Sink vs. diagonal patterns as mechanisms for attention switch and oversmoothing prevention

Peter Súkeník [1] **Cristina López Amado** [1] **Christoph H. Lampert** [*1] **Marco Mondelli** [*1]

Abstract

This paper studies the role of sinks and diagonal patterns as attention switch and anti-oversmoothing mechanisms. We analyze geometric conditions under which sinks can be represented, showing a necessary alignment between the embedding of the sink and all other embeddings. Next, we refine the current understanding of the role of sinks in oversmoothing prevention: we specify the conditions under which dense attention provably smooths more than sparse attention, and empirically verify that such conditions are often satisfied in practice. We further prove an equivalence between sinks and hard attention switch, in which the output of the attention is identically 0. Finally, we relax the hard attention switch by allowing token self-communication: we provide a quantitative comparison of the costs of representing sinks vs. diagonal patterns, showing why sinks are favored in pretrained transformers. The introduction and analysis of diagonal patterns and the generalization of the attention switch close the gap between what oversmoothing prevention requires and what sinks provide, while also establishing when and why attention layers act like MLPs if token communication is not necessary.

1. Introduction

Softmax self-attention is the key component of transformers (Vaswani et al., 2017) that enables communication between tokens and, hence, makes it possible for non-recurrent architectures to solve sequential modeling tasks. Unlike in GNNs (Scarselli et al., 2008), the full-graph structure of attention allows for much denser and long-range communication patterns, a core feature that made transformers outperform vanilla RNN-like architectures (Bengio et al., 1993) in language modeling. Thus, it came as a surprise when (Xiao et al., 2024) revealed that many tokens in many attention heads do not communicate at all. More precisely, (Xiao et al., 2024) showed heads in which all tokens in a sequence idly attended to the first, beginning-of-sequence (BOS) token. This pattern was termed *attention sinks* due to the resemblance of visualized attention maps (a vertical line) to sinks. Sinks have since been reproduced in different language settings (Gu et al., 2025), vision transformers (Darcet et al., 2024), vision-language models (Luo et al., 2026), multi-modal models (Kang et al., 2025) and diffusion-based models (Wen et al., 2025; Rulli et al., 2025).

Sinks have received significant attention from both practitioners and theoreticians. Practitioners primarily studied how to take practical advantage of their existence or how to mitigate their emergence. (Xiao et al., 2024) used sinks explicitly to generate long-form content without any fine-tuning and suggested to assign a dedicated sink token during pre-training. Similar concepts were discussed in the follow-ups (Li et al., 2024; Cai et al., 2025). (Yu et al., 2024) presented mixed results where not all sinks improve the accuracy. On the negative side, sinks have been a target of blame in the context of resource effectiveness and interpretability (Kang et al., 2025), hallucination (Tu et al., 2026; Wang et al., 2025), quantization issues (Bondarenko et al., 2023), and training stability (Qiu et al., 2025). Thus, some works (Qiu et al., 2025; Gu et al., 2025) suggested methods to mitigate them, mostly by relaxing the softmax-normalization constraint.

Theoreticians focused on explaining *why* attention sinks emerge (see Section 2 for an overview). Some works (Barbero et al., 2024; 2025) claimed that sinks help transformers reduce oversmoothing and oversquashing. Others (Guo et al., 2025; Ran-Milo, 2026) connected sinks with the active-dormant mechanism (which we call attention switch), where the head is inactive on some tokens and outputs zero. However, existing oversmoothing results do not cover all architectural components of transformers, and they do not exclude other sparse attention patterns, such as diagonal attention. Similarly, (Guo et al., 2025) neither proves that sinks are the only mechanism for attention switch, nor it

[1]Institute of Science and Technology Austria (ISTA). Correspondence to: Peter Súkeník <peter.sukenik@ista.ac.at>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

argues why switching off attention means that it outputs 0, instead of e.g. acting as an MLP on idle tokens. Finally, it remains unclear how heads exhibit the sink from a geometric point of view. Our work closes the outlined gaps via the following contributions:

- Section 4 studies how sinks are represented. We show that, for the attention layer to be able to represent the sink, it suffices that all token embeddings are aligned (either positively or negatively) with the sink embedding. We empirically verify this is the case in pretrained transformers.
- Section 5 provides the first quantitative analysis of oversmoothing considering all architectural components. We analyze a single attention step and derive conditions under which increasing attention density increases or decreases token cosine similarity. Our analysis shows that, in principle, skip connections and value-output transformations may counteract the mixing induced by attention, even under uniform attention. However, measurements on trained LLMs show that this is rare in practice: in most heads, token similarity increases as attention becomes more uniform.
- While earlier work (Guo et al., 2025; Ran-Milo, 2026) defines an attention switch when the output on the residual stream is 0, we regard this as a hard attention switch and define a relaxation, dubbed the *soft attention switch*, which only prohibits inter-token communication, but doesn't force zero output. Soft switches allow for diagonal attention, which we empirically verify to be common in transformers.
- Section 6 shows that, under mild conditions (which we verify in pretrained LLMs), a head is a hard attention switch if and only if it is a sink. This equivalence goes beyond the results of (Guo et al., 2025), which suggest sinks only as a possible mechanism for hard attention switch and show the emergence of sinks close to convergence (i.e., when attention already exhibits some amount of sink behavior).
- Section 7 focuses on the soft attention switch, and it establishes a trade-off between sink and diagonal attention in terms of the ℓ_2 cost necessary to represent them. Because sinks are provably cheaper in a wide range of hyperparameters and distributions, this supports their larger relative prevalence w.r.t. diagonal attention in pretrained transformers.

