No Universal Mechanism for Attention Sink in Transformers: Evidence from GPT-2

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 different architectures.

1 Introduction

2 Related Work

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)). Some models, such as Mamba-based models as Endy et al. (2025) shows, don't exhibit attention sink.

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)) and LMMs (Kang et al., 2025). 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.

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).

For autoregressive generation, given query $q_t^{(i)}$ and keys $\{k_j^{(i)}\}_{j\le t}$, the attention weights are $\alpha_{tj} = \mathrm{softmax}_j\big((q_t^{(i)})^\top k_j^{(i)} / \sqrt{d}\big)$ 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 divides the feature dimension across h heads, computing attention independently within each head's subspace before concatenating outputs. For simplicity of analysis, our experiments treat W_k and b_Q in their original form prior to head-wise reshaping.

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 . 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* (EPE) 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 it captures the net positional signal that emerges after the first layer's transformation. Specifically, we observe experimentally that adding EPE_i to the output of the first layer (when no positional encoding was initially provided) produces approximately the same effect as the standard approach of adding the raw positional encoding p_i before the first layer. This equivalence demonstrates that EPE_i represents the effective contribution of positional information after being processed through the network's initial transformations (Experiments demonstrating this can be found in).

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

$$s_{t\to j}^{(i)} = (W_q^{(i)} x_t^{(i)})^\top (W_k^{(i)} x_j^{(i)}) + (W_q^{(i)} x_t^{(i)})^\top b_K^{(i)} + (b_Q^{(i)})^\top (W_k^{(i)} x_j^{(i)}) + (b_Q^{(i)})^\top b_K^{(i)}. \quad (151)$$

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*) 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.

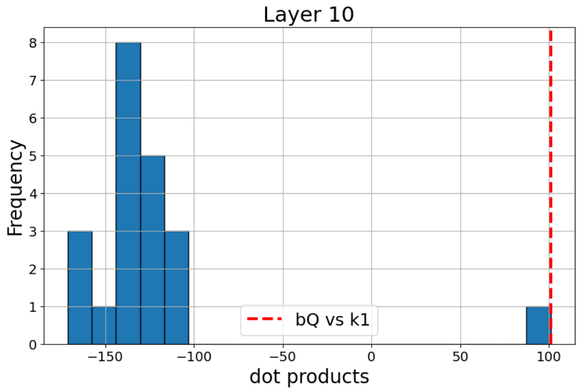


Figure 1: Distribution of bias terms $\Delta_j^{(10)}$ across positions. The first-position term $\Delta_1^{(10)}$ (red) centers at $\approx$ 100, while all other positions (blue) center at ≈ -140, demonstrating a learned preference for the first token.

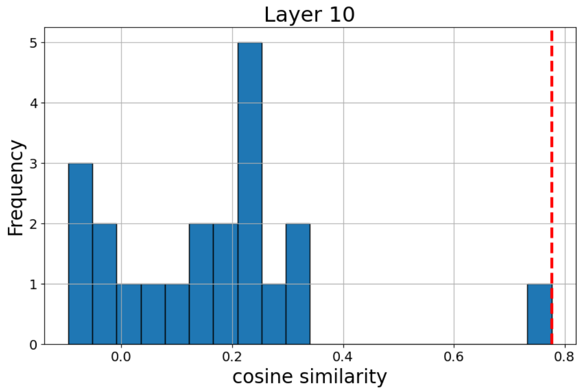


Figure 2: Cosine similarity between query bias $b_Q^{(10)}$ and $W_k^{(10)}$EPE. $W_k^{(10)}\mathrm{EPE}_1$ (red) shows strong positive alignment (≈ 0.7), while other positions (blue) cluster near -0.2.

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 $W_k^{(i)}\mathrm{EPE}_1$ 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 (see Appendix A.2 for results across all layers).

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 $W_k^{(i)}$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 $W_k^{(10)}\mathrm{EPE}_1$ exhibits strong positive alignment with $b_Q^{(10)}$, while other positions cluster near zero (comprehensive results across all layers in Appendix A.3).

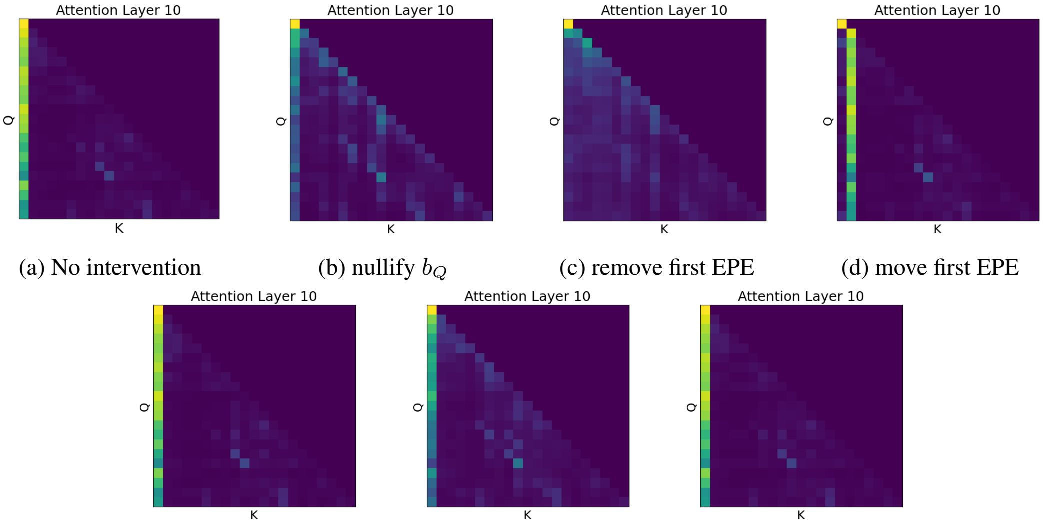
4.2.3 Coordinate-Level Structural Analysis

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 $|\mathrm{EPE}_1[d]|$ exactly where $|\gamma^{(i)}[d]|$ is large.

We identify coordinates with conspicuously large absolute values in EPE_1 (see Appendix A.4 for details). For each such coordinate d, we compare $|\gamma^{(i)}[d]|$ against the mean of the all of the columns. 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 (detailed results for all layers in Appendix A.5)

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. Full intervention



(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.

Layer	**Baseline (rand)**	d=138	d=447
layer 7	1.12±2.701	12.453	18.17
layer 9	1.23±3.225	17.846	26.014
layer 11	1.403±4.002	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 baseline mean $\pm$ two standard deviations across all coordinates. The massive-EPE_1 coordinates consistently exceed the baseline by wide margins, demonstrating that EPE_1 is irregularly large precisely where the bias projection has strong influence.

results across all layers are provided in Appendix A.6.

- **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 (complete layer-wise results in Appendix A.6.2).

- **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 (full layer analysis in Appendix A.6.3).

- **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. (extensive validation across layers in Appendix A.6.4).

- **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. (consistent findings across all layers in Appendix A.6.5).

- **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. (comprehensive layer-wise analysis in Appendices A.6.6 and A.6.7).

5 Conclusions

Attention sinks are a robust emergent behavior that appears across a wide range of Transformer architectures and modalities , but the mechanisms behind them differ across architectures. In the GPT-

2–style sub-architecture we studied, we identify a concrete implementation pathway: 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 position 1 (EPE_1), and (iii) structure in the key projection aligned with the large-magnitude coordinates of EPE_1. Crucially, this circuit cannot account for sinks in architectures that lack these components—for example, models without learned query biases or models using alternative positional schemes such as RoPE or ALiBi , all of which have been shown to exhibit attention sinks . This implies that while attention sinks are robust as a phenomenon, they are not governed by a single universal mechanism.

Implications The lack of a single universal mechanism reveals attention sinks as an optimization-friendly attractor: when multiple representational pathways exist, training reliably discovers circuits that implement the sink behavior. This has important implications for both understanding and controlling these phenomena. First, it suggests that attention sinks may serve a fundamental computational role that emerges regardless of specific architectural choices. Second, it indicates that effective mitigation strategies must be mechanism-aware rather than one-size-fits-all. Naive interventions targeting individual components (e.g., shrinking query biases) will likely fail, as optimization can compensate through alternative pathways. Instead, successful approaches must either address the underlying computational pressures that drive sink formation, or develop architecture-specific interventions tailored to each mechanism.

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.

