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

Transformers routinely display an *attention sink*: a persistent tendency to allocate disproportionate attention mass to early (often first) positions independent of semantic content (Xiao et al., 2023; Gu et al., 2025). The effect is robust across many language and vision architectures. It has been observed across training stages and hyperparameters (Gu et al., 2025; Guo et al., 2024), across model families and datasets (Xiao et al., 2023), and under diverse positional encodings—including absolute and learnable embeddings, ALiBi, RoPE, and even no explicit positional encodings (Press et al., 2021; Su et al., 2021; Irie et al., 2019; Gu et al., 2025). Similar sink-like patterns have also been reported in large multimodal systems and vision transformers (Kang et al., 2025; Wang et al., 2025; Feng and Sun, 2025). Together, these results indicate a robust, recurring phenomenon rather than a quirk of any single training recipe.[1]

The practical stakes are significant. Attention sinks can reduce effective context use, lower accuracy and calibration (Yu et al., 2024; Guo et al., 2024), aggravate numerical error and hinder quantization (Sun et al., 2024; Lin et al., 2024), and obscure interpretability by dominating attention maps (Guo et al., 2024). Understanding when sinks arise and how to control them is therefore directly relevant for model performance and interpretability.

We study the sink mechanistically in GPT-2–style Transformers with learned query biases and absolute positional embeddings (Radford and Narasimhan, 2018). Combining descriptive measurements with targeted causal interventions, we tie the first-token sink to three interacting components: learned query bias, first-layer transformation of positional information (captured by effective positional encoding, EPE), and structure in the key projection. To establish causality, we pair each measurement with targeted interventions showing the sink weakens, disappears, or moves accordingly; additionally, when alternative explanations are plausible, we ablate them and find the sink persists, isolating the causal pathway.

Finally, we situate these findings within the broader ecosystem. Many popular architectures do not include the components our GPT-2 analysis identifies as central in this setting: they omit learned attention biases and use alternative positional schemes—RoPE or ALiBi—instead of absolute embeddings, or even omit explicit positional encodings (NoPE) (Touvron et al., 2023; Chowdhery et al., 2022; Su et al., 2021; Press et al., 2021; Irie et al., 2019). Yet such models also robustly exhibit attention sinks (Gu et al., 2025; Xiao et al., 2023). Thus, the GPT-2 mechanism we uncover cannot account for their sinks. The behavior is robust across families, but the implementation pathway depends on architecture. This has two implications: first, attention sinks may play a fundamental computational role that arises irrespective of

[1] Some non-transformer architectures have been reported to have little to no sink (Endy et al., 2025).

particular architectural choices. Second, effective mitigation should be mechanism-aware rather than a one-size-fits-all approach.

2 Preliminaries

2.1 Attention mechanism

Let $X^{(i)} = [x_1^{(i)}, \dots, x_n^{(i)}]$ denote the input to attention layer i, where $x_t^{(i)} \in \mathbb{R}^d$ is the representation for position t (after LayerNorm). We denote projection matrices and 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: $q_t^{(i)} = x_t^{(i)}W_q^{(i)} + b_Q^{(i)}$, $k_t^{(i)} = x_t^{(i)}W_k^{(i)} + b_K^{(i)}$, and $v_t^{(i)} = x_t^{(i)}W_v^{(i)} + b_V^{(i)}$. Some architectures include biases (Zhang et al. (2022), Yang et al. (2024), Zeng et al. (2024)) , others omit them (Touvron et al. (2023), Dehghani et al. (2023), (Chowdhery et al., 2022)).

For autoregressive generation, attention weights are $\alpha_{tj} = \mathrm{softmax}_j\big(q_t^{(i)}(k_j^{(i)})^\top/\sqrt{d}\big)$ where the softmax is over valid positions $j \le t$. For simplicity, our experiments treat W_k and b_Q in their original form prior to head-wise reshaping.

2.2 Positional encoding

Attention layers are invariant to input permutations, lacking inherent awareness of token order. To address this, Transformers incorporate positional information through various schemes (Su et al. (2021), Press et al. (2021), Irie et al. (2019)). We focus on learned absolute positional encodings: a set of trainable vectors $\{p_i\}_{i=1}^L \subset \mathbb{R}^d$, where i is the token position and L is the sequence length. These are added to token embeddings e_i: $x_i^{(0)} = e_i + p_i$.

2.2.1 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 $\mathrm{MLP}^{(1)}$ is the first layer's feed-forward network applied to raw positional encoding p_i. We term this "effective" because it captures the net positional signal emerging after first-layer transformation. Experimentally, adding EPE_i to the first layer's output (when no positional encoding was initially provided) has roughly the same effect as adding p_i before the first layer, demonstrating that EPE_i represents the effective positional contribution (see section A.1 for details).

3 Methodology and Results

We first describe the mechanism underlying attention sinks in models with learnable query biases and absolute positional encodings. Then, we provide evidence through experimental analyses and causal interventions.

Throughout, we use the sentence "My name is Ozymandias, king of kings: Look on my works, ye Mighty, and despair!" as input, though our analysis generalizes to similar-length inputs.

3.1 The Mechanism behind Attention Sinks

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

$$s_{t\to j}^{(i)} = (x_t^{(i)}W_q^{(i)})(x_j^{(i)}W_k^{(i)})^\top + (x_t^{(i)}W_q^{(i)})b_K^{(i)\top} + (b_Q^{(i)})(x_j^{(i)}W_k^{(i)})^\top + (b_Q^{(i)})b_K^{(i)\top}. \tag{141}$$

The third term, $\Delta_j^{(i)} \triangleq b_Q^{(i)}W_k^{(i)\top}x_j^{(i)\top}$, 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 $b_Q^{(i)}W_k^{(i)\top}$. 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. The underlying reason for the large $\Delta_1^{(i)}$ is the effective positional encoding EPE_1. EPE_1 has very large absolute values on a small set of coordinates (a phenomenon called *massive activations* (Sun et al., 2024)) which are exactly those coordinates where $b_Q^{(i)}W_k^{(i)\top}$ 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.

3.2 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 3.1. In section 3.2.1 we show that $\Delta_1^{(i)}$ is conspicuously large relative to other positions across multiple layers. We then investigate its underlying cause and show in section 3.2.2 that $\mathrm{EPE}_1W_k^{(i)}$ exhibits strong alignment with vector $b_Q^{(i)}$ in deep layers. In section 3.2.3 we establish that EPE_1 exhibits massive

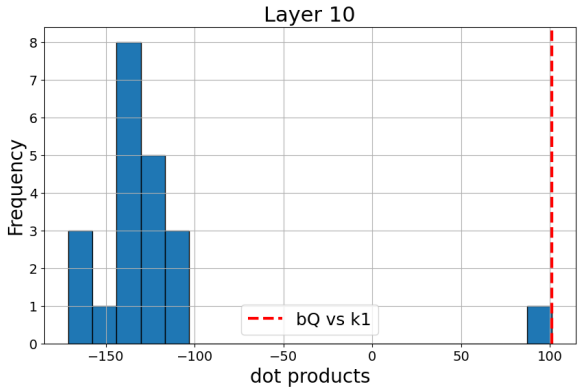


Figure 1: Distribution of bias terms $\Delta_j^{(10)}$ across positions. The first-position term $\Delta_1^{(10)}$ (red) centers at ≈ 100, while all other positions (blue) center at ≈ -140, demonstrating a learned preference for the first token.

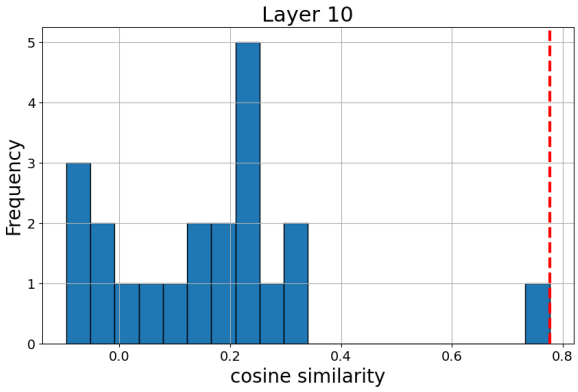


Figure 2: Cosine similarity between query bias $b_Q^{(10)}$ and $\text{EPE}W_k^{(10)}$. $\text{EPE}_1 W_k^{(10)}$ (red) shows strong positive alignment (≈ 0.7), while other positions (blue) cluster near -0.2.

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: $\gamma^{(i)} = b_Q^{(i)} W_k^{(i)\top}$ at coordinates where EPE_1 has massive activations (dims 138, 447) versus the baseline mean ± 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.

activations precisely at coordinates where the bias projection $b_Q^{(i)} W_k^{(i)\top}$ has high magnitude. Finally, in section 3.2.4 we use causal interventions to verify that disrupting any component abolishes the sink while transplanting components transfers it to new positions.

3.2.1 Bias Term Magnitude Analysis

First, We verify that $\Delta_j^{(i)} \triangleq b_Q^{(i)} W_k^{(i)\top} x_j^{(i)\top}$ is anomalously large for position 1. Histograms of $\Delta_j^{(i)}$ across positions j show $\Delta_1^{(i)}$ as a consistent distinct outlier. Figure 2 shows this for layer 10, where $\Delta_1^{(i)}$ substantially exceeds other positions (results across all layers are in Appendix A.2).