2. Related work

Attention sinks and diagonal patterns. (Guo et al., 2025) argues that sinks implement the active-dormant mechanism and justifies the emergence of sinks in a near-convergence regime on a bigram-backcopy task. (Zhang et al., 2025) claims that sinks play a more active role in computation by tagging future tokens with a shared embedding which can be exploited geometrically in subsequent layers, e.g., for averaging. Similarly, (Behrens et al., 2024) shows that sinks are sometimes necessary for token counting and empirical histogram tasks. (Barbero et al., 2024; 2025) study sinks from the representational viewpoint and show that they help prevent oversmoothing and oversquashing, while (Queipo-de Llano et al., 2025) connects sinks with compression valleys arguing they are both manifestations of massive activations in the residual stream. (Ruscio et al., 2025) proposes a geometric interpretation in which sinks act as reference-frame anchors, simplifying the geometry and subsequent attention allocation. (Qiu et al., 2026) argues that attention and residual sinks rescale and refine non-outlier features and attention weights. The contemporaneouswork (Ran-Milo, 2026) proves that sinks are necessary in an adaptation of the bigram-backcopy task from (Guo et al., 2025). We note that this paper relies on the following assumption: if a pair of token embeddings x_t, x_q appears at positions t, q in a sequence with non-vanishing probability, then the same pair must appear with non-vanishing probability precisely at positions $2, 3$. Our theory in Section 6 explicitly takes into account positional embeddings, and it does not require such a strong assumption. Moreover, the main theorem of (Ran-Milo, 2026) can be recovered from the more general version of our Proposition 6.1 presented in Appendix C for context lengths smaller than dimension. Finally, while diagonal patterns have been discussed e.g. in (Hankemeier & Schilling, 2026; Zhai, 2026), to the best of our knowledge we are the first to explicitly consider them in the context of oversmoothing and sinks.

Oversmoothing. (Dong et al., 2021) proves that rank collapse occurs doubly exponentially with depth in multi-head transformers without skip connections, layer normalization, or MLPs. Furthermore, while the role of skip connections and MLPs is considered, (Dong et al., 2021) does not analyze their joint effect. (Geshkovski et al., 2023) shows that token representations converge to a clustered configuration for single-head attention without layer normalization or MLPs, assuming that query, key and value matrices are shared across layers. (Dovonon et al., 2025) analyzes a similar setting without such assumptions on learnable matrices, showing that oversmoothing can be avoided for some configurations. Oversmoothing at initialization is considered in (Noci et al., 2022; Giorlandino & Goldt, 2025). (Wu et al., 2024) shows that oversmoothing occurs in single-head self-attention without skip connections regardless of the attention mask, and layer normalization may prevent it. (Karagodin et al., 2024) extends the clustering results of (Geshkovski et al., 2023) to causal masking, while still making strong assumptions on learnable matrices. Related

work studies the loss of token information as context length increases (rather than as depth increases): (Saada et al., 2024) proves rank collapse at initialization for single-head attention without skip connections, layer normalization, or MLPs; (Barbero et al., 2024; 2025) study oversquashing, i.e., the phenomenon in which the output is less sensitive to specific input tokens. Overall, prior work lacks a characterization of causal self-attention beyond initialization that jointly accounts for normalization and skip connections under weak assumptions on learnable matrices, and our work addresses this gap.

3. Problem setup

Model and task. We use the shorthand $[n] := \{1, \dots, n\}$. Let $\|\cdot\|_2$ and $\|\cdot\|_F$ be Euclidean and Frobenius norms. We denote the all-one vector of length T by $\mathbf{1} \in \mathbb{R}^T$. For a matrix M, we denote its i-th row by $M_{i,:}$ and its j-th column by $M_{:,j}$ or M_j. We consider single-head causal self-attention transformers $f_\Theta = \mathrm{lin}_{L+1} \circ \mathrm{RMS} \circ \mathrm{B}_L \circ \mathrm{B}_{L-1} \circ \cdots \circ \mathrm{B}_1$. Each block is the composition of self-attention and MLP sub-blocks with residual connections: $\mathrm{B}_l = (\mathrm{id} + \mathrm{lin}_{l,2} \circ \sigma \circ \mathrm{lin}_{l,1} \circ \mathrm{RMS}) \circ (\mathrm{id} + \mathrm{ATTN}_l \circ \mathrm{RMS})$, with RMS denoting the RMSNorm layer (in our theoretical results we assume it has non-learnable parameters), id the identity mapping, lin a fully-connected layer, σ the ReLU layer and ATTN the single-head attention defined below. For an input embedding matrix $X^{(1)} = X_e + X_p \in \mathbb{R}^{d \times (T+1)}$, with X_e, X_p token and positional embeddings, $X^{(l)}$ is the output of the first $l-1$ blocks and $Z^{(l)} = \mathrm{RMS}(X^{(l)})$. An extension to relative positional embeddings is discussed in Section 7. The attention layer ATTN_l computes the (row-wise) attention score matrix $A^{(l)}_{ij} = \frac{\exp(S^{(l)}_{ij})}{\sum_{k=0}^{i} \exp(S^{(l)}_{ik})}$ if $j \le i$ and 0 otherwise, where $S^{(l)} = (W_Q^{(l)} Z^{(l)})^\top (W_K^{(l)} Z^{(l)})/\sqrt{d}$ is the score matrix and $W_Q, W_K \in \mathbb{R}^{d \times d}$ are query and key matrices. Then, $\mathrm{ATTN}_l(Z^{(l)}) = W_{VO}^{(l)} Z^{(l)} (A^{(l)})^\top$, where $W_{VO}^{(l)} \in \mathbb{R}^{d \times d}$ is the multiplication of the value and the output matrix. When focusing on a single layer, we drop layer-wise indexing. We consider transformers pretrained by minimizing $\mathbb{E}_{\rho(X)} \mathcal{L}(f_\Theta(X), Y) + \lambda \|\Theta\|_2^2$, where $\rho(X)$ is the training data empirical distribution and $\mathcal{L}$ is a loss (e.g., next-token cross-entropy).