Nir Endy, Idan Daniel Grosbard, Yuval Ran-Milo, Yonatan Slutzky, Itay Tshuva, and Raja Giryes. 2025. Mamba knockout for unraveling factual information flow. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23457–23477, Vienna, Austria. Association for Computational Linguistics.

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*.

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 Further Experiments

A.1 Effective positional encoding demonstration

In this section, we illustrate that EPE_i roughly captures the net positional signal that is added to the input after the first layer's transformation when adding the positional encoding p_i to the input $x_i^{(0)}$. To that end, we define an approximation of result of processing the input by first MLP:

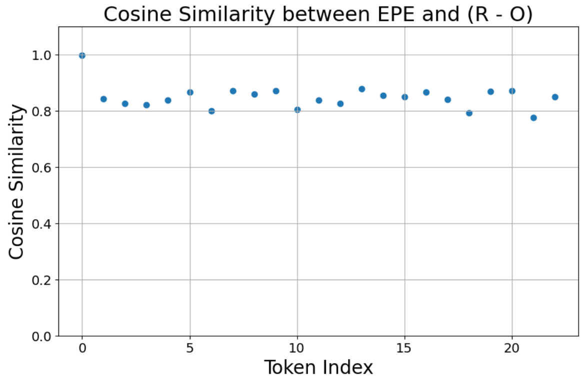


Figure 4: Coordinate values of EPE_i for the first token (replicating the distribution described in section 4.2.3). Most coordinates are near zero; a small set exhibits extremely large magnitudes ("massive activations").

$R_i := x_i^{(0)} + MLP^{(1)}(x_i^{(0)})$. We then define the input without positional information $e_i := x_i^{(0)} - p_i$, and the result of processing the input without the positional information: $O_i = e_i + MLP^{(1)}(e_i)$. We then compare EPE_i to the difference $R_i - O_i$, computing the cosine similarity for each token i. This directly measures whether the incremental contribution caused by adding p_i aligns in direction with EPE_i. The results range between 0.776 at the lowest and 0.996 at the highest. (Figure 4) These similarities are very high, indicating that EPE_i represents the effective contribution of positional information after being processed through the network's initial transformations.

A.2 Bias Term Magnitude Across All Layers

This section reproduces the bias-term magnitude analysis from section 4.2.1 across all layers: we plot the distribution of $\Delta_j^{(i)} = (b_Q^{(i)})^\top W_k^{(i)} x_j^{(i)}$ across positions for each layer (cf. fig. 1). In most layers, the first-position term $\Delta_1^{(i)}$ is a conspicuous outlier, indicating a strong prior to attend to position 1 (see Figure 5).

A.3 EPE-Bias Alignment Across All Layers

This section repeats the alignment analysis from section 4.2.2: for each layer i and position j we compute $\cos\left(b_Q^{(i)}, W_k^{(i)}\mathrm{EPE}_j\right)$, highlighting position 1. In most layers, position 1 shows strong positive alignment while other positions do not (see Figure 6).

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	18.199	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±3.618	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 baseline mean ± two standard deviations across all coordinates. Massive-EPE_1 coordinates consistently exceed the baseline, indicating that EPE_1 is irregularly large precisely where the bias projection is large.

A.4 Identifying Massive Activations in First-Position EPE

This section explains how we identify coordinates with unusually large absolute values in EPE_1. We select these coordinates by visual inspection of the EPE_1 coordinate distribution, choosing dimensions whose magnitudes are conspicuously larger than the rest (see Figure 7). Each such selected dimension exhibits the coordinate-level phenomenon described in section 4.2.3 (i.e., large $|\gamma^{(i)}[d]|$ and a strong contribution to the source-agnostic shift).

A.5 Coordinate-Level Alignment Across All Layers

This section tabulates $\gamma^{(i)} = (W_k^{(i)})^\top b_Q^{(i)}$ at coordinates where $|\mathrm{EPE}_1|$ is conspicuously large, mirroring the coordinate-level analysis in section 4.2.3. Values are compared against the baseline mean ± two standard deviations across all coordinates (see Table 2).

A.6 Intervention Results Across All Layers

This section reproduces the intervention analyses from section 4.2.4 across all layers, including the baseline and five targeted interventions. Each subsection mirrors the corresponding main-text figure and shows the layer-wise attention maps.

A.6.1 Baseline: No Intervention

We show attention maps with no intervention (cf. fig. 3a), demonstrating the prevalence of the first-position sink across layers (see Figure 8).

A.6.2 Intervention 1: Nullifying Query Bias

We zero b_Q (cf. fig. 3b), which substantially diminishes the sink across layers.

A.6.3 Intervention 2: Replacing First Position EPE

We swap EPE_1 with another position's EPE (cf. fig. 3c), which removes the first-position sink.

A.6.4 Intervention 3: Transplanting EPE to New Position

We transplant EPE_1 from position 1 to 2 (cf. fig. 3d), which induces a sink at position 2.

A.6.5 Intervention 4: Nullifying BOS Token

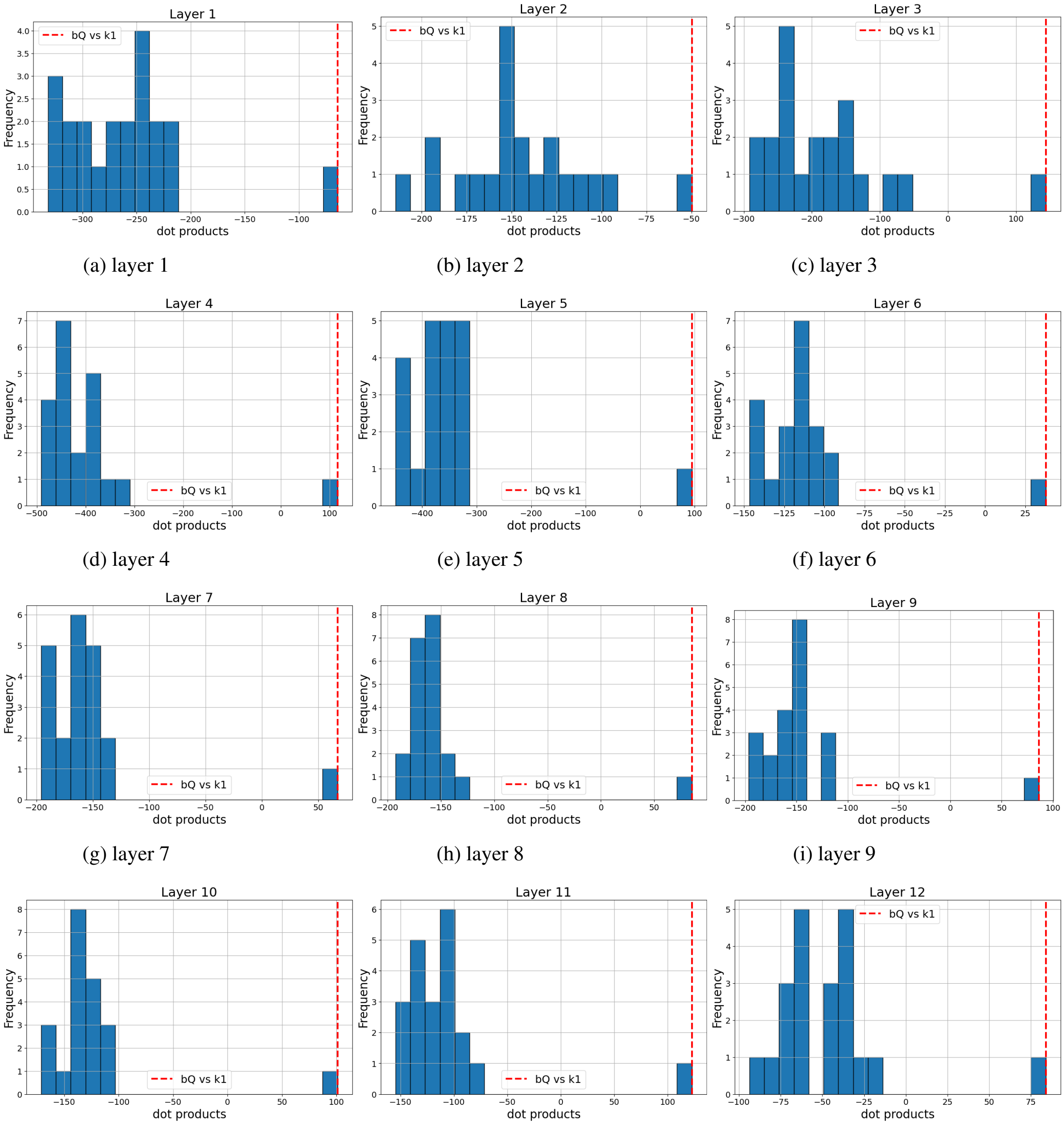
We zero the BOS token embedding prior to adding positional signals (cf. fig. 3e); the sink persists.