3.2.2 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 that end, we examine alignment between $\text{EPE}W_k^{(i)}$ and query bias $b_Q^{(i)}$. Figure 1 shows $\text{EPE}_1 W_k^{(10)}$ strongly aligns with $b_Q^{(10)}$, while other positions cluster near zero (full results for all layers are in Appendix A.3).

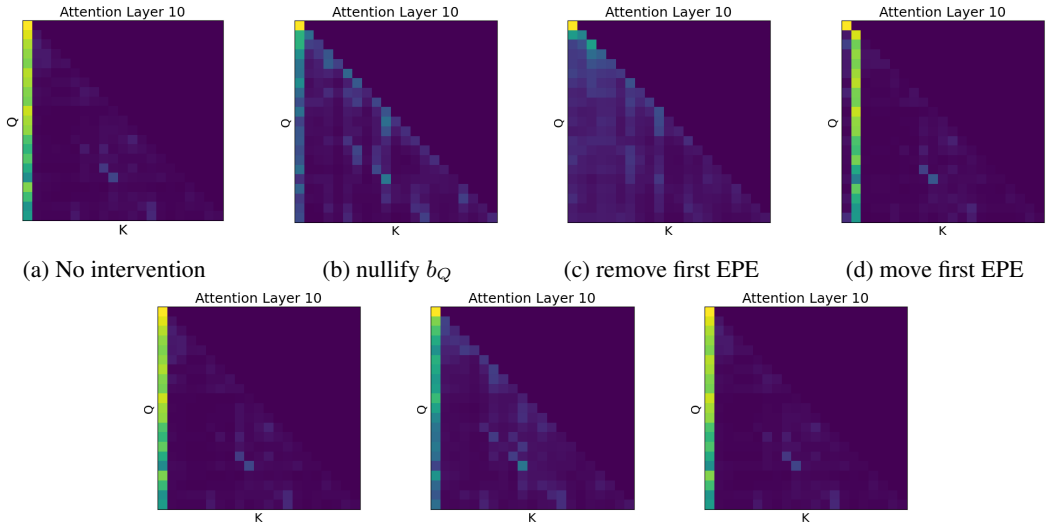
3.2.3 Coordinate-Level Structural Analysis

Massive coordinates of EPE_1 should coincide with coordinates favored by the bias projection. Let $\gamma^{(i)} = b_Q^{(i)} W_k^{(i)\top} \in \mathbb{R}^d$; its entry $\gamma^{(i)}[d]$ measures coordinate d's contribution to source-agnostic shift $\Delta_j^{(i)}$. 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 other rows. Table 1 shows massive coordinates (d=138, 447) substantially exceed baseline, confirming EPE_1 is large exactly where bias projection is large (full results for all layers are in Appendix A.5).

3.2.4 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 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.
- **Intervention 2 — Replace EPE_1 (specificity of the positional signal).** Swap EPE_1 with another position's EPE; the first-position sink disappears



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

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

(fig. 3c), indicating that EPE_1 is critical to induce a sink.

- **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.
- **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.
- **Intervention 5 — Zero W_k at bias-projection coordinates (structural pathway is necessary).** Zero W_k rows at massive-EPE_1 coordinates compared to zeroing W_k rows 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.

4 Conclusions

Attention sinks are robust across Transformer architectures and modalities, but mechanisms differ by architecture. In GPT-2–style models, we identify a concrete implementation pathway: interaction between (i) a learned query bias, (ii) the first-layer transformation of positional information, and (iii) structure in the key projection. Crucially, this circuit cannot account for sinks in architectures lacking these components—models without learned query biases or that use alternative positional schemes (RoPE, ALiBi, or no positional encodings)—yet these also 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.

5 Limitations

5.1 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.

5.2 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.

5.3 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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim M. Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, and 23 others. 2023. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*.

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*.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I. Jordan, and Song Mei. 2024. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms. *ArXiv*, abs/2410.13835.

Kazuki Irie, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. Language modeling with deep transformers. *ArXiv*, abs/1905.04226.

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.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

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*.

Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, Hui Xue, Jialing Tao, Ranjie Duan, and Jiexi Liu. 2025. Mirage in the eyes: Hallucination attack on multi-modal large language models with only attention sink. *Preprint*, arXiv:2501.15269.

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

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

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.

Team Glm Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, and 36 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *ArXiv*, abs/2406.12793.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.

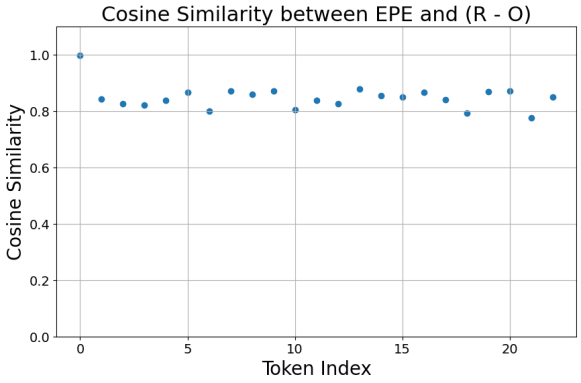


Figure 4: Coordinate values of EPE_i for the first token (replicating the distribution described in section 3.2.3). Most coordinates are near zero; a small set exhibits extremely large magnitudes ("massive activations").

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: $R_i := x_i^{(0)} + MLP^{(1)}(x_i^{(0)})$. 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 3.2.1 across all layers: we plot the distribution of $\Delta_j^{(i)} \triangleq b_Q^{(i)} W_k^{(i)\top} x_j^{(i)\top}$, across positions for each layer (cf. fig. 1). In most layers, the first-position term $\Delta_1^{(i)}$ is a conspicu-

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: $\gamma^{(i)} = b_Q^{(i)} W_k^{(i)\top}$ 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.

ous 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 3.2.2: for each layer i and position j we compute $\cos\left(b_Q^{(i)}, \mathrm{EPE}_j W_k^{(i)}\right)$, highlighting position 1. In most layers, position 1 shows strong positive alignment while other positions do not (see Figure 6).

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 3.2.3 (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 $\gamma^{(i)} = b_Q^{(i)} W_k^{(i)\top}$ at coordinates where $|\mathrm{EPE}_1|$ is conspicuously large, mirroring the coordinate-level analysis in section 3.2.3. 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 3.2.4 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.

B 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.

B.1 Positional Encoding in Transformers

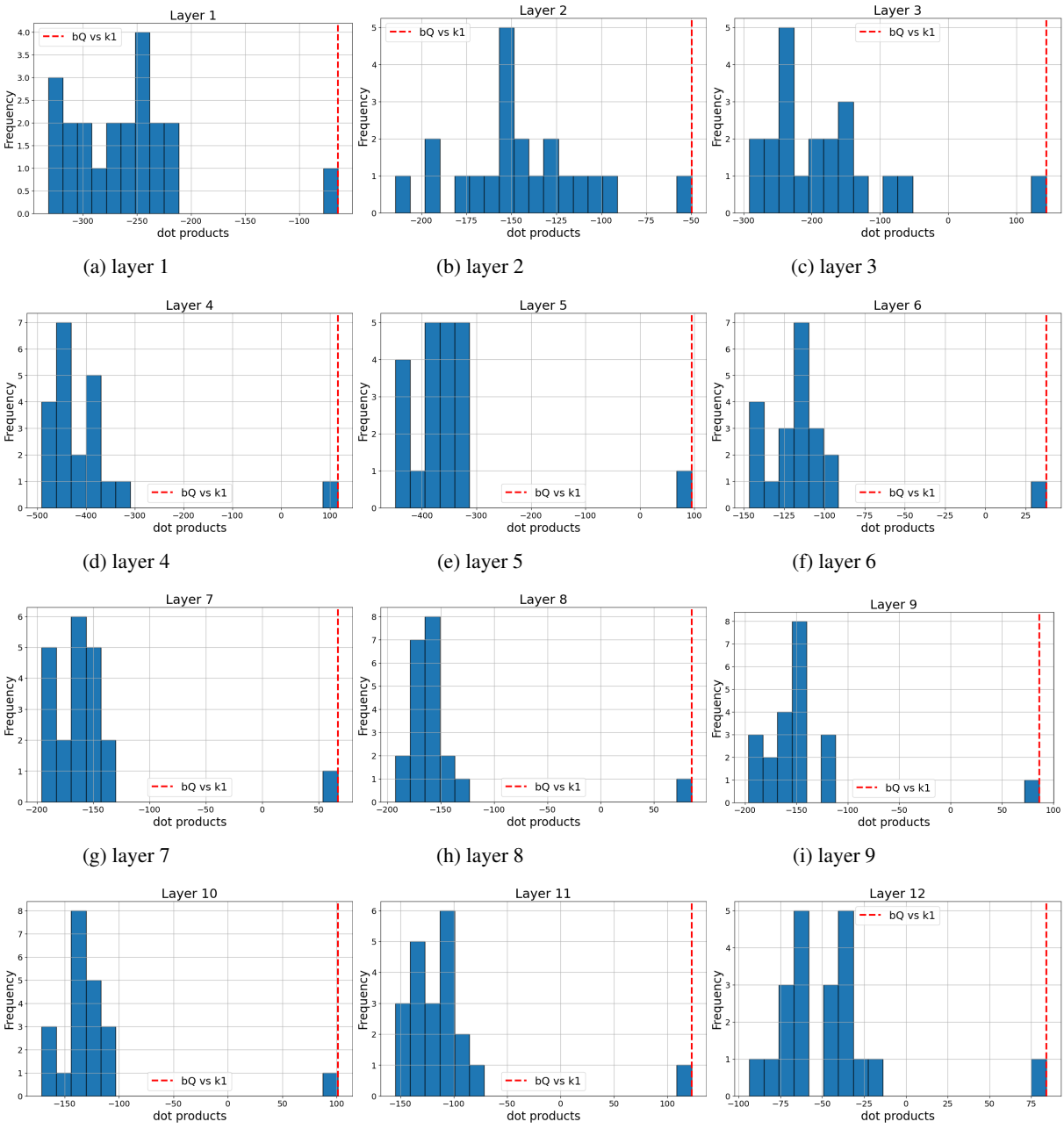
By design, the self-attention mechanism has no inherent sense of token order. To address this, Transformers must be augmented with positional information. The original Transformer used fixed sinusoidal embeddings (Vaswani et al., 2017). Many models in the GPT family (Radford and Narasimhan, 2018), including 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). Irie et al. (2019) found that deep Transformer language models do not require positional encoding (NoPE - no positional encoding)

