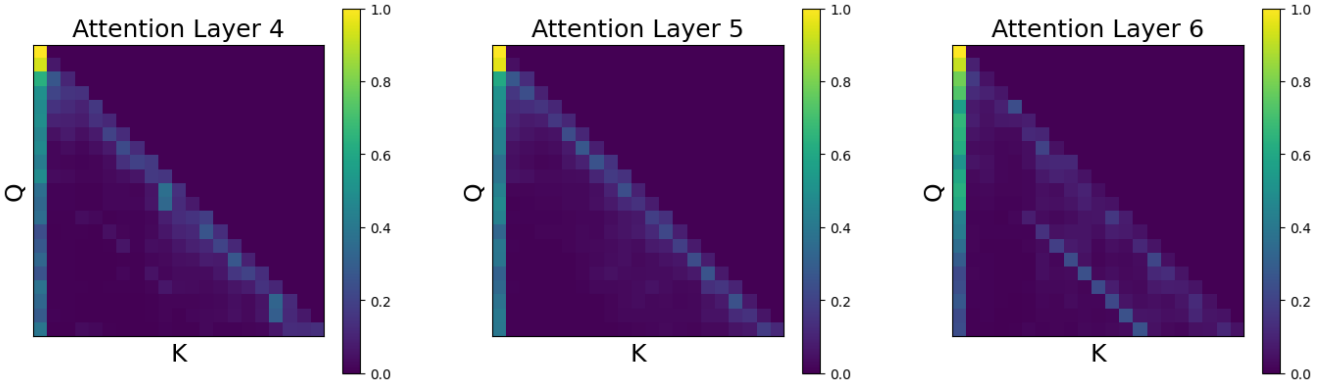
Figure 11: Intervention 4: Attention maps for all layers with zeroing out the BOS token. The attention sinks remains