Attention sink, diagonal attention and attention switch mechanisms. A head exhibits *attention sink*, if a fraction of context embeddings attends exclusively to BOS. This does not need to be global for the entire sequence. Instead, it suffices that individual tokens within the sequence attend to BOS. A head exhibits *diagonal attention*, if a fraction of context embeddings attends exclusively to themselves. As for the sink, this property is also defined on the individual token level as opposed to the sequence level. A head exhibits a *hard attention switch*, if its output on a fraction of context embeddings is zero. A head exhibits a *soft attention switch*, if its output on a fraction of context embeddings is either zero or equal to the context embeddings themselves.

Oversmoothing measure. We quantify token similarity via average cosine similarity. For a token matrix $X \in \mathbb{R}^{d \times T}$, the similarity between X_i and X_j is $\rho_{ij}(X) := \frac{(X_i)^\top X_j}{\|X_i\|_2 \|X_j\|_2}$, and the average cosine similarity of X is $\rho(X) = \frac{2}{T(T-1)} \sum_{i=1}^{T} \sum_{j=i+1}^{T} \rho_{ij}$. Other notions of oversmoothing studied in the literature include input similarity (Wang et al., 2022; Wu et al., 2024) and rank collapse (Shi et al., 2022; Noci et al., 2022; Giorlandino & Goldt, 2025) (see (Dovonon et al., 2025) for details).

Empirical head pattern labeling. We empirically classify a head as a certain type when the relative attention mass on the corresponding pattern exceeds a threshold (e.g., a head is a *sink* if BOS attention weights make up $\ge 40\%$ of the total mass). For dual-pattern heads, e.g. sink–lower-diagonal heads, we require both the joint pattern mass to exceed the threshold and the sub-patterns to be relatively balanced (e.g., the lower-diagonal mass is $\ge 10\%$ of the joint sink–lower-diagonal mass, and the joint mass is $\ge 60\%$ of the total mass).

4. Warm-up: How transformers represent sinks

While sinks frequently appear in many heads per block (Ivanitskiy et al., 2025; Gu et al., 2025), it is not a priori clear when the sink can be geometrically represented (e.g., if two embedding vectors are opposite to each other, they cannot be both positively aligned with the sink's key). To address this, the result below provides conditions under which a group of tokens attends to the sink. The proof is deferred to Appendix A.

Proposition 4.1. *Let $Z \in \mathbb{R}^{d \times (T+1)}$ be a matrix of input embeddings $z_0, z_1, \dots, z_T$, and let $\mathcal{J} \subset [T]$ be a set of indices. Then, there exists $W \in \mathbb{R}^{d \times d}$ such that $z_j^T W z_0 > z_j^T W z_i$ for all $j \in \mathcal{J}$ and $i \in [T]$ if and only if the two sets of vectors $\{z_0 - z_i,\, i \in [T]\}$ and $\{z_j,\, j \in \mathcal{J}\}$ lie in half-spaces.*

The existence of W s.t. $z_j^T W z_0 > z_j^T W z_i$ for all $j \in \mathcal{J}$ and $i \in [T]$ is equivalent to the head being able to represent a sink. Thus, Proposition 4.1 shows that, if the sink is representable, one can project the sink furthest away from other embeddings. Moreover, context embeddings attending to the sink must lie in a cone. A simple sufficient condition is $z_0^T z_i \le 0$ for all $i \in [T]$, i.e., context embeddings are jointly rotated against the BOS embedding.

In Figure 1, we verify empirically that this is indeed the case for all middle layers of three pretrained transformers (gpt2-small, Llama 3.1-8B and Gemma 7B). We measure the cosine similarity between BOS embedding and all other

embeddings across multiple sequences and datasets. The maximum over all cosine similarities does not cross 0 in the inner layers of the models, allowing each attention head to produce any sink pattern.