B.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. They show the sink emerges under different kinds of positional encoding, such as absolute PE, learnable PE, NoPE, ALiBi and more. They showed that the attention sink occurs on various models without learnable attention bias. Xiao et al. (2023) also show that attention sinks emerge in decoder-only LMs. A similar phenomenon has been observed in LMMs (Large Multimodal Models) (Kang et al. (2025), (Wang et al., 2025)) and ViTs (Visual transformers) (Feng and Sun, 2025). [YRM: This is possibly the most important part of the RW. This should be with much more details and more citations, so saying specially that he demonstrated this in model without query bias and absolute positional encoding]. Some studies have found the attention sink could have a negative impact on the achievable accuracy, inference, quantization and interpretability of LLMs (Yu et al. (2024), Guo et al. (2024)) [YRM: Can you add a bit more details here? how does it hurt inference/quantization and interpretability?]. Feng and Sun (2025) shows that attention sinks distort ViT's ability to effectively process image patches. 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.

B.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.



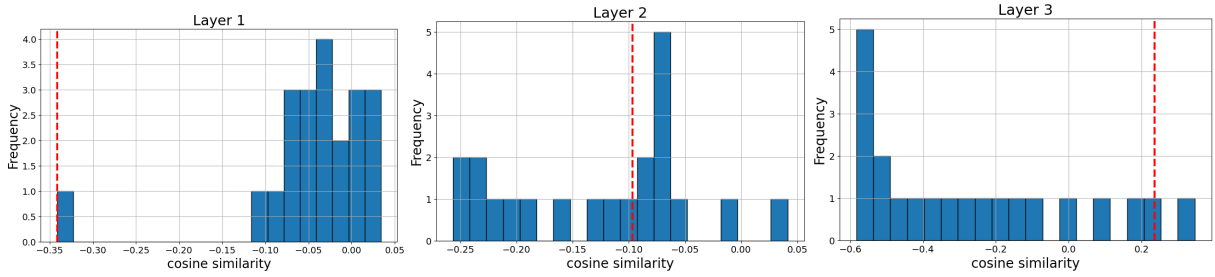
(a) layer 1 (b) layer 2 (c) layer 3

(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

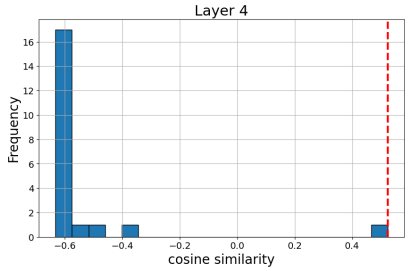
Figure 5: Bias-term distributions Δ_j across positions j for each layer i (replicating fig. 1). Red denotes the first-position term $\Delta_1^{(i)}$; blue denotes all other positions. In most layers, the red distribution is shifted far to the right, evidencing an anomalously large first-position bias.



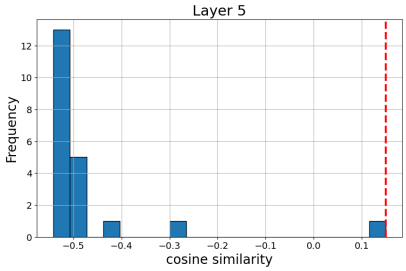
(a) layer 1

(b) layer 2

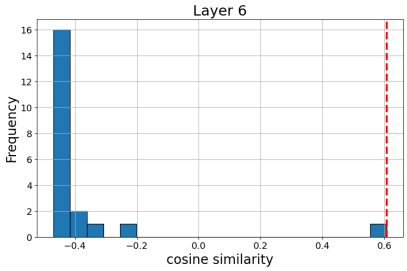
(c) layer 3



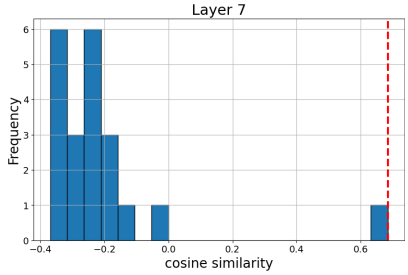
(d) layer 4



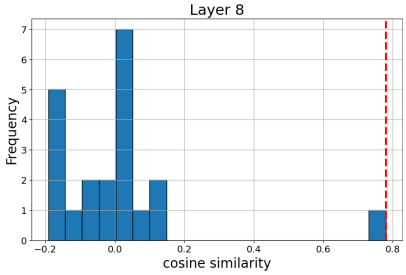
(e) layer 5



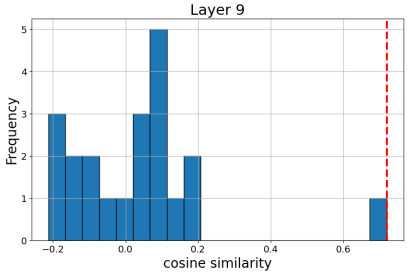
(f) layer 6



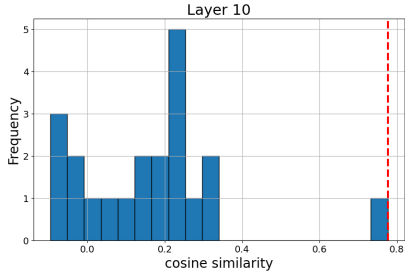
(g) layer 7



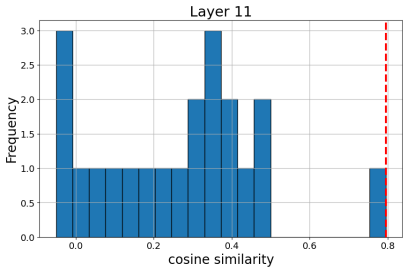
(h) layer 8



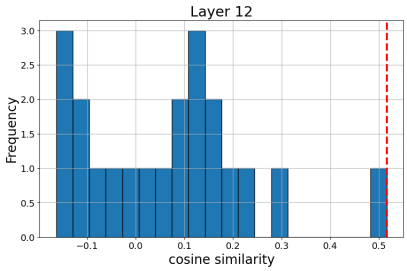
(i) layer 9



(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 $\cos\left(b_Q^{(i)}, \mathrm{EPE}_j W_k^{(i)}\right)$. Red marks position j=1; blue marks all other positions. Position 1 shows strong positive alignment while other positions do not, indicating that $\mathrm{EPE}_1 W_k^{(i)}$ is specifically aligned with $b_Q^{(i)}$.

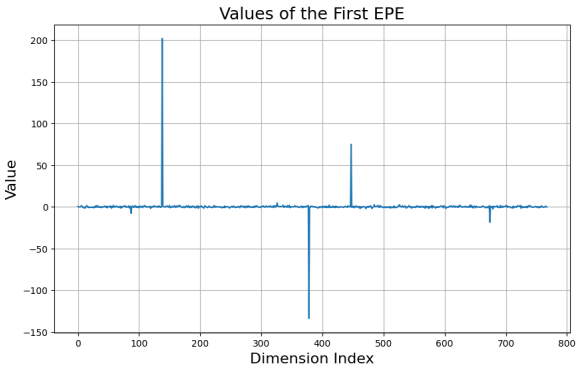
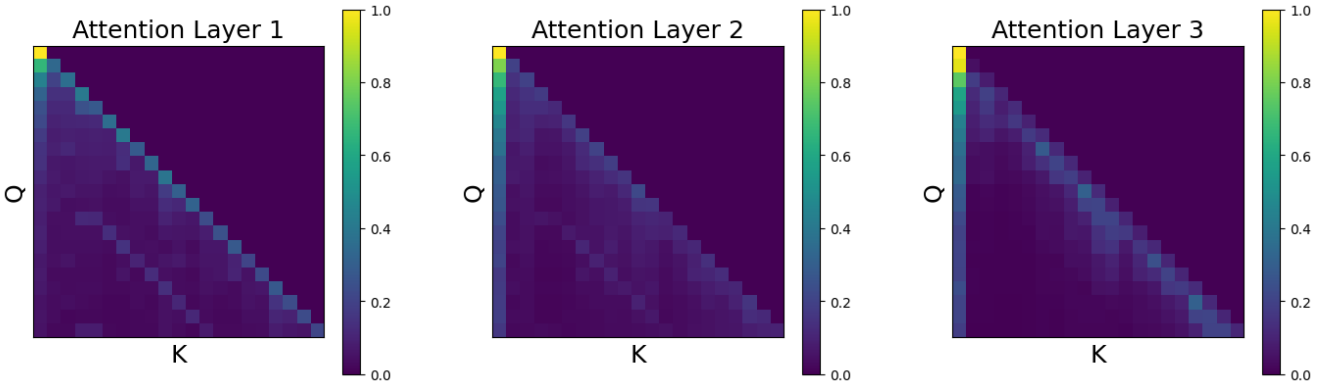


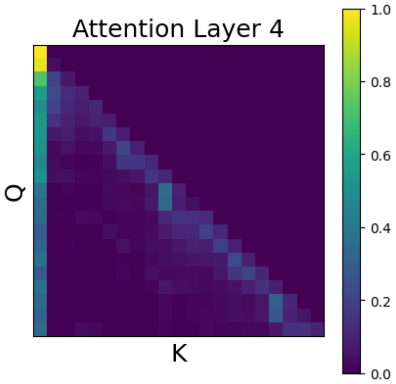
Figure 7: Coordinate values of EPE_1 for the first token (replicating the distribution described in section 3.2.3). Most coordinates are near zero; a small set exhibits extremely large magnitudes (“massive activations”).