(a) layer 1

(b) layer 2

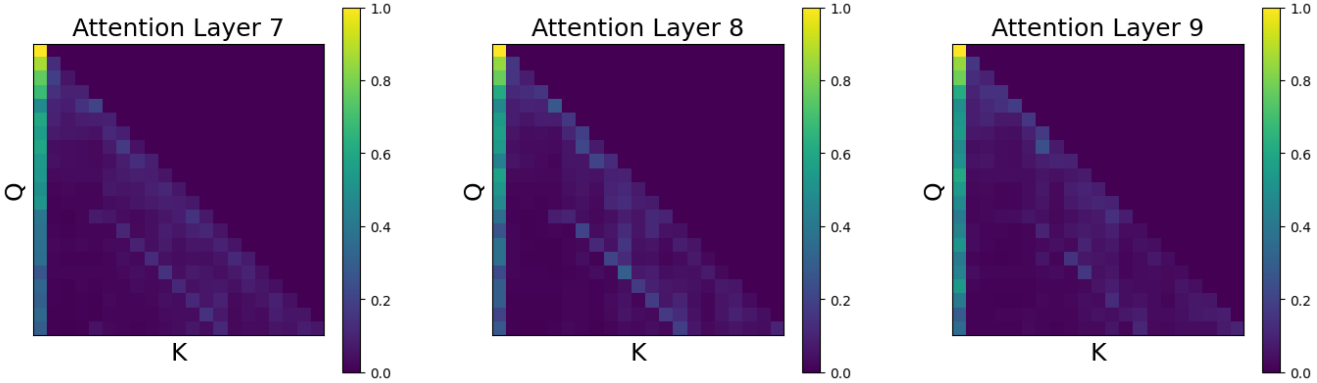
(c) layer 3



(d) layer 4

(e) layer 5

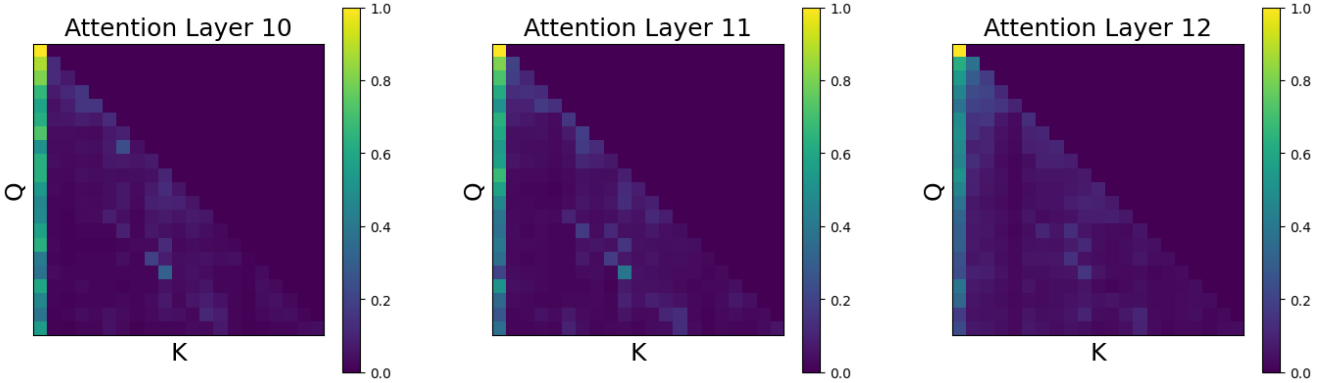
(f) layer 6



(g) layer 7

(h) layer 8

(i) layer 9

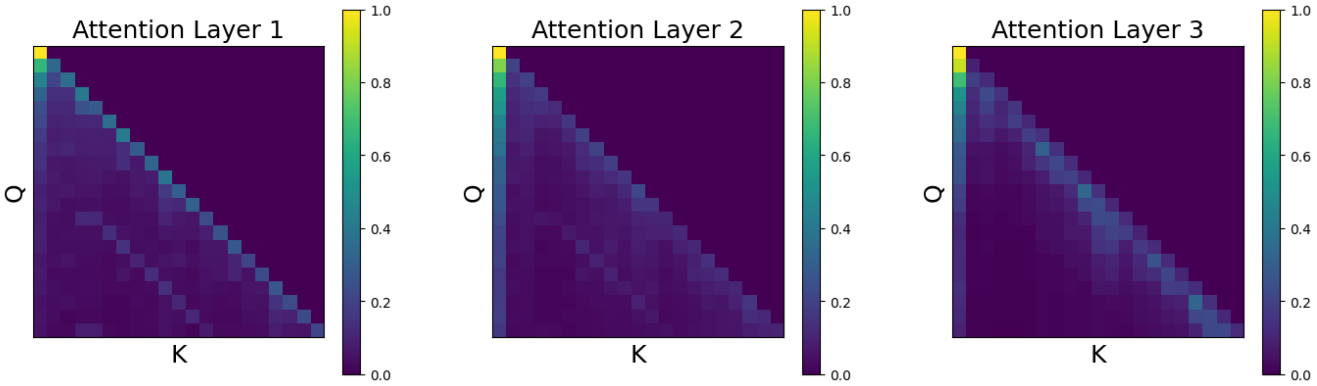


(j) layer 10

(k) layer 11

(l) layer 12

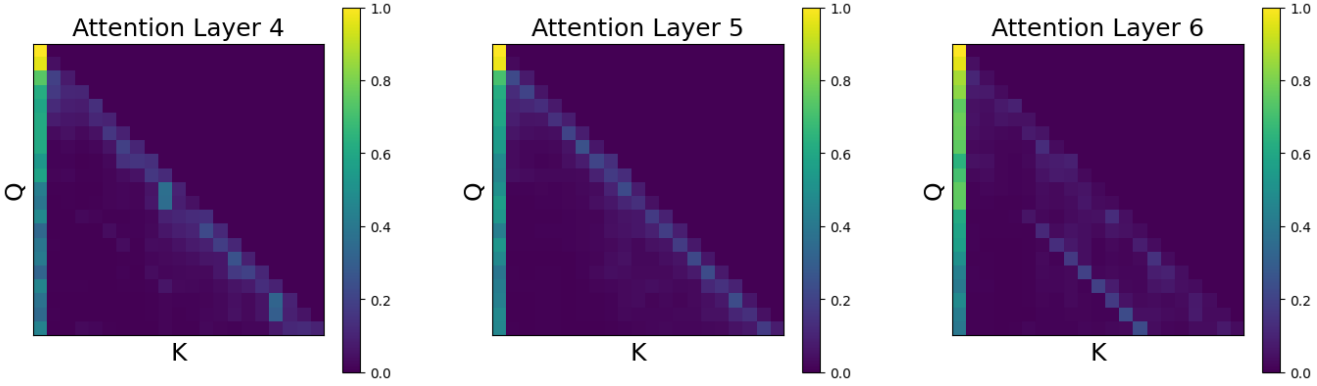
Figure 12: Intervention 5: Attention maps for all layers zeroing out W_k at bias-projection coordinates. The sink is reduced.