5. Sparse attention patterns as an oversmoothing prevention mechanism

In this section, we investigate whether sparse attention is necessary to avoid oversmoothing, or instead other transformer components can counteract the effect of near-uniform attention patterns. As shown in Appendix B.1, in the presence of skip connections and layer normalization, for a fixed token matrix, one can choose $W_{VO}^{(l)}$ to prevent oversmoothing even under uniform attention. However, the key question is whether a single choice of $W_{VO}^{(l)}$ prevents oversmoothing over the full range of token matrices arising from the data distribution, rather than only for particular inputs.

Theoretical analysis. We study how token similarity changes after a single step of attention, under a distributional assumption on token representations. For simplicity, we omit the layer index l ($X_i^{(l)} \equiv X_i$, $W_{VO}^{(l)} \equiv W$). As token representations tend to exhibit positive cosine similarity across layers (see Figure 4 in Appendix E.1), we model the normalized token representations as $Z_i = \bar{z} + \epsilon_i$. Here, $\bar{z}$ is the average token representation and ϵ_i is such that $\mathbb{E}[\epsilon_i] = 0$. In addition, we assume that each unnormalized token can be written as $X_i = \beta Z_i$. To study how the sparsity of the attention matrix affects oversmoothing, for $\lambda \in [0, 1]$ we define $Y(\lambda) = \beta Z + WZ((1-\lambda)I + \lambda A_u)^\top$, where A_u is the $T \times T$ uniform causal attention matrix, i.e., $(A_u)_{ij} = 1/i$ for $j \leq i$. The matrix $(1-\lambda)I + \lambda A_u$ interpolates between the identity ($\lambda = 0$) and uniform attention ($\lambda = 1$). Accordingly, $Y(\lambda)$ represents the token matrix after one attention update with skip connection, when the attention pattern varies from no mixing to uniform causal mixing, while keeping token representations and weight matrix fixed. This isolates the effect of attention sparsity on token mixing. We approximate the expected pairwise cosine similarity of the token matrix $Y(\lambda)$ by

$$\mathbb{E}[\rho_{ij}(Y(\lambda))] \approx \hat{\rho}_{ij}(Y(\lambda)) := \frac{\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)]}{\sqrt{\mathbb{E}[\|Y_i(\lambda)\|^2]}\sqrt{\mathbb{E}[\|Y_j(\lambda)\|^2]}}. \tag{1}$$

This yields the approximation $\hat{\rho}(Y(\lambda)) = \frac{2}{T(T-1)} \sum_{i=1}^{T} \sum_{j=i+1}^{T} \hat{\rho}_{ij}(Y(\lambda))$ for the average cosine similarity. This approximation is expected to be accurate in high dimensions when the bilinear terms $Y_i(\lambda)^\top Y_j(\lambda)$, $\|Y_i(\lambda)\|^2$, $\|Y_j(\lambda)\|^2$ are sufficiently concentrated around their means.

The result below provides sufficient conditions under which $\hat{\rho}(Y(\lambda))$ increases/decreases with λ, i.e., as the attention pattern becomes denser. The proof is deferred to Appendix B.2.

Theorem 5.1. *Assume $\mathbb{E}[\epsilon_i \epsilon_i^\top] = \Sigma_V$ for all $i \in [T]$ and $\mathbb{E}[\epsilon_j \epsilon_i^\top] = \Sigma_C$ for all $i \neq j$. Let $B := \Sigma_V - \Sigma_C$ and $C = \bar{z}\bar{z}^\top + \Sigma_C$. Then, for all $\lambda \in [0, 1]$ and $i < j$, $k \in [T]$, we have*

$$\begin{aligned}
&\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)] = \beta^2 \operatorname{tr}(C) + 2\beta \operatorname{tr}(CW) + \operatorname{tr}(CW^\top W) \\
&+ \frac{\lambda}{j} \left(\beta \operatorname{tr}(BW) + \operatorname{tr}(BW^\top W)\right), \\
&\mathbb{E}[\|Y_k(\lambda)\|^2] = \beta^2 \operatorname{tr}(B + C) + 2\beta \operatorname{tr}((B + C)W) \\
&+ \operatorname{tr}((B + C)W^\top W) - 2\left(1 - \frac{1}{k}\right)(\beta \operatorname{tr}(BW) \\
&+ \operatorname{tr}(BW^\top W))\lambda + \left(1 - \frac{1}{k}\right) \operatorname{tr}(BW^\top W)\lambda^2.
\end{aligned} \tag{2}$$

Thus, assuming $\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)] > 0$ $\forall i, j \in [T]$, we have that (i) *if $\beta \operatorname{tr}(BW) > 0$, $\hat{\rho}(Y(\lambda))$ is strictly increasing in λ; and* (ii) *if $\beta \operatorname{tr}(BW) + \operatorname{tr}(BW^\top W) < 0$, $\hat{\rho}(Y(\lambda))$ is strictly decreasing.*