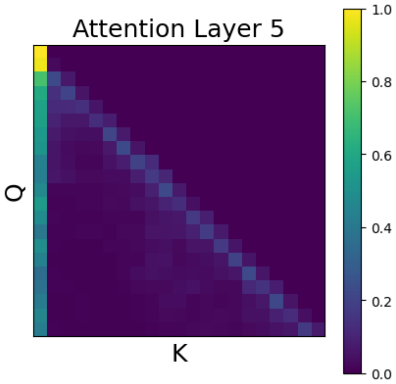
(a) layer 1

(b) layer 2

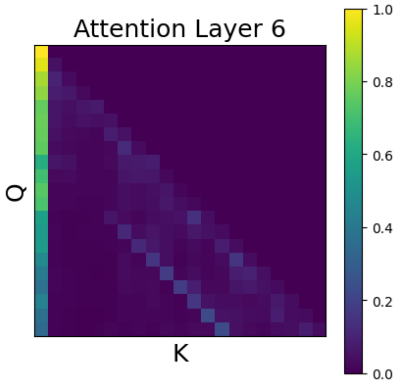
(c) layer 3



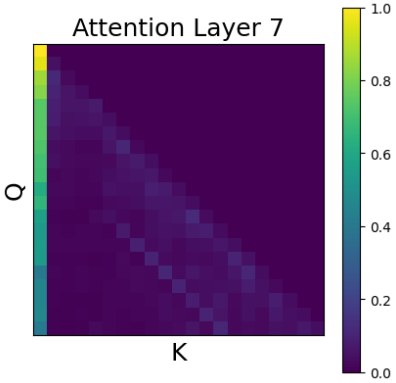
(d) layer 4



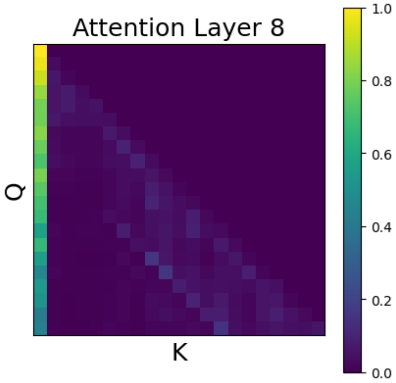
(e) layer 5



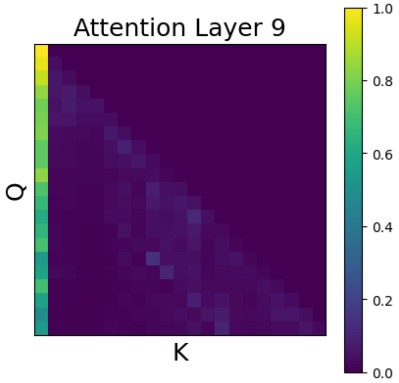
(f) layer 6



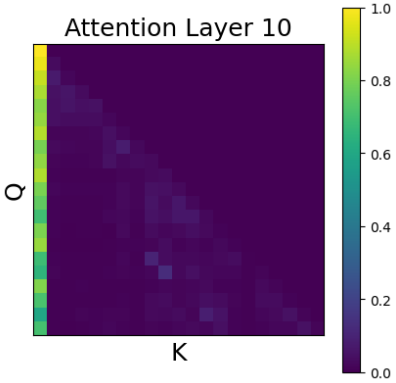
(g) layer 7



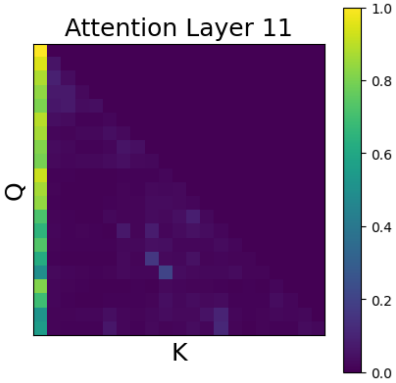
(h) layer 8



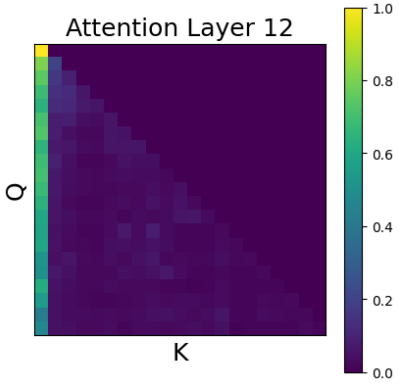
(i) layer 9



(j) layer 10

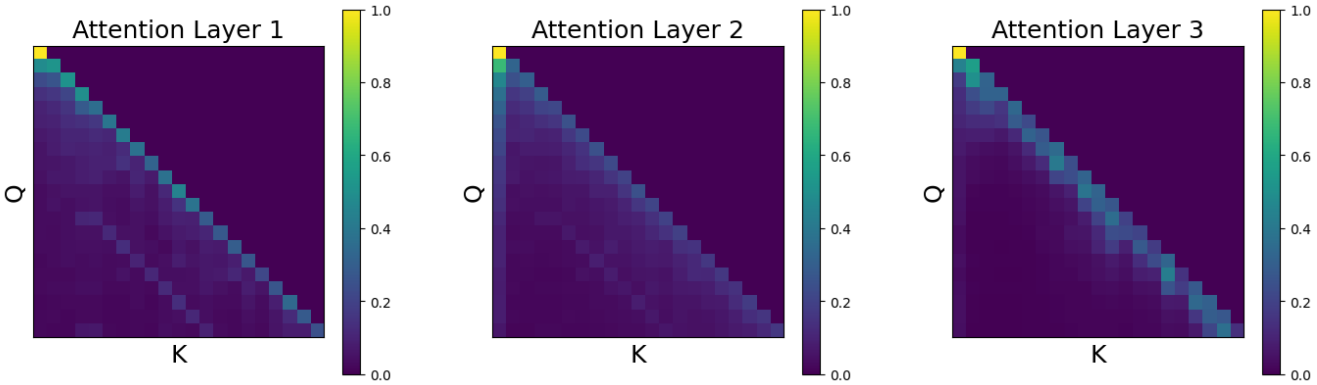


(k) layer 11



(l) layer 12

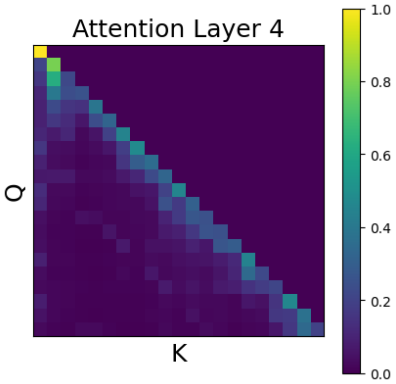
Figure 8: Attention maps for all layers with no intervention (replicating fig. 3a). A prominent first-position sink is visible in most layers.