A.6.6 Intervention 5: Nullifying Massive Activation Coordinates

We zero W_k columns at massive-EPE_1 coordinates (cf. fig. 3f), which reduces the sink far more than zeroing random columns.

A.6.7 Intervention 5 Control: Nullifying Random Coordinates

As a control, we zero an equal number of random W_k columns (cf. fig. 3g); the sink largely remains.



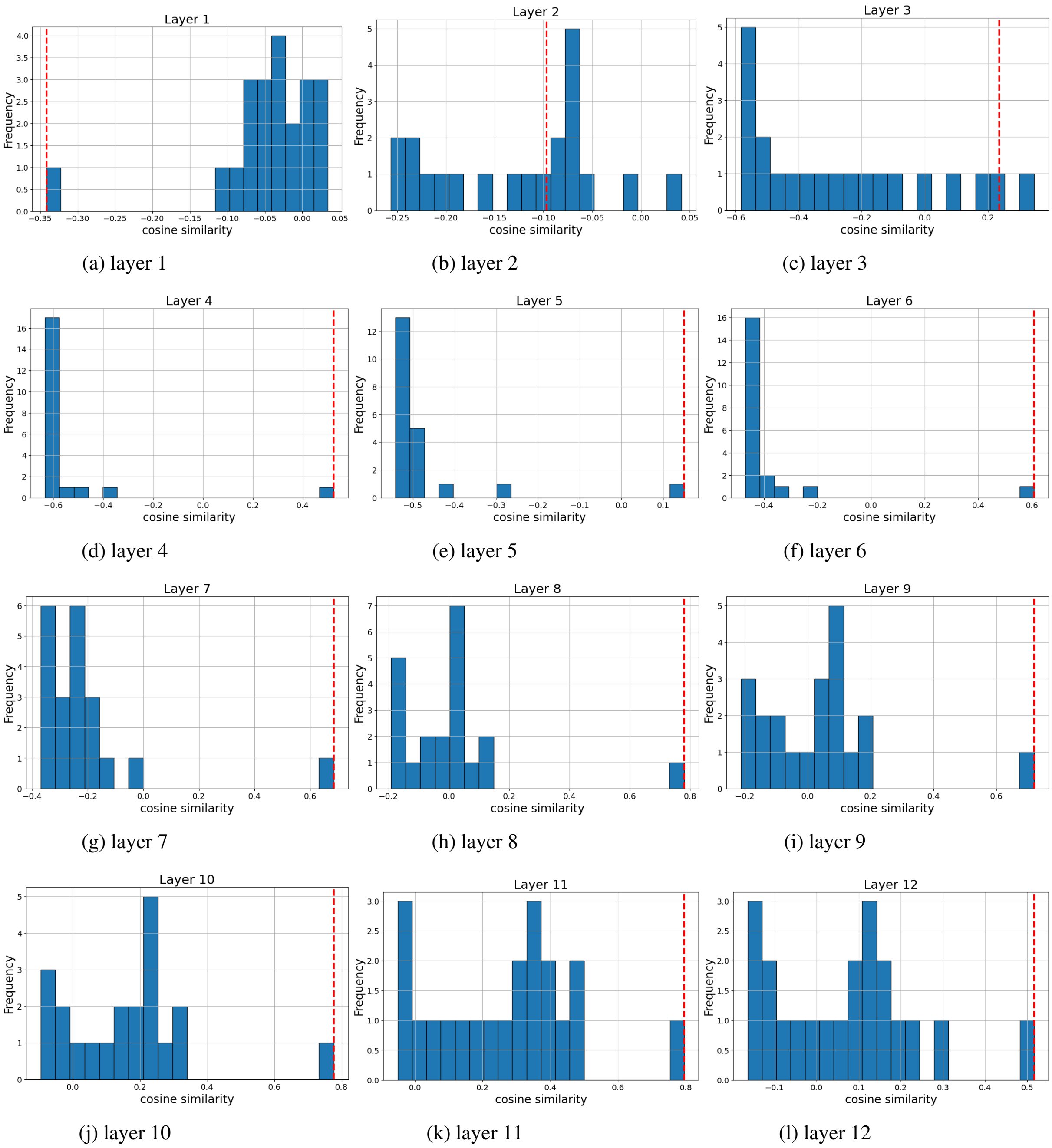
(a) layer 1 (b) layer 2 (c) layer 3

(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 5: Bias-term distributions Δ_j across positions j for each layer i (replicating fig. 1). Red denotes the first-position term $\Delta_1^{(i)}$; blue denotes all other positions. In most layers, the red distribution is shifted far to the right, evidencing an anomalously large first-position bias.



(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 6: Cosine similarity between the query bias and EPE-projected keys across layers and positions (replicating fig. 2). For each layer i and position j, we plot $\cos\left(b_Q^{(i)},\, W_k^{(i)}\mathrm{EPE}_j\right)$. Red marks position $j{=}1$; blue marks all other positions. Position 1 shows strong positive alignment while other positions do not, indicating that $W_k^{(i)}\mathrm{EPE}_1$ is specifically aligned with $b_Q^{(i)}$.

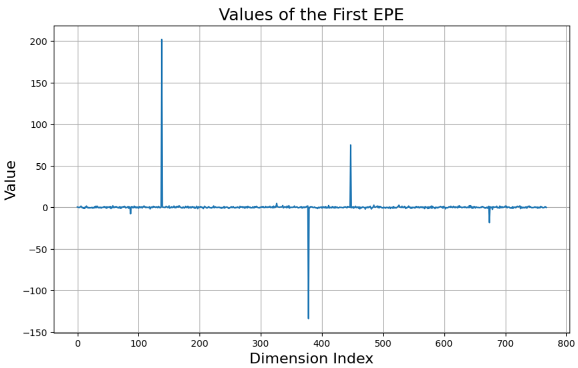
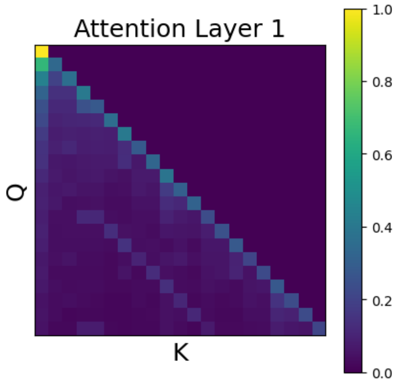
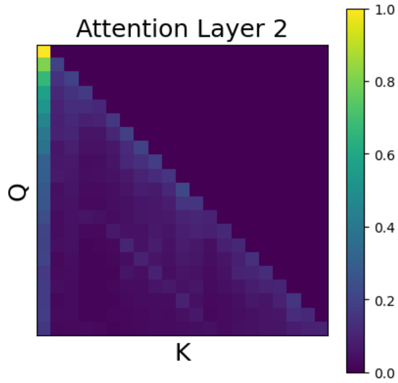


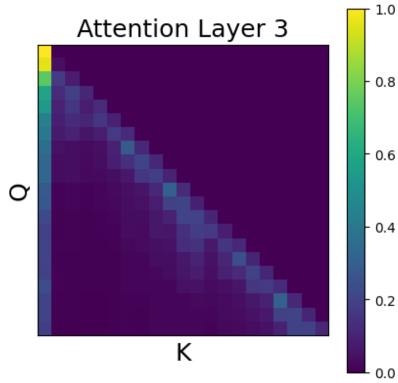
Figure 7: Coordinate values of EPE_1 for the first token (replicating the distribution described in section 4.2.3). Most coordinates are near zero; a small set exhibits extremely large magnitudes (“massive activations”).