This result shows that denser attention does not, by itself, imply stronger oversmoothing, and the behavior depends on W. Compared to (Karagodin et al., 2024) which proves a clustering result for $W = I$, we show that W can counteract the mixing effect of attention. Theorem 5.1 also sheds light on the role of skip connections in mitigating oversmoothing, which has been discussed in the literature (Dong et al., 2021; Dovonon et al., 2025; Noci et al., 2022). This effect depends jointly on the skip-connection strength, the token distribution, and the transformation W: in the absence of skip connections ($\beta = 0$), condition (ii) (i.e., $\beta \operatorname{tr}(BW) + \operatorname{tr}(BW^\top W) < 0$ implying $\hat{\rho}(Y(\lambda))$ decreasing) never holds; furthermore, when $\operatorname{tr}(BW) < 0$, strengthening the skip connection decreases $\beta \operatorname{tr}(BW) + \operatorname{tr}(BW^\top W)$, which favors $\hat{\rho}(Y(\lambda))$ to be decreasing. A further interpretation of these conditions is in Appendix B.2 (Remark B.4). We note that the requirement $\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)] > 0$ is not restrictive, since token representations typically have positive cosine similarity across layers. While generalizations to position-dependent scaling ($X_i = \beta_i Z_i$) and general covariances ($\mathbb{E}[\epsilon_j \epsilon_i^\top] = \Sigma_{ij}$) are possible, the experiments discussed below show that the approximation obtained under the current assumptions is already very close to the empirical quantity.

Experimental analysis. We empirically assess whether the approximation in (2) matches the empirical average cosine similarity. Then, we study whether there are weight matrices W that, even under uniform attention, counteract the mixing induced by attention, as part *ii* of Theorem 5.1 suggests. We consider four models (LLaMA3-8B, Gemma-7B, GPT2-XL, Mistral-7B), average across heads and over a batch of 50

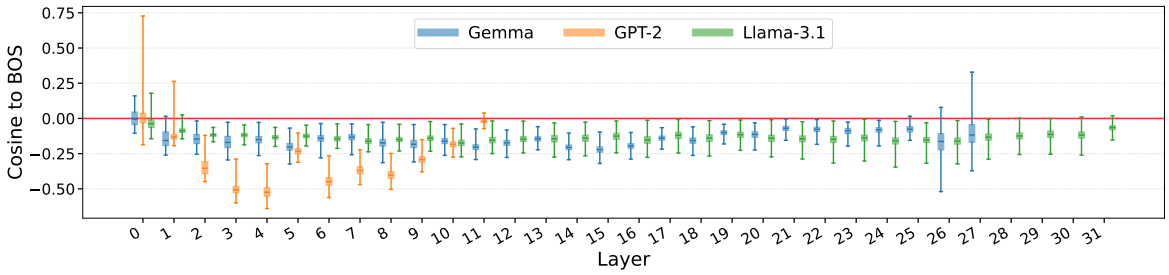

Figure 1. Cosine similarity between BOS and all other 256 token embeddings at the input to the attention layer, pooled across heads and 50 sequences across 4 language datasets (`TinyStories` (Eldan & Li, 2023), `tinyshakespeare` (Karpathy, 2015), `WikiText` (Merity et al., 2017a), and `CodeSearchNet-Python` (Husain et al., 2019)). The whiskers of the box plot point to absolute minimum and maximum.

sequences. We report results for the C4 dataset (Raffel et al., 2020), and show in Figure 5 in Appendix E.1 that the same conclusions hold for other datasets. For each attention head, W is the composition of learnable normalization scaling, value matrix, and output projection matrix.

The left two plots of Figure 2 show an excellent agreement between empirical average cosine similarity and the approximation (2) . This holds both for X and $X + WZ$, across several models and context lengths. In fact, the approximation remains accurate not only after averaging across heads, but also for most individual attention heads, see Figure 7 in Appendix E.1.

The right two plots of Figure 2 show the empirical average cosine similarity of token representations in LLaMA3-8B for different attention components. The attention contribution is rescaled to match the magnitude of the original token matrix, since in practice contributions of different heads are summed within each layer (see (11) in Appendix E.1). Comparing curves corresponding to X, $X + WZ$ and $X + WZA^\top$, we note that the application of W decreases the average cosine similarity, which is then increased by the attention matrix. On average, the effect of the attention matrix is slightly stronger than that of W, i.e., the cosine similarity for $X + WZA^\top$ is larger than that for X. Comparing third and fourth plots, we note that the effect of the attention matrix is more pronounced (i.e., the gap between the curves for $X + WZA^\top$ and X is bigger), when averaging only over heads closer to being uniform. Finally, the curve for $X + WZA_u^\top$ is obtained by replacing the attention matrix with a uniform one, which leads to a significant increase in cosine similarity. Figure 6 in Appendix E.1 shows that this behavior is consistent across models and datasets, and Figure 10 provides a finer head-wise view of this phenomenon. Figures 8-9 show that *(a)* the monotonicity of $\hat{\rho}(Y(\lambda))$ predicted by Theorem 5.1 replicates well in practice, and that *(b)* roughly half of the heads satisfy condition (i) (which leads to $\hat{\rho}(Y(\lambda))$ increasing), whereas few heads satisfy condition (ii) (which leads to $\hat{\rho}(Y(\lambda))$ decreasing). This suggests that oversmoothing increases as attention becomes more uniform.

Summary. Sparse attention is not the only way to prevent oversmoothing: skip connections and value-output transformations can counteract the effect of attention. In trained transformers, however, this rarely happens, and denser attention typically increases oversmoothing.