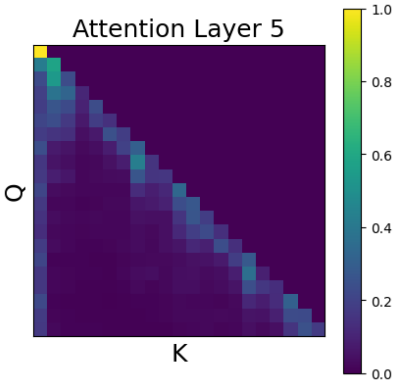
(a) layer 1

(b) layer 2

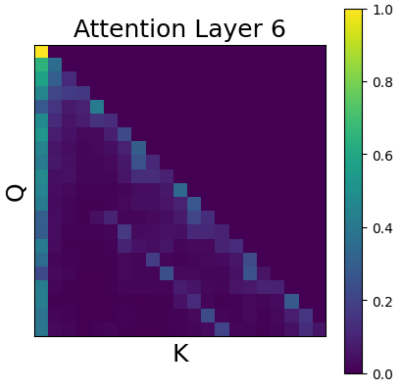
(c) layer 3



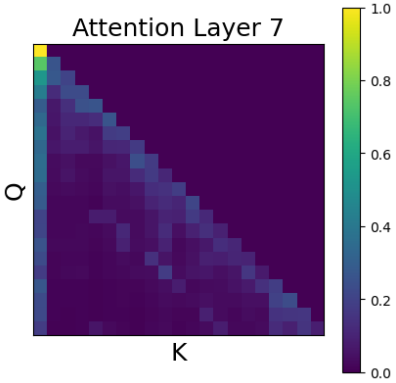
(d) layer 4



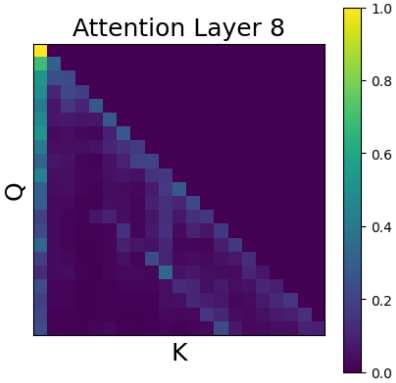
(e) layer 5



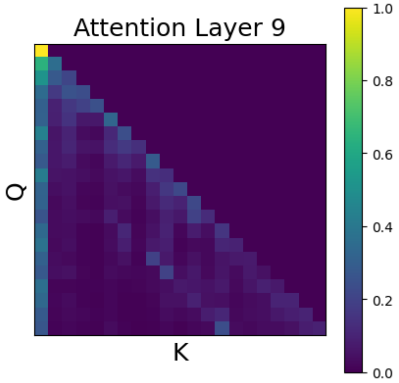
(f) layer 6



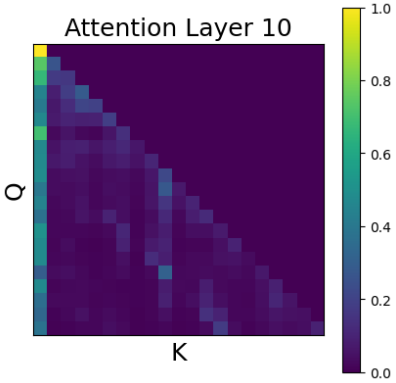
(g) layer 7



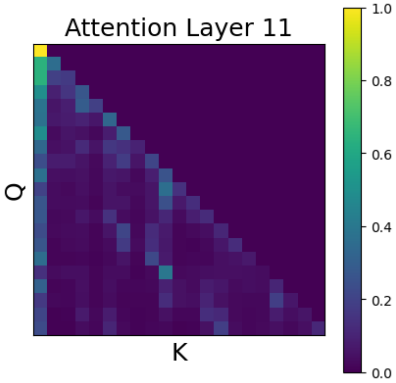
(h) layer 8



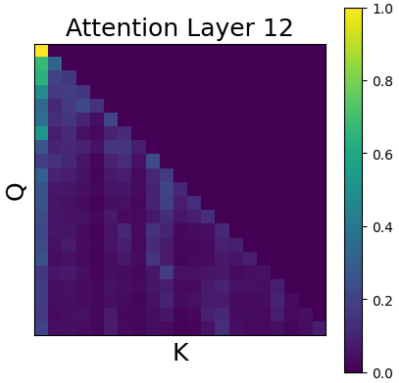
(i) layer 9



(j) layer 10

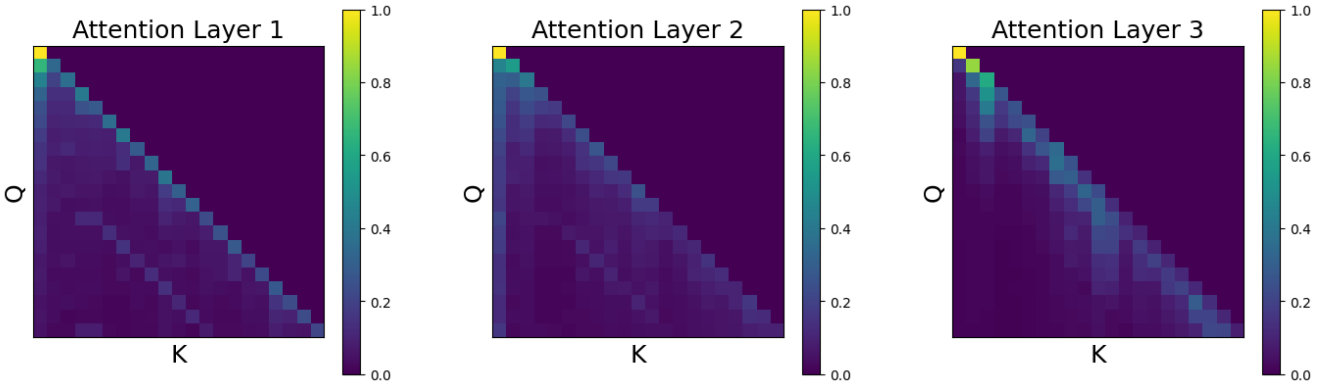


(k) layer 11



(l) layer 12

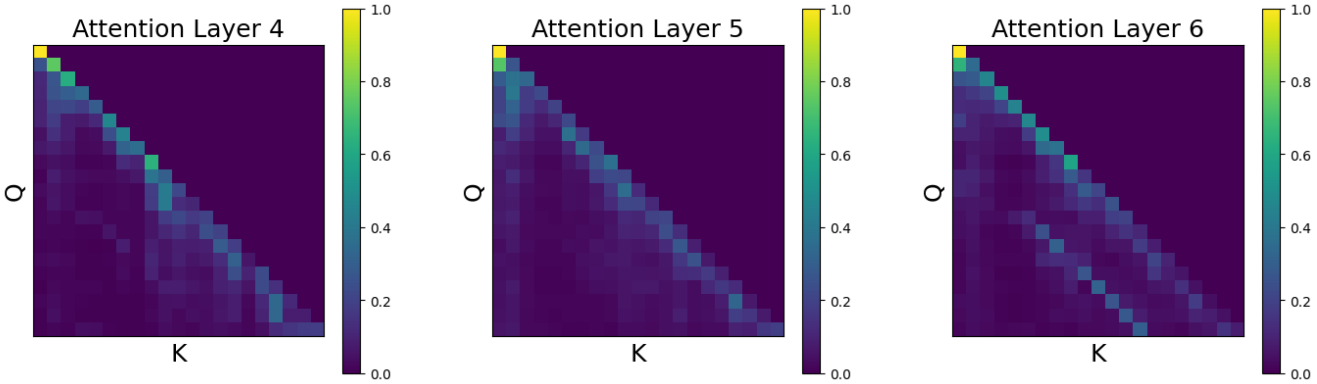
Figure 9: Attention maps for all layers with b_Q set to zero (replicating fig. 3b). The sink is substantially reduced across layers.