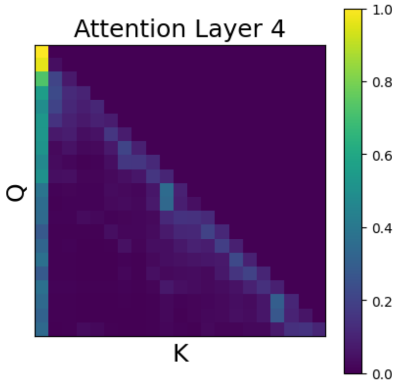
(a) layer 1



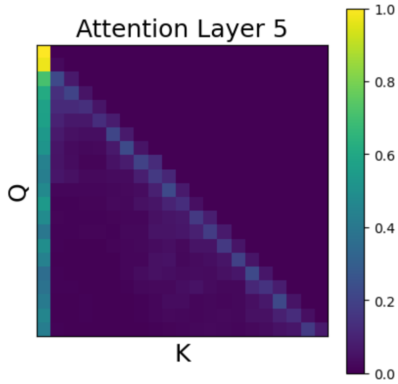
(b) layer 2



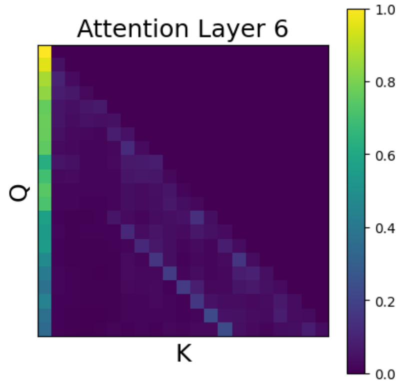
(c) layer 3



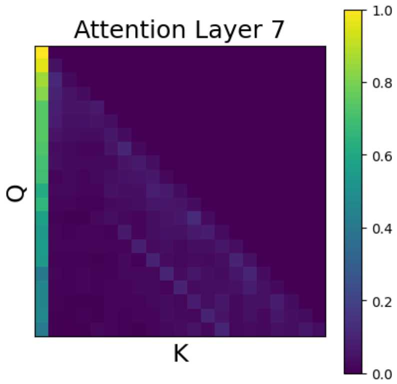
(d) layer 4



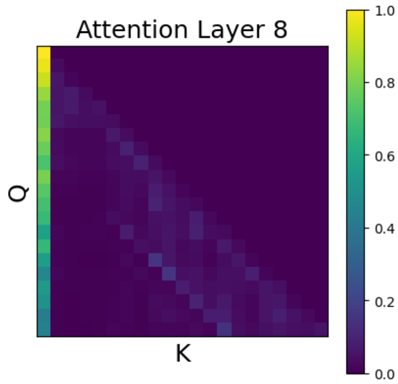
(e) layer 5



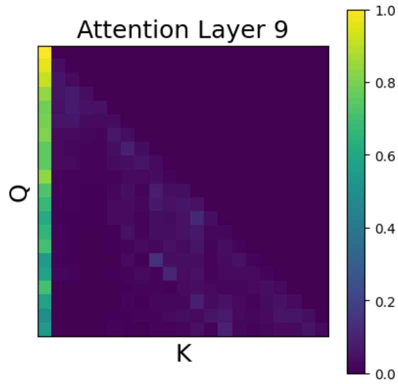
(f) layer 6



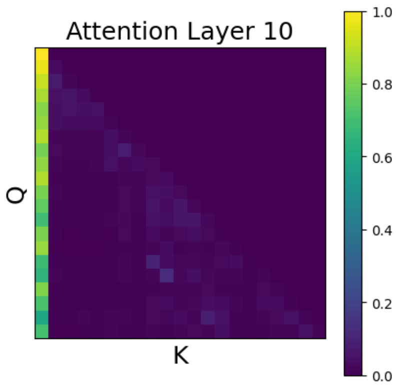
(g) layer 7



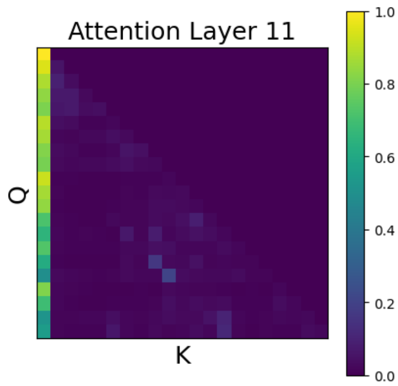
(h) layer 8



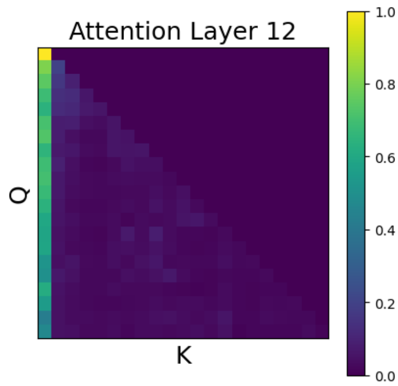
(i) layer 9



(j) layer 10

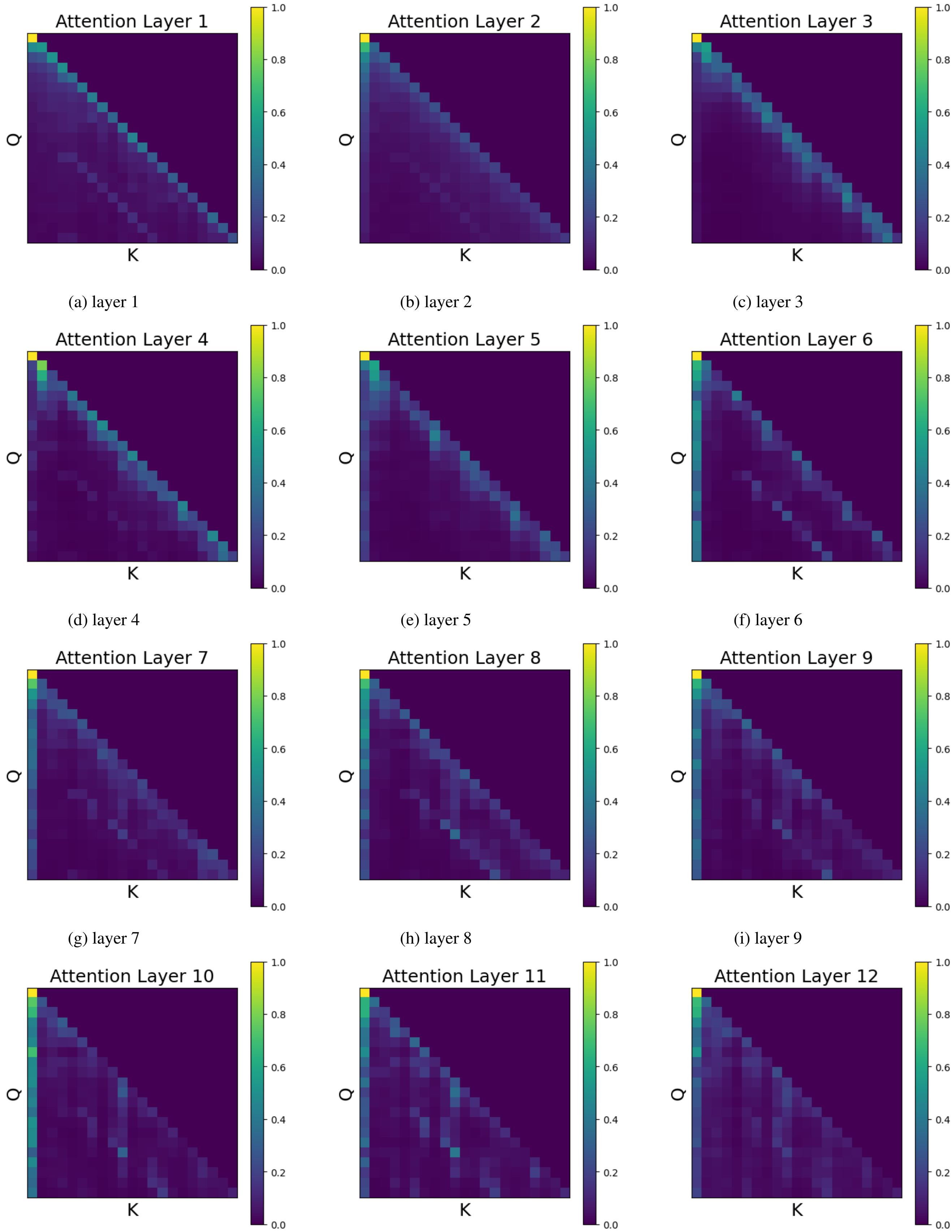


(k) layer 11



(l) layer 12

Figure 8: Attention maps for all layers with no intervention (replicating fig. 3a). A prominent first-position sink is visible in most layers.