6. Equivalence between sink and hard attention switch

In this section, we prove that in globally optimal transformers, the sink is the *only* possible mechanism for hard attention switch. We generalize the input-output distribution of (Guo et al., 2025): for each context position t, let the set of possible input context embeddings x_t be $\mathcal{S}_t = \mathcal{D}_t \cup \mathcal{C}_t$, where embeddings in $\mathcal{D}_t$ are dormant (i.e., the output of the head on them is 0) and embeddings in $\mathcal{C}_t$ are copy-paste (i.e., their output on the head is any past context embedding x_s except BOS). We assume $W_{VO}x_0 = 0$ as in (Guo et al., 2025) which also supports it empirically (see Figure 1(b) therein). While (Guo et al., 2025) claims that sink is *the* solution for hard attention switch, there could, in theory, be other solutions that also implement it. In particular, a context embedding that has zero update in the attention layer could: *(i)* attend to a convex combination of past context embeddings that sums up to zero, or *(ii)* attend to itself while having zero value embedding. In the result below, we identify geometric conditions guaranteeing that neither of these two options is possible and, therefore, the only solution to switch off the head is to use the sink.

Proposition 6.1. *Consider the input-output distribution above. Assume that, for all input embedding sequences, either the sequence is full column-rank or it lies in a strict half-space. Assume also that the rank of the span of all output embeddings equals the rank of the span of all input embeddings. Then, in order to realize the hard attention switch, dormant tokens have to attend fully to BOS.*

This result reveals why the hard attention switch has to be implemented via sinks: tokens cannot mix up value embeddings of past tokens or themselves to zero, because their own context embeddings might be used in the future.

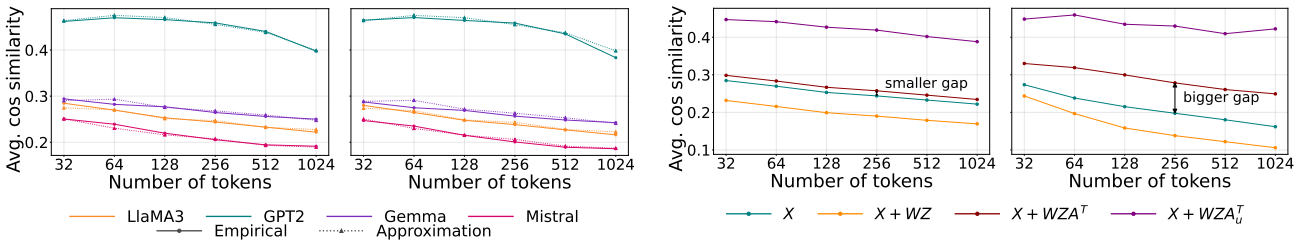

Figure 2. **Left:** empirical average cosine similarity (solid) and theoretical approximation (2) (dashed) for X (first plot) and $X + WZ$ (second plot), across models and context lengths. **Right:** empirical average cosine similarity in LLaMA3-8B for different components of the attention step, averaging over all heads (first plot) and over heads with uniformity coefficient (see (12) in Appendix E.1) larger than 0.6 (second plot). Cosine similarity generally increases as the attention matrix becomes denser.

Sink is then the only option if either *(i)* input embeddings lie in a strict half-space, or *(ii)* both input embeddings and the value matrix on the span of input embeddings are full rank. Remarkably, both these conditions empirically hold in pretrained transformers: the first is verified in Figure 1, and the second in Figure 12 in Appendix E.2. We also note that a sufficient condition for the weight matrix to be full rank under the statement's conditions is that all embedding dimensions are useful (i.e., copy-pasted) for some future context tokens. Finally, while Proposition 6.1 shows that the hard attention switch must be implemented via the sink, the converse statement is obvious: in optimal transformers, if the head exhibits an attention sink (with zero BOS value embedding), then it must implement the attention switch. In fact, the only alternative is that all output embeddings of the head are zero, which contradicts the optimality of the transformer.

Proposition 6.1 follows from the more general Theorem C.1 which does not require a precise copy-paste pattern, and it directly implies that the sink is the only solution for the distribution of (Guo et al., 2025), see Corollary C.3. Both Theorem C.1 and Corollary C.3 are in Appendix C. There, we discuss the expected sample complexity due to the span equality condition (Remark C.2), and we show that heads implementing the behavior predicted by our theory are ubiquitous in pretrained LLMs (Remark C.4).

Summary. Sinks are the only mechanism that can implement the hard attention switch under mild geometric assumptions (which empirically hold in pretrained transformers) and for a rather general input-output distribution (which includes the ones from (Guo et al., 2025; Ran-Milo, 2026)).

7. Sink vs. diagonal patterns for soft attention switch

Section 6 proves that sinks are necessary for *hard* attention switch. However, Section 5 only gives evidence that *dense* attention patterns (and, hence, significant inter-token communication) lead to oversmoothing. This leaves a gap between controlling oversmoothing and the current notion of attention switch. Note that oversmoothing would be prevented if the transformer implemented a ResNet by setting attention weights on the diagonal to 1 and the rest to 0. This diagonal pattern can be interpreted as an attention switch-off, since there is no inter-token communication.

Similarly to sinks, diagonal attention is common in pretrained LLMs. Across the sequences and datasets considered in Figure 1, we find that $6\%, 0.6\%$ and 9.5% of the heads of gpt2-medium, Llama 3.1-8B and Gemma 7B have $\geq 40\%$ of the attention mass on the main diagonal (in Llama the number represents $> 10\%$ of all non-sink heads). Thus, we introduce a notion of *soft* attention switch allowing for both sink and diagonal attention, and we study the trade-off between these two patterns.