(a) layer 1

(b) layer 2

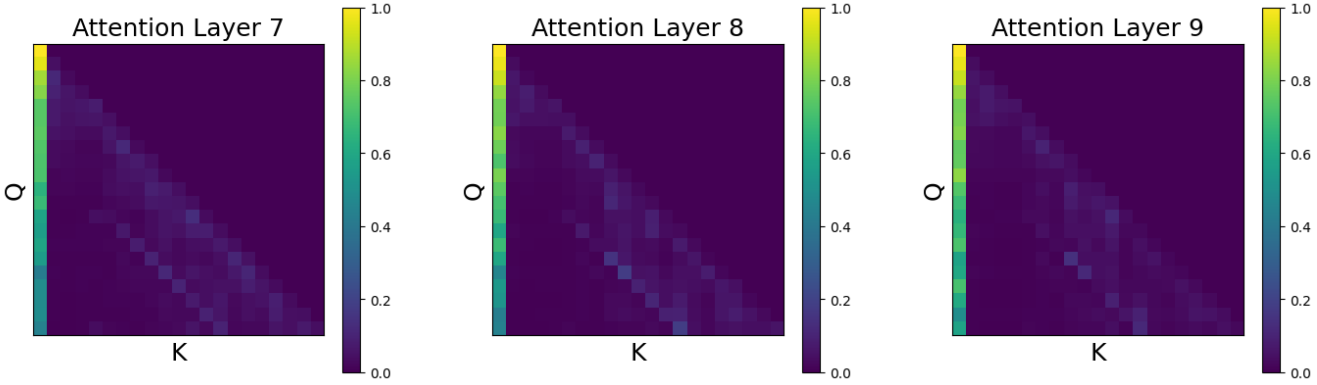
(c) layer 3



(d) layer 4

(e) layer 5

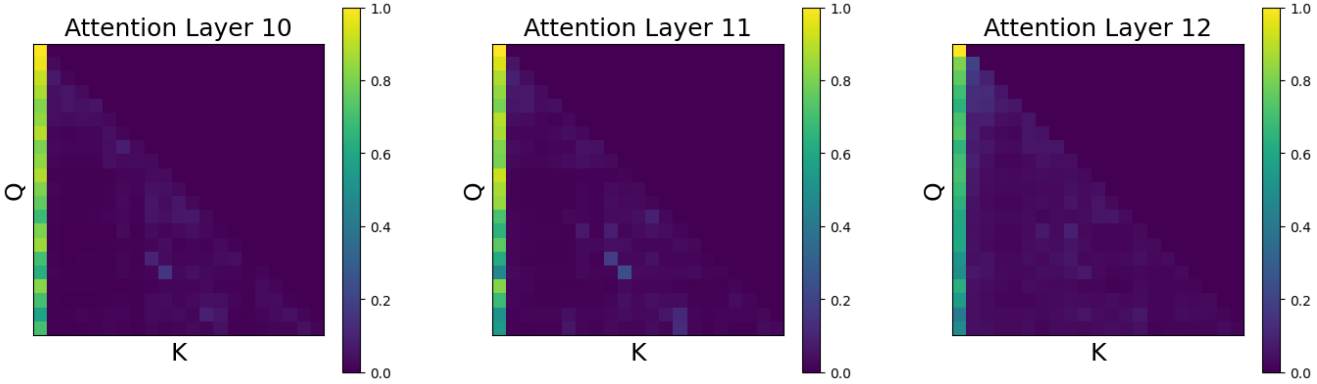
(f) layer 6



(g) layer 7

(h) layer 8

(i) layer 9



(j) layer 10

(k) layer 11

(l) layer 12

Figure 13: Intervention 5: Attention maps for all layers zeroing out W_k at random coordinates. The sink remains.