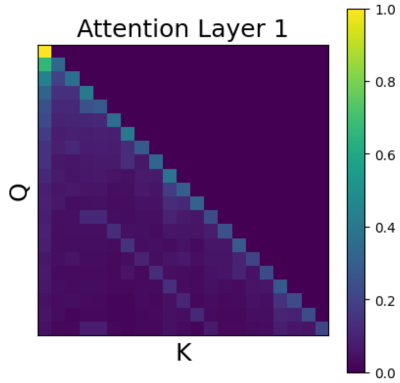
(a) layer 1 (b) layer 2 (c) layer 3

(d) layer 4 (e) layer 5 (f) layer 6

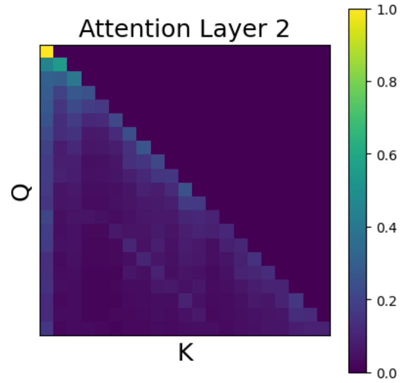
(g) layer 7 (h) layer 8 (i) layer 9

(j) layer 10 (k) layer 11 (l) layer 12

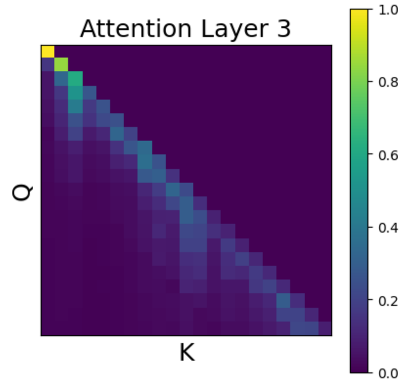
Figure 9: Attention maps for all layers with b_Q set to zero (replicating fig. 3b). The sink is substantially reduced across layers.



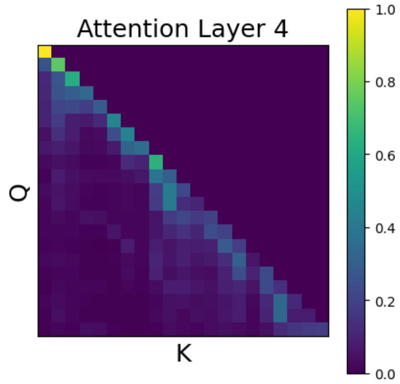
(a) layer 1



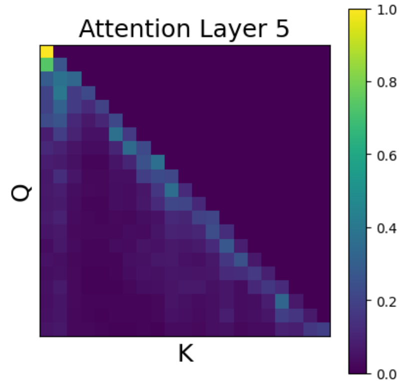
(b) layer 2



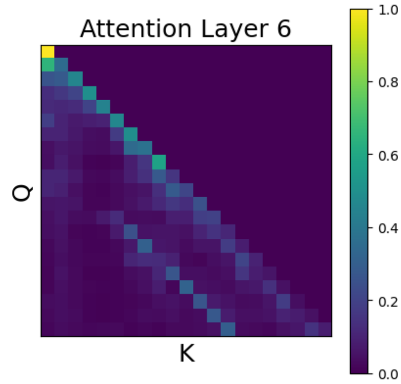
(c) layer 3



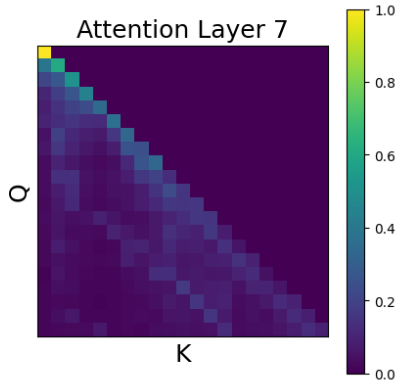
(d) layer 4



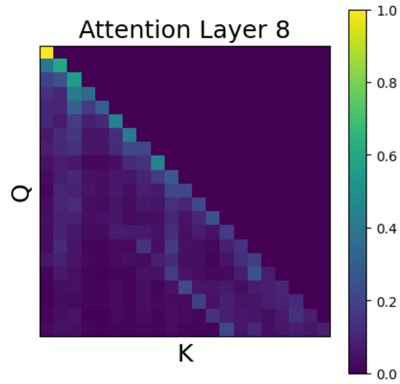
(e) layer 5



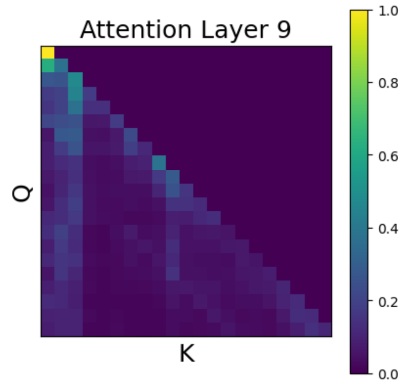
(f) layer 6



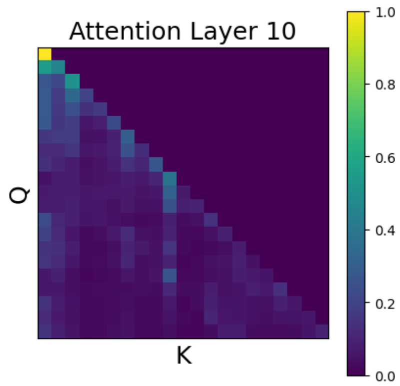
(g) layer 7



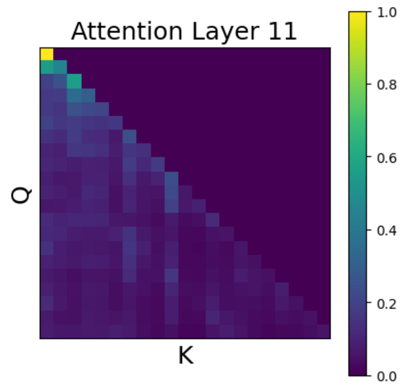
(h) layer 8



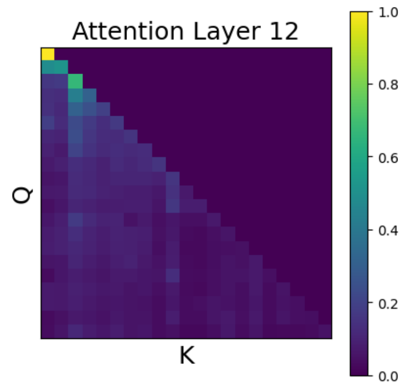
(i) layer 9



(j) layer 10

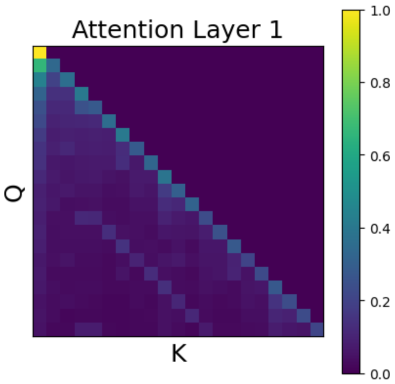


(k) layer 11

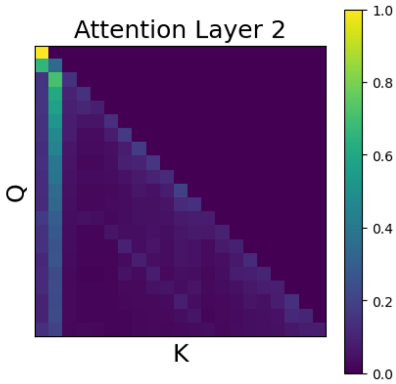


(l) layer 12

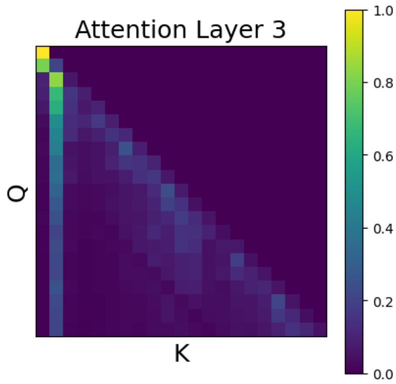
Figure 10: Attention maps for all layers after swapping EPE_1 with another position's EPE (replicating fig. 3c). The first-position sink disappears.