The first aspect in this trade-off is the ℓ_2 norm of the matrix $W_{QK} \equiv W_Q^T W_K$ representing either sink or diagonal attention. For simplicity, assume the sequence is the identity I_{T+1}. To represent the sink, it suffices to set $W_{QK} = \kappa \mathbf{1}_{T+1} e_1^T$, with e_1 the first element of the canonical basis and κ large enough to make attention weights approximately 1 after the softmax. This matrix has a single non-zero singular value given by $\kappa\sqrt{T+1}$. To represent the diagonal attention, one needs to set $W_{QK} = \kappa I_{T+1}$, which has $T+1$ singular values equal to κ. As the regularization cost of W_{QK} is its nuclear norm (Kobayashi et al., 2024), the costs of sink and diagonal attention are respectively $\kappa\sqrt{T+1}$ and $\kappa(T+1)$. Hence, the sink is cheaper due to its low-rank structure.

The second aspect of the trade-off is the re-use of resources in fully-connected layers. If the attention pattern is diagonal, the switched-off tokens receive a non-trivial update in the residual stream, which is missing in the sink case. There are then two options. *(i)* The update is *useful* for tokens in the following layers. Here the sink, which ignores the update, compensates via the MLP and incurs an extra cost. Hence, there is a non-trivial trade-off between the two patterns, depending on the token geometry and the role of the head. *(ii)* The update is *not useful* as it is inconsistent with messages sent to the following tokens. Here the diagonal attention is not desirable even when it is cheaper, as it harms the next-token prediction capabilities. Hence, there is a clear

preference towards the sink.

In the following, we focus on the case of *useful* updates where there is no clear preference towards the sink. We analyze two settings with different generality and compare the costs of a single-head attention block that represents the same task via either the diagonal or the sink attention patterns.

Large context lengths favor sinks for the backcopy task. We adapt the bigram-backcopy task from (Guo et al., 2025) to better suit a single-block optimality analysis (rather than an end-to-end result for transformers with a single layer) – keeping the backcopy part of the distribution which remains well-defined, while dropping the irrelevant bigram part. More precisely, let the input embedding be $x_{ti} = \sqrt{d/2}(h_i + p_t)$, where h_i is the i-th context embedding and p_t the positional embedding. All h_i's and p_t's are assumed to be orthonormal. The context embeddings are split into two groups: the dormant set $\mathcal{D}$ and the copy-paste set $\mathcal{C}$ – plus the BOS, always indexed by 0. The output of the block on the BOS token x_{00} equals x_{00} itself; the output for a dormant embedding x_{tc} equals $2x_{tc}$; and the output for a copy-paste embedding x_{tc} equals $x_{tc} + x_{t-1,i}$, where i is the index of the preceding context. We say that the block implements a sink if the dormant embeddings attend to the BOS in the attention layer. We say it implements a diagonal attention if the dormant embeddings attend to themselves. As above, let κ be large enough so that, if the dominant attention logit is larger than all other logits by at least κ, we can treat the dominant weight after softmax as exactly 1. Note that heads of this type occur in pretrained transformers, as displayed in Figure 13 in Appendix E.3.

Theorem 7.1. *Let a single-head transformer block represent exactly the input-output distribution above. Let c_1 be such that $c_1 \geq d/(T + |\mathcal{C}| + |\mathcal{D}| + 2)$. We have that, if*

$$\begin{aligned} &\frac{2\kappa}{\sqrt{d}}\left(\sqrt{2|\mathcal{D}|} + T - 1\right) + 3T + 3|\mathcal{D}| + |\mathcal{C}| \\ &+ 4c_1(T + |\mathcal{C}| + |\mathcal{D}| + 2)|\mathcal{C}| - 2 \\ &< \frac{\kappa}{\sqrt{3d}} \min\{|\mathcal{C}|, |\mathcal{D}|\}^{1/2} T^{3/2}, \end{aligned} \tag{3}$$

then it is cheaper to represent this mapping with a sink rather than a diagonal attention.

This result (proved in Appendix D) provides an analytic condition, in terms of context length T and number of dormant and copy-paste tokens $|\mathcal{D}|, |\mathcal{C}|$, for a sink to incur lower cost than a diagonal attention. First, we note that the LHS of (3) scales linearly in T, while the RHS scales as $T^{3/2}$. Hence, larger context lengths favor sinks, in agreement with (Barbero et al., 2025). Second, the LHS of (3) scales quadratically in $|\mathcal{D}|, |\mathcal{C}|$, while the RHS scales as $\min\{|\mathcal{C}|, |\mathcal{D}|\}^{1/2}$. Hence, dense copy-paste patterns favor diagonal attention. Finally, a large dimension d reduces the RHS of (3) while keeping the LHS of order $c_1(T + |\mathcal{C}| + |\mathcal{D}|)|\mathcal{C}|$ and thus favoring diagonal attention. This can be deduced from the fact that a large d makes the W_{QK} matrix cheaper compared to fully-connected layers (perfectly aligned vectors of length $\sqrt{d}$ have dot-product equal to d, while the softmax only divides by $\sqrt{d}$).