(a) layer 1

(b) layer 2

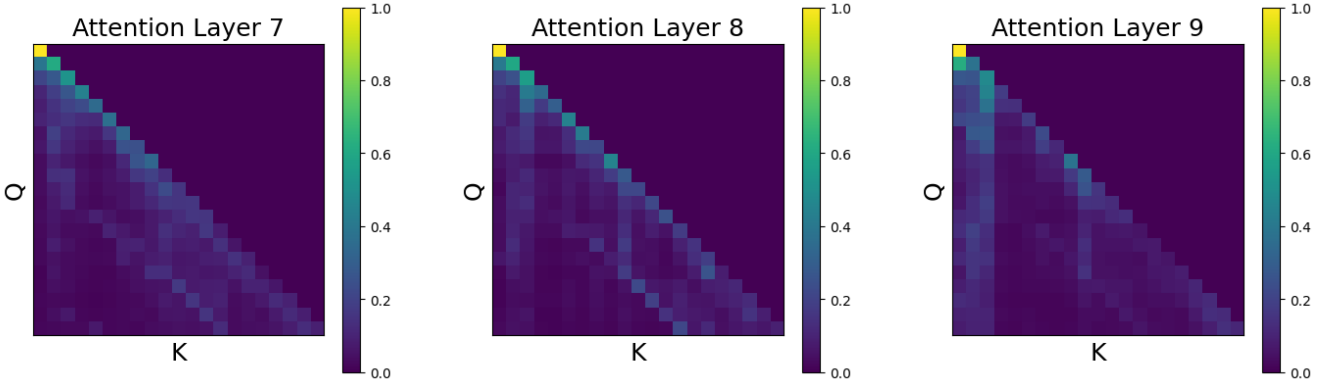
(c) layer 3



(d) layer 4

(e) layer 5

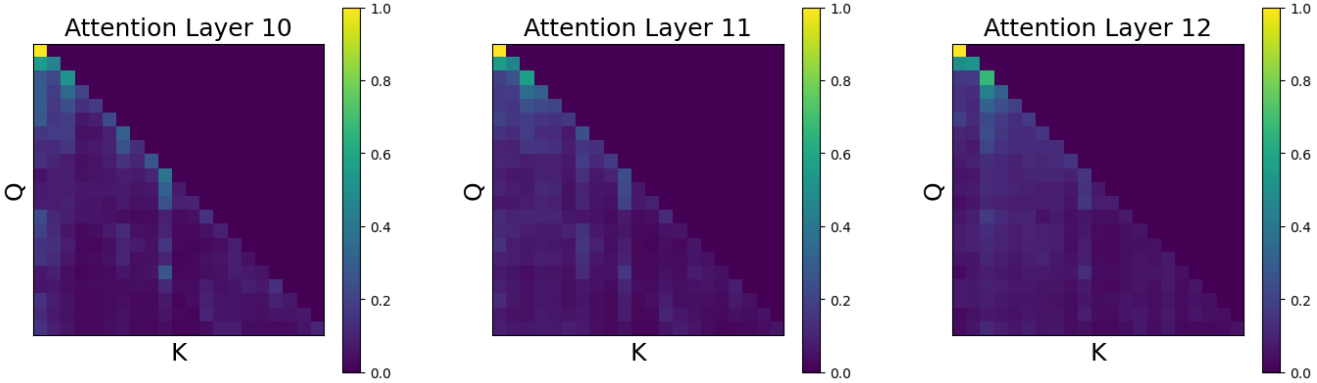
(f) layer 6



(g) layer 7

(h) layer 8

(i) layer 9

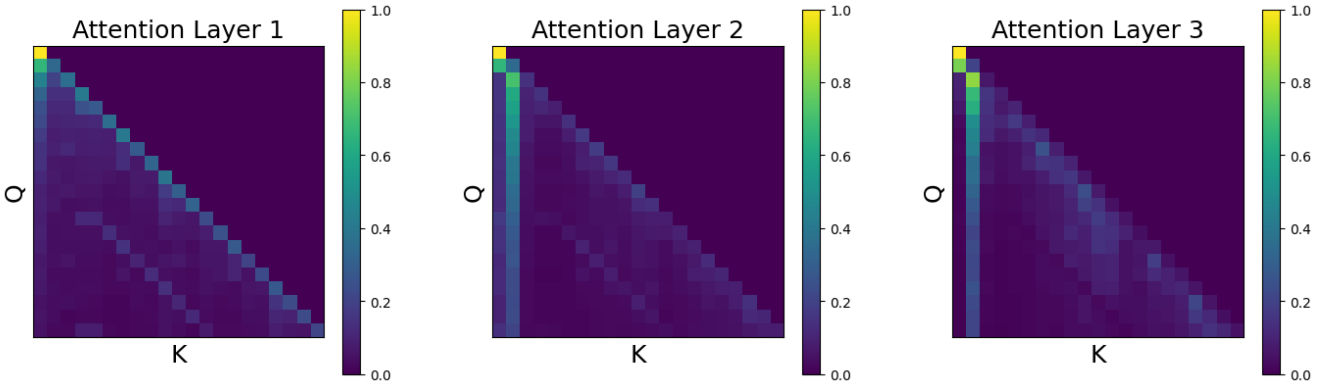


(j) layer 10

(k) layer 11

(l) layer 12

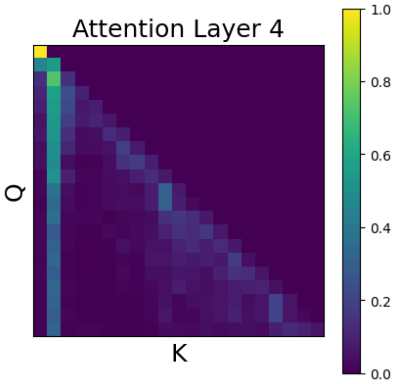
Figure 10: Attention maps for all layers after swapping EPE_1 with another position's EPE (replicating fig. 3c). The first-position sink disappears.



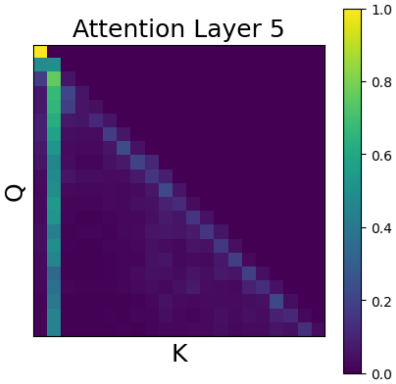
(a) layer 1

(b) layer 2

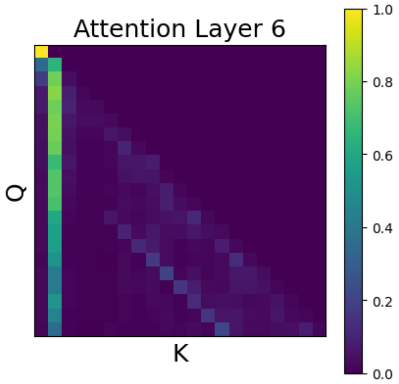
(c) layer 3



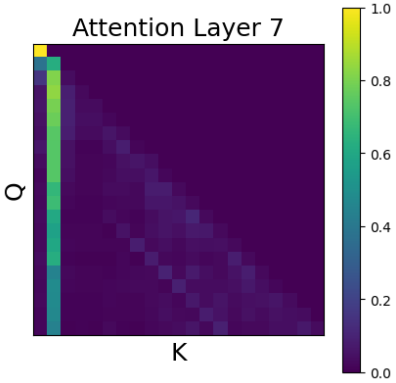
(d) layer 4



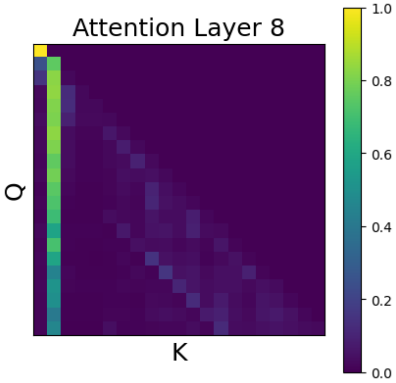
(e) layer 5



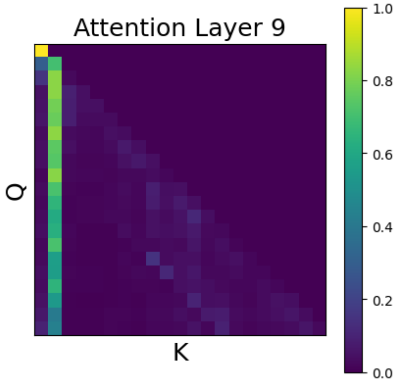
(f) layer 6



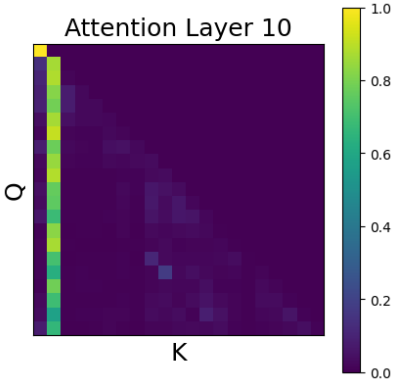
(g) layer 7



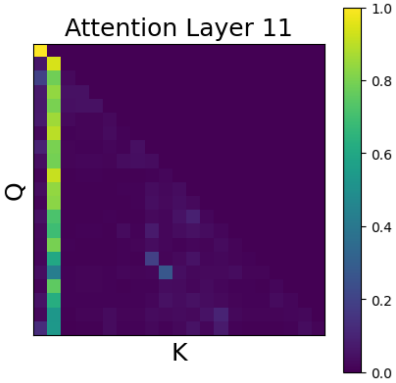
(h) layer 8



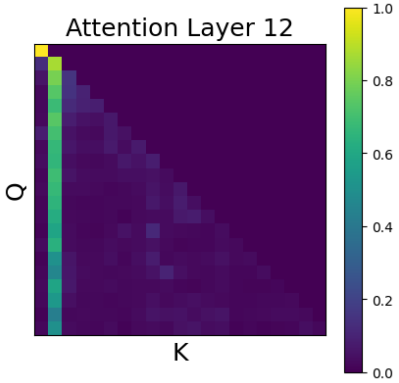
(i) layer 9



(j) layer 10

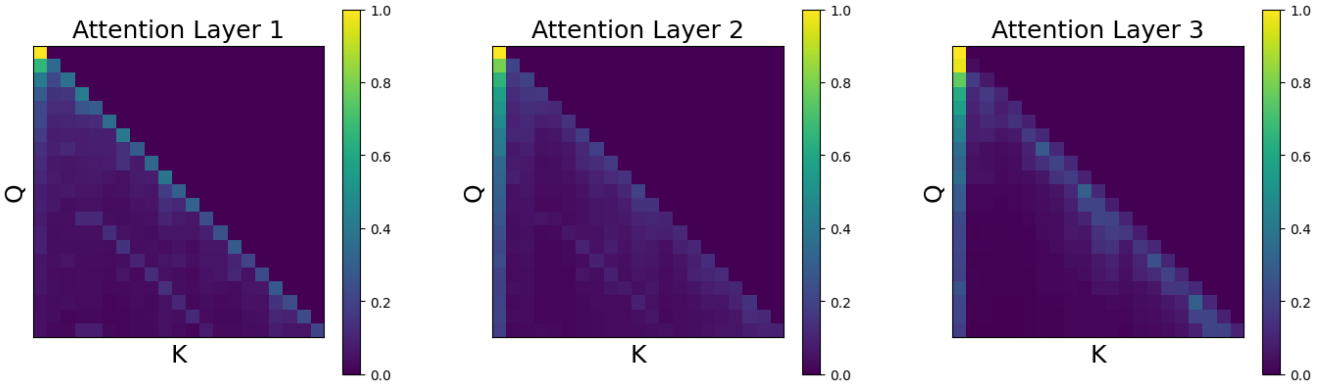


(k) layer 11



(l) layer 12

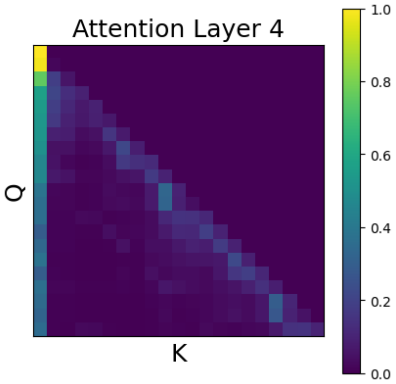
Figure 11: Attention maps for all layers after moving EPE_1 from position 1 to 2 (replicating fig. 3d). A strong sink forms at position 2.