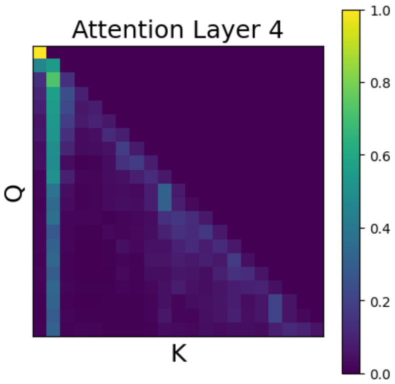
(a) layer 1



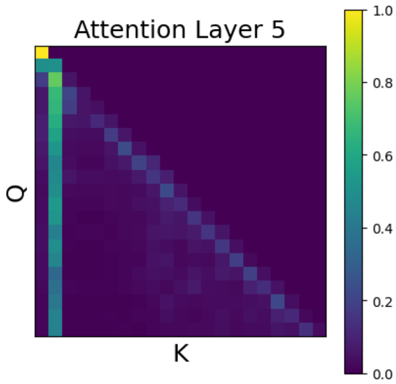
(b) layer 2



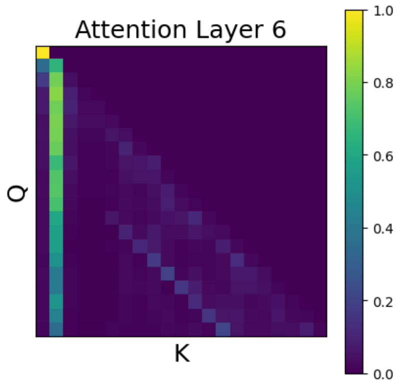
(c) layer 3



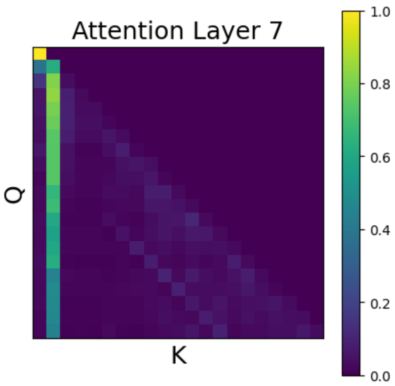
(d) layer 4



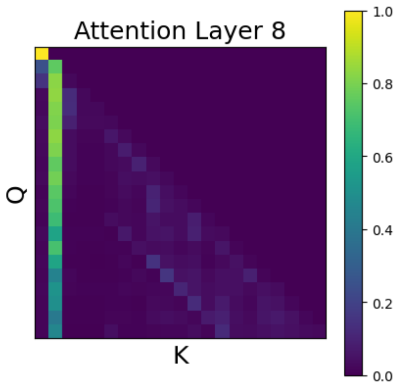
(e) layer 5



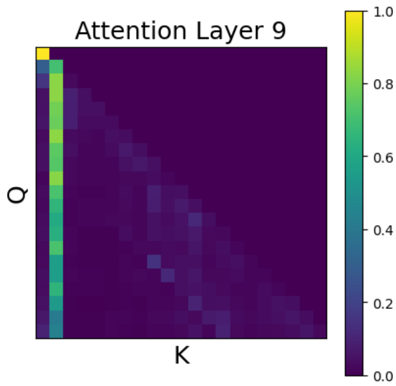
(f) layer 6



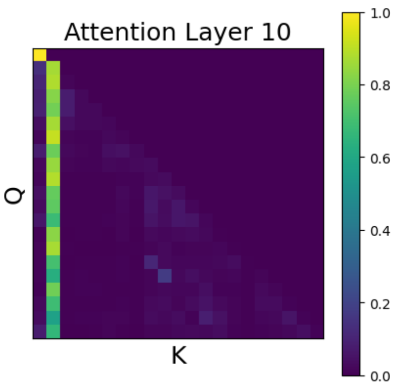
(g) layer 7



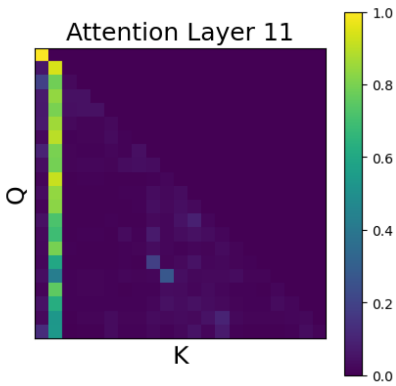
(h) layer 8



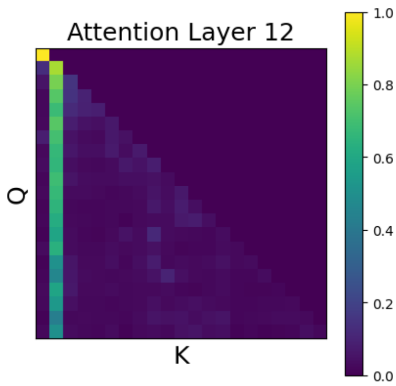
(i) layer 9



(j) layer 10

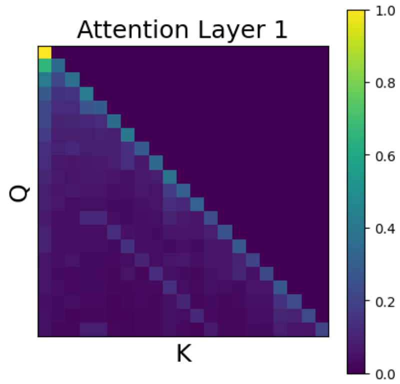


(k) layer 11

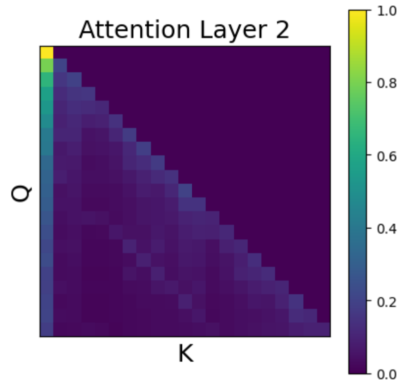


(l) layer 12

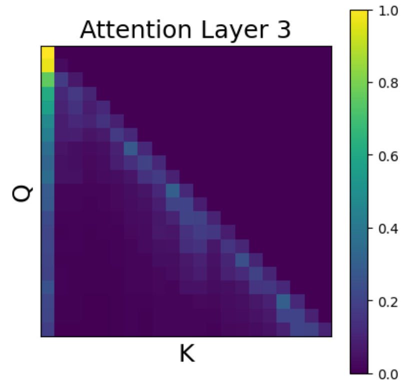
Figure 11: Attention maps for all layers after moving EPE_1 from position 1 to 2 (replicating fig. 3d). A strong sink forms at position 2.



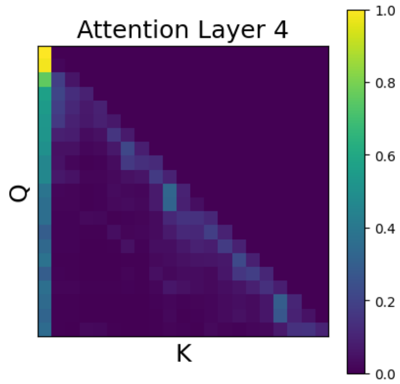
(a) layer 1



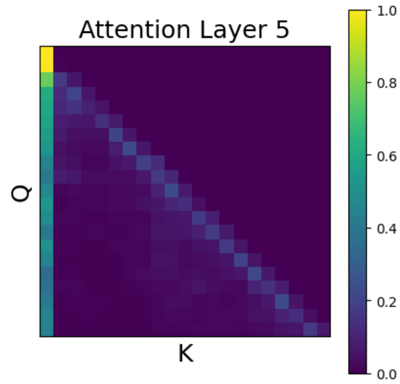
(b) layer 2



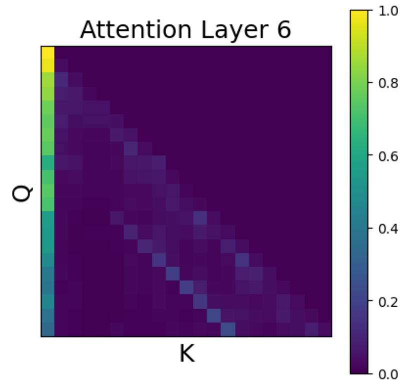
(c) layer 3



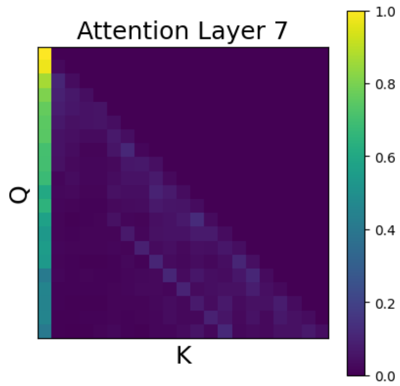
(d) layer 4



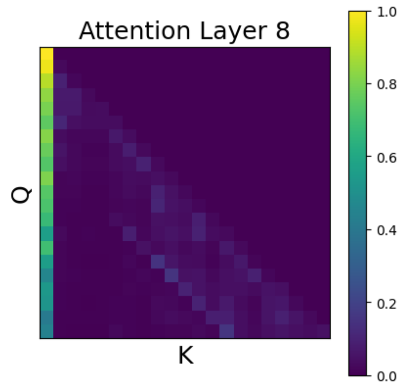
(e) layer 5



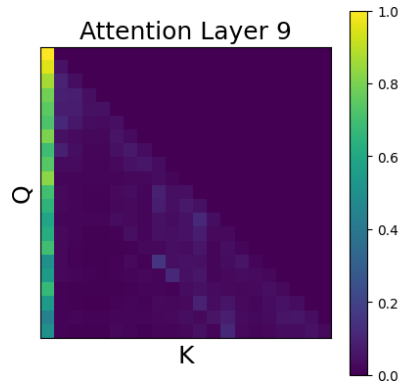
(f) layer 6



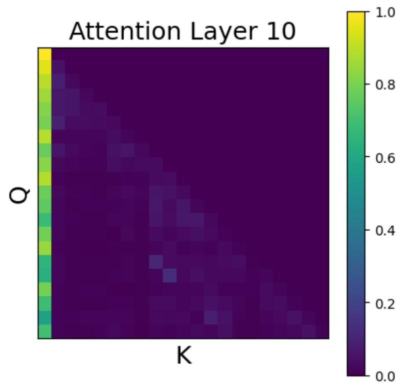
(g) layer 7



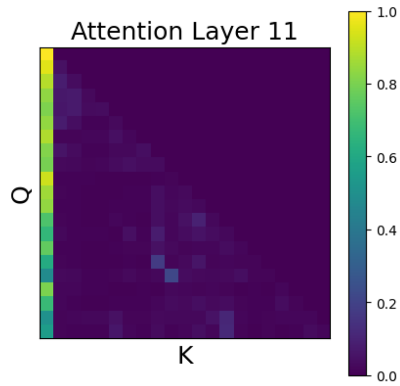
(h) layer 8



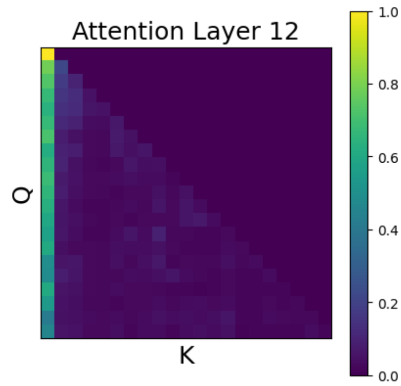
(i) layer 9



(j) layer 10

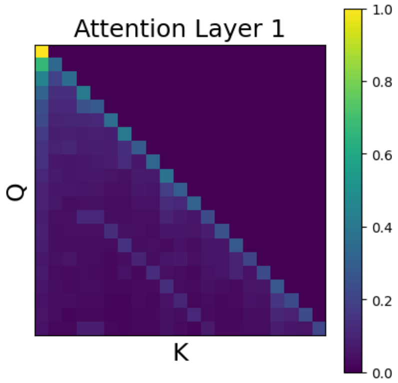


(k) layer 11

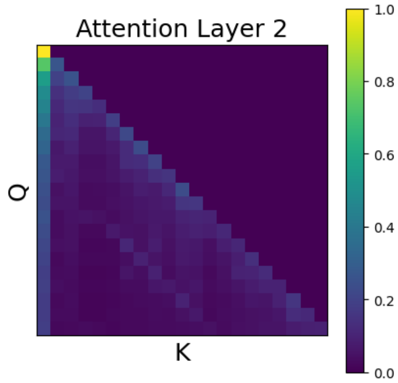


(l) layer 12

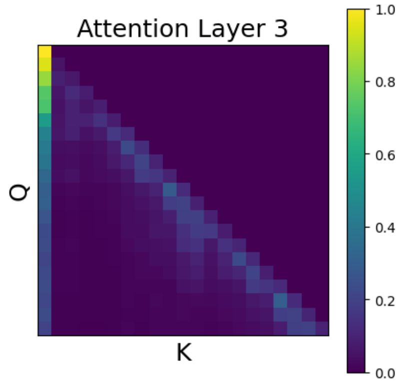
Figure 12: Attention maps for all layers after zeroing the BOS token embedding (replicating fig. 3e). The sink remains.