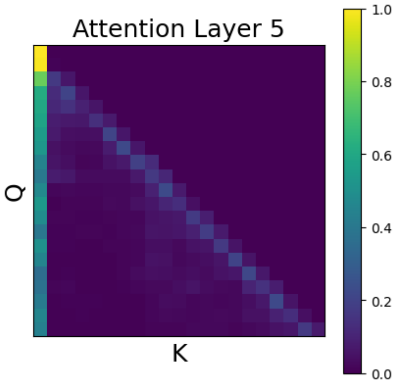
(a) layer 1

(b) layer 2

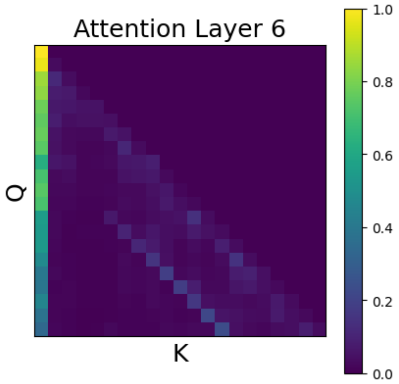
(c) layer 3



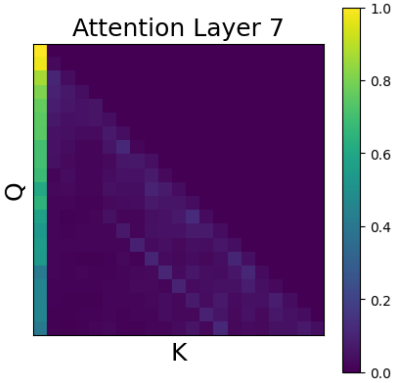
(d) layer 4



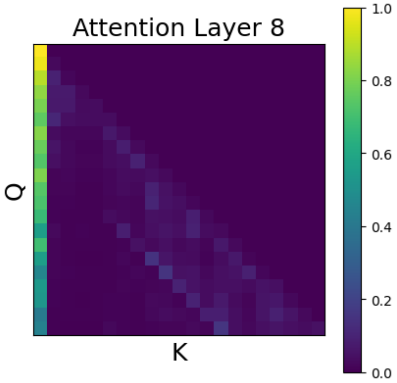
(e) layer 5



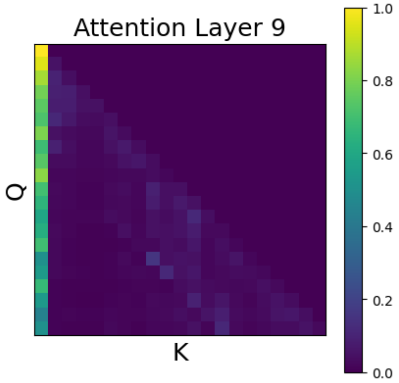
(f) layer 6



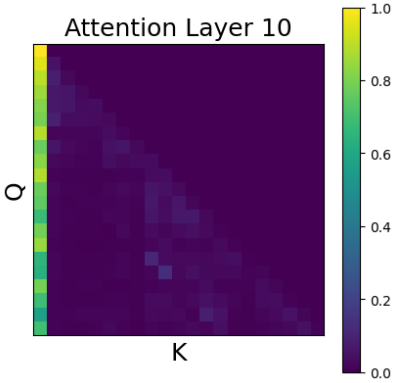
(g) layer 7



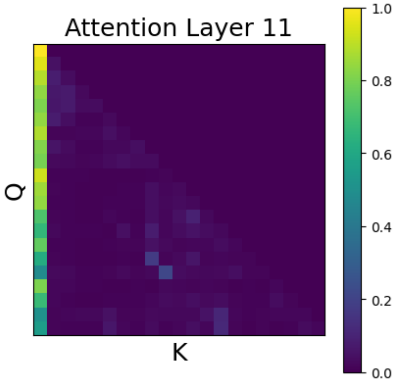
(h) layer 8



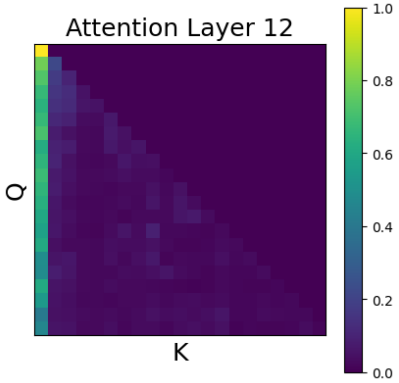
(i) layer 9



(j) layer 10

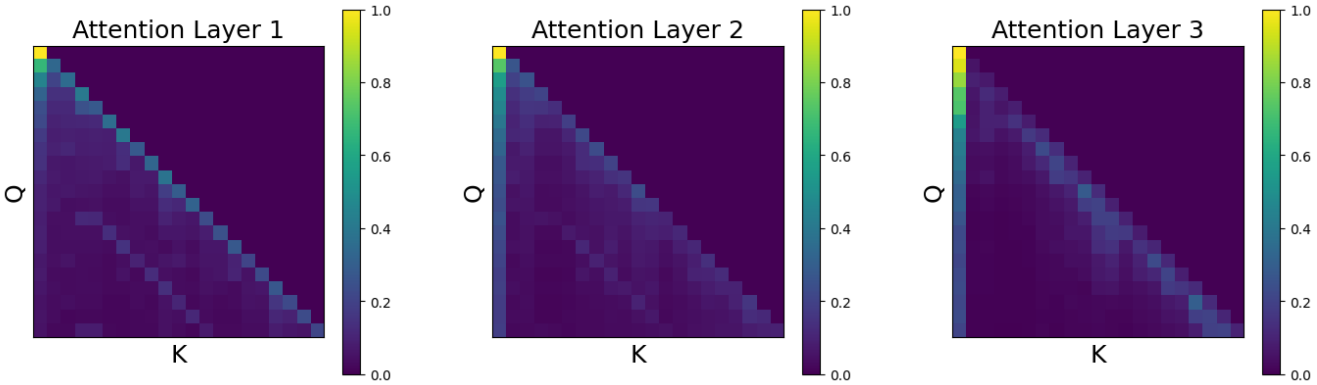


(k) layer 11



(l) layer 12

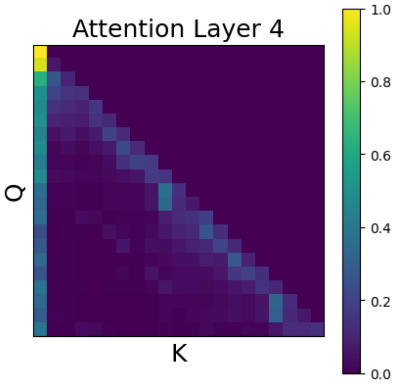
Figure 12: Attention maps for all layers after zeroing the BOS token embedding (replicating fig. 3e). The sink remains.