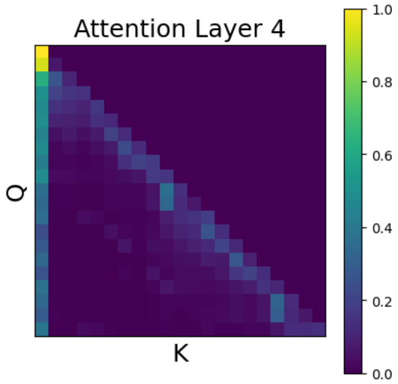
(a) layer 1



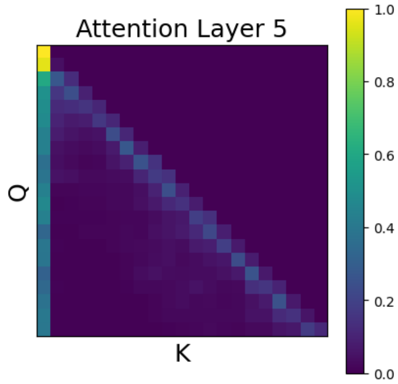
(b) layer 2



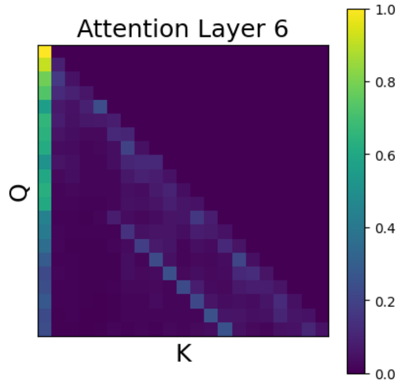
(c) layer 3



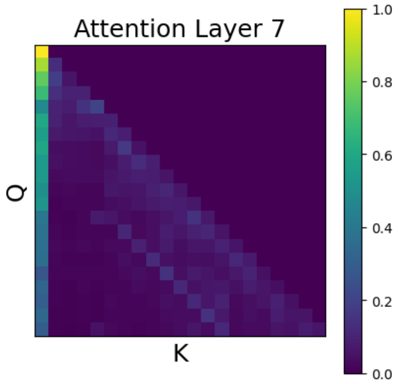
(d) layer 4



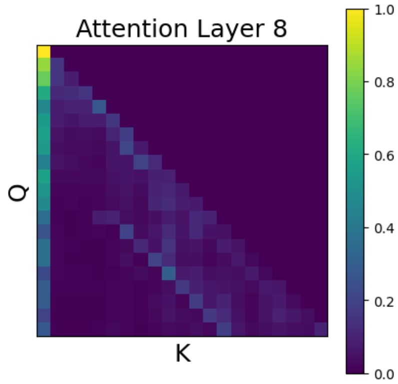
(e) layer 5



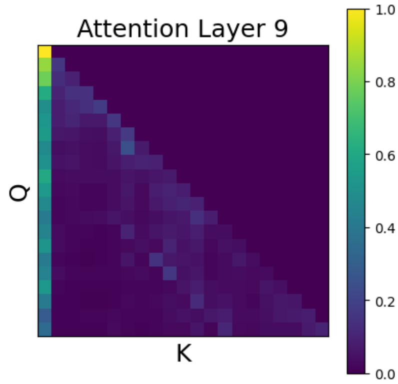
(f) layer 6



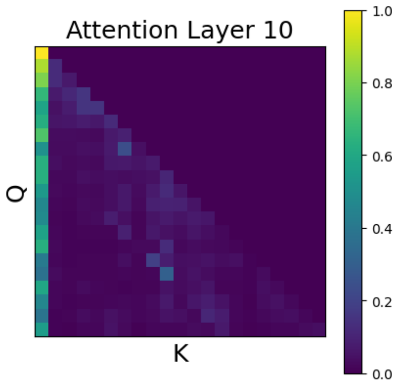
(g) layer 7



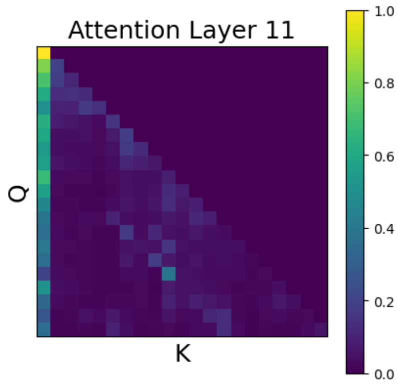
(h) layer 8



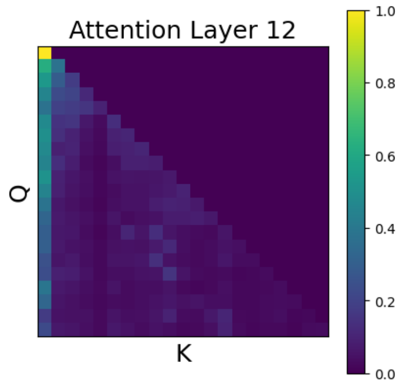
(i) layer 9



(j) layer 10

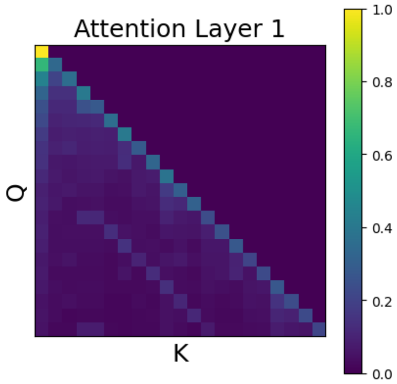


(k) layer 11

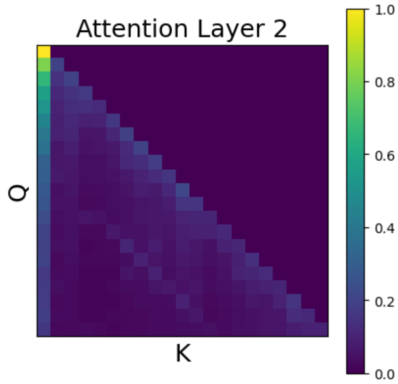


(l) layer 12

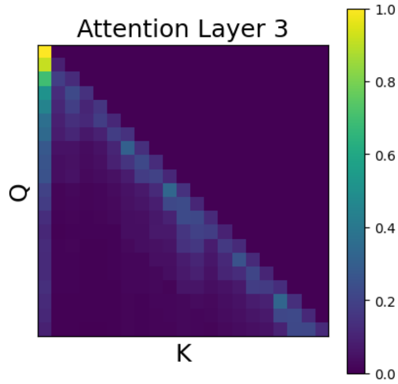
Figure 13: Attention maps for all layers after zeroing W_k at massive-EPE$_1$ coordinates (replicating fig. 3f). The sink is markedly reduced compared to random-coordinate zeroing.



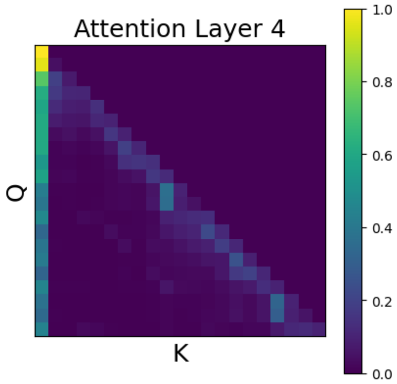
(a) layer 1



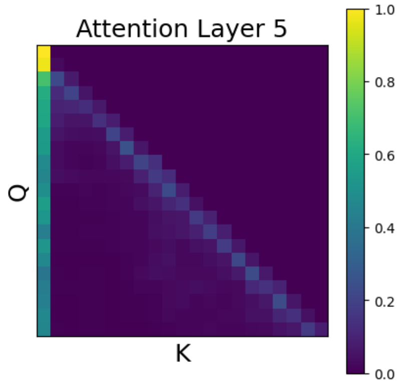
(b) layer 2



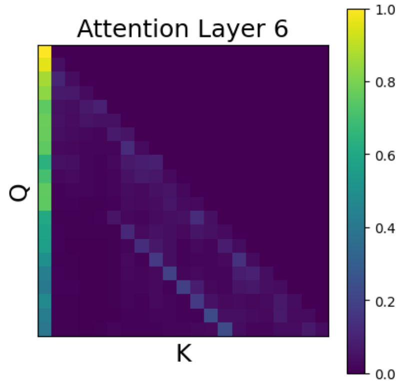
(c) layer 3



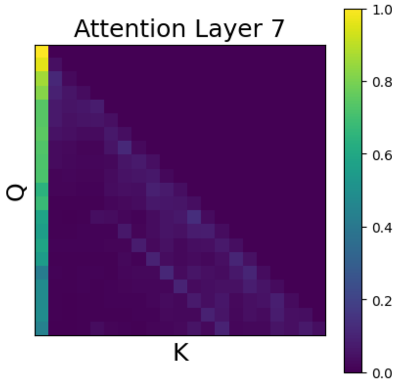
(d) layer 4



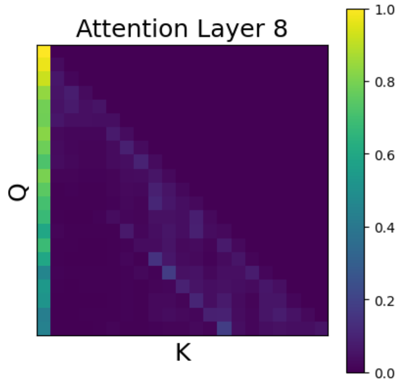
(e) layer 5



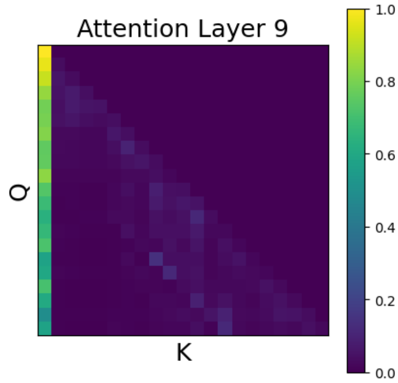
(f) layer 6



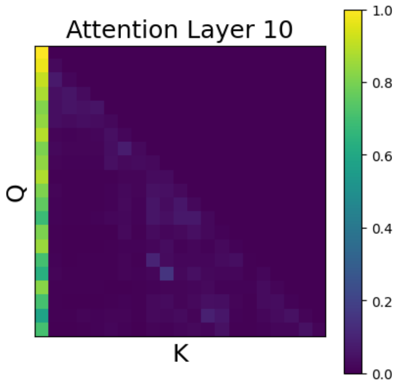
(g) layer 7



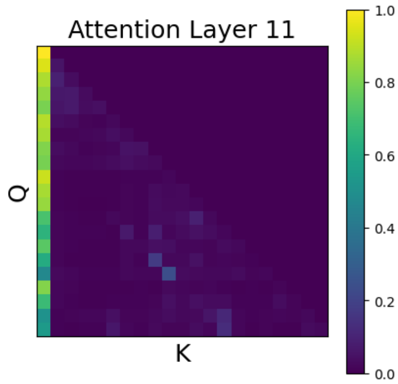
(h) layer 8



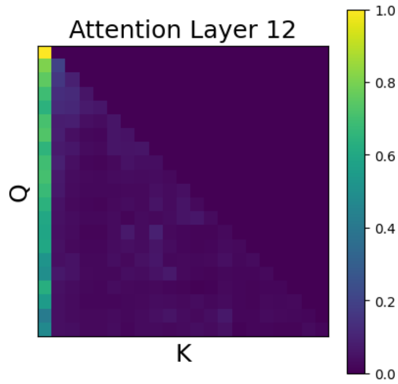
(i) layer 9



(j) layer 10



(k) layer 11



(l) layer 12

Figure 14: Attention maps for all layers after zeroing random W_k coordinates (replicating fig. 3g). The sink remains.