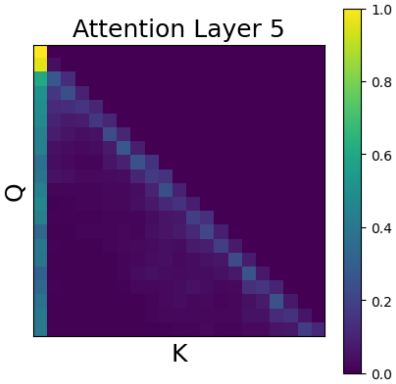
(a) layer 1

(b) layer 2

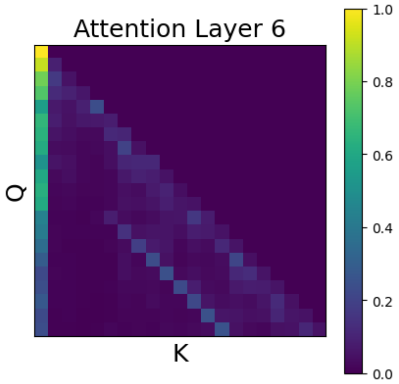
(c) layer 3



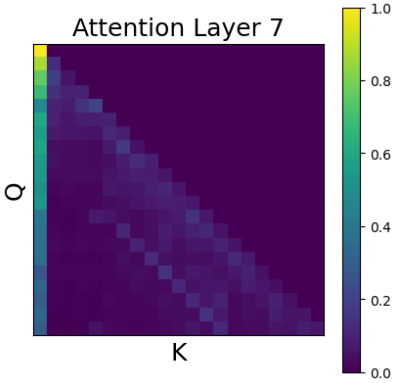
(d) layer 4



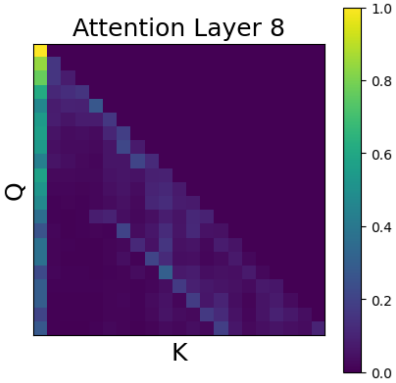
(e) layer 5



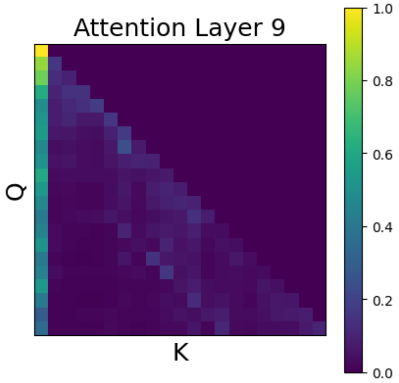
(f) layer 6



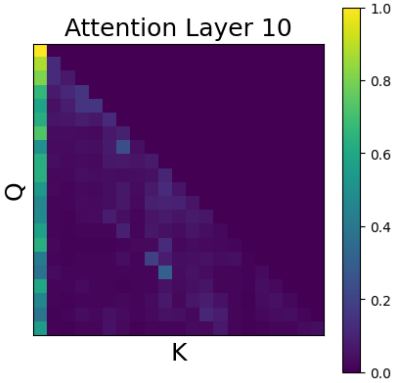
(g) layer 7



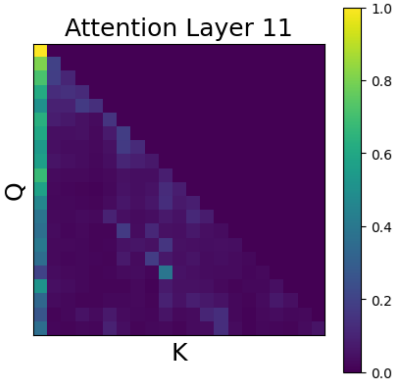
(h) layer 8



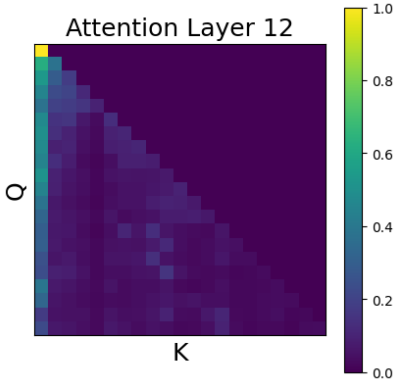
(i) layer 9



(j) layer 10

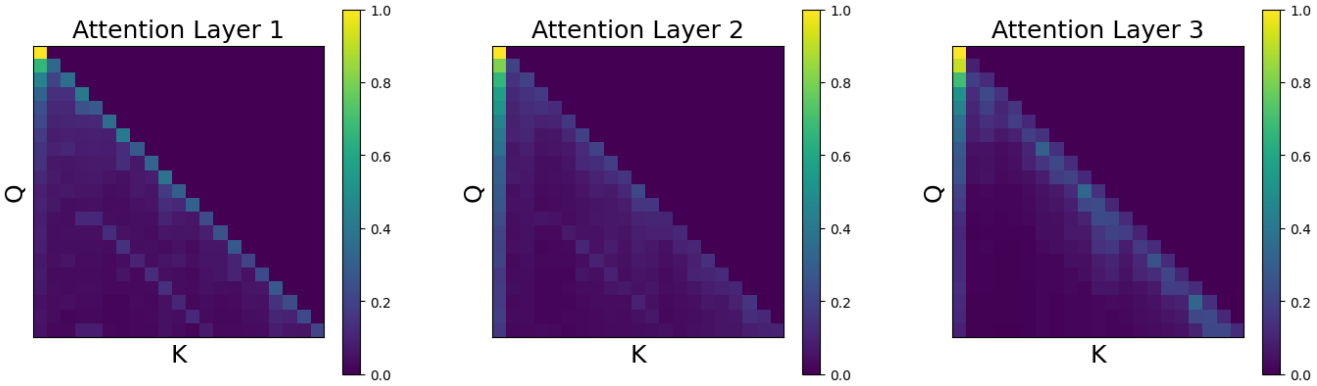


(k) layer 11



(l) layer 12

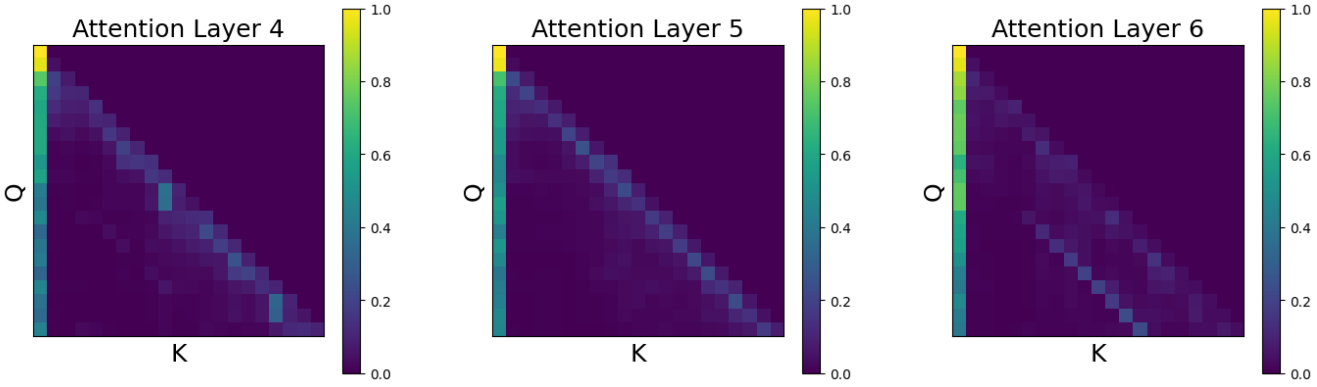
Figure 13: Attention maps for all layers after zeroing W_k at massive-EPE$_1$ coordinates (replicating fig. 3f). The sink is markedly reduced compared to random-coordinate zeroing.



(a) layer 1

(b) layer 2

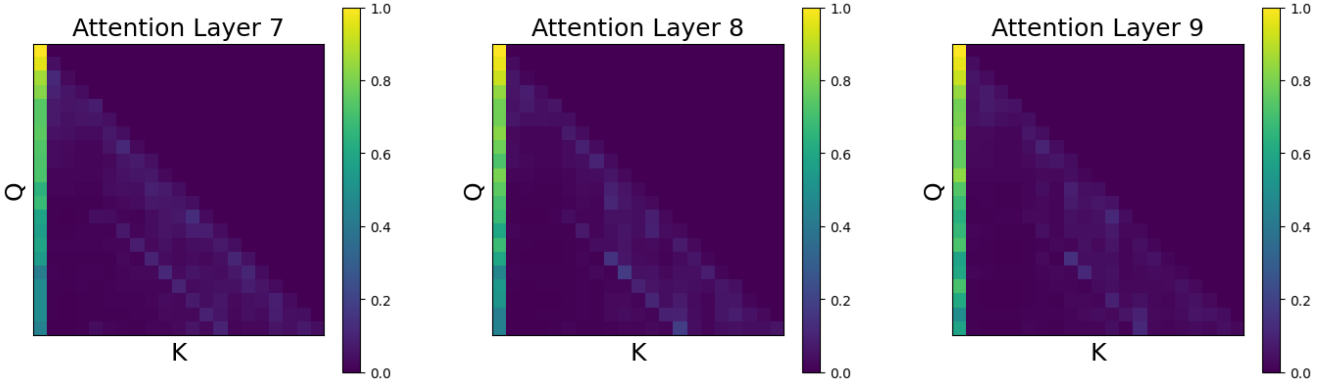
(c) layer 3



(d) layer 4

(e) layer 5

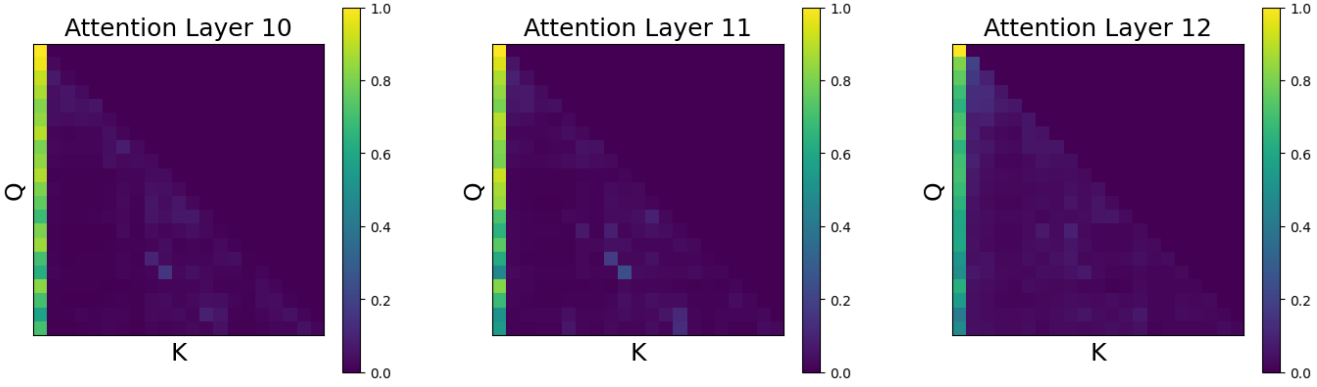
(f) layer 6



(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.