Figure 3 compares the total cost of a single transformer block exhibiting either diagonal or sink attention patterns. We train on the distribution described above, set $|\mathcal{D}| = |\mathcal{C}| = 5, d = T + |\mathcal{D}| + |\mathcal{C}|$ and use an additional $1/\sqrt{d}$ normalization of the softmax parameters to de-bias the results across context lengths. The plot confirms the dependence on T predicted by Theorem 7.1: on a logarithmic scale, the sink slope is close to 1, while the diagonal slope is close to $3/2$.

Trade-off for a generic input-output mapping. We next consider a generalized distribution, leading to a more nuanced trade-off between sink and diagonal attention. Let $C \geq 1$ be the number of copy-paste groups. Let $\mathcal{S}$ denote the set of d-dimensional vectors $\{b_{\text{BOS}}, \bar{d}_1, \ldots, \bar{d}_C, c_1, \ldots, c_C\}$, where b_{BOS} is the BOS token, $\bar{d}_1, \ldots, \bar{d}_C$ are the dormant group directions, and $c_1, \ldots, c_C$ the copy-paste context embeddings. Vectors in $\mathcal{S}$ have the same length (i.e., $\|b_{\text{BOS}}\|_2 = \|\bar{d}_c\|_2 = \|c_c\|_2 = \sqrt{d}$ for every c) and are near-orthogonal (i.e., $|\langle x, y\rangle| \leq \phi d$, for any two distinct vectors x, y in $\mathcal{S}$). Each dormant context in group c has the form $d_{ci} = \lambda \bar{d}_c + \eta_{ci}$, where $\eta_{ci} \perp \mathcal{S}$, $\|\eta_{ci}\|_2 = \delta\sqrt{d}$ for $\delta \in [0, 1)$ and $\lambda := \sqrt{1 - \delta^2}$. For each group c, let $\mathcal{P}_c$ be the set of unordered pairs $\{i, j\}$ s.t. both relative orders of d_{ci} and d_{cj} occur in the data. Denote $\Sigma_D := \sum_{c=1}^{C} \sum_{\{i,j\} \in \mathcal{P}_c} (d_{ci} - d_{cj})(d_{ci} - d_{cj})^\top$, $r_{\text{eff}}(\Sigma_D) := \frac{\text{tr}(\Sigma_D)}{\|\Sigma_D\|_2}$ and r_η as the rank of all noise vectors.

The output of the block on the BOS token is itself; the output on every dormant context d_{ci} is $(1 + \phi_{ci})d_{ci}$, with $\phi_{ci} \in [1 - \epsilon, 1 + \epsilon]$ and $\epsilon \in [0, 1)$ a tolerance parameter; the output on a copy-paste context c_c is c_c if there is no dormant context from group c before it in the sequence, and otherwise $c_c + \phi_c \frac{1}{|R_c|} \sum_{\tilde{c} \in R_c} \tilde{c}$, where R_c is the set of preceding dormant contexts from group c and $\phi_c \in [1 - \epsilon, 1 + \epsilon]$. Note that ϕ_c, ϕ_{ci} might depend on the particular sequence and position of the copy-paste context. We say that the block implements a sink if: *(i)* the BOS embedding attends only to itself; *(ii)* every dormant embedding attends to b_{BOS}; and *(iii)* every copy-paste embedding attends uniformly to the preceding dormant embedding from its own group, if any such embedding is present, and otherwise attends to b_{BOS}. We say that the block implements a diagonal attention if: *(i)* the BOS embedding attends only to itself; *(ii)* every dormant embedding attends to itself; and *(iii)* every copy-paste embedding attends uniformly to the preceding dormant embedding from its own group, if any such embedding is present, and otherwise attends to itself. Finally,

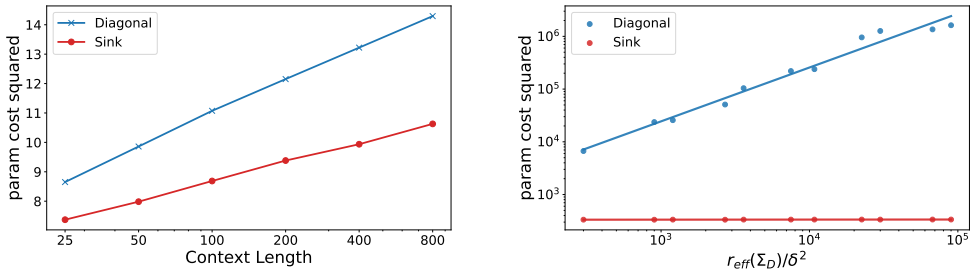

Figure 3. Frobenius squared cost of a single transformer block if the attention pattern is either sink or diagonal. **Left:** Backcopy task of Theorem 7.1 and cost as a function of the context length. Results are averaged over 3 independent seeds, and the variance of the results is negligible. **Right:** Generic task of Theorem 7.2 and cost as a function of $r_{\text{eff}}(\Sigma_D)/\delta^2$.

we assume that, for every copy-paste token c_c and every dormant key d_{ki}, the score depends only on the centroid component of the key: since $d_{ki} = \lambda \bar{d}_k + \eta_{ki}$ and $\eta_{ki} \perp \mathcal{S}$, we write this as $c_c^\top W_{QK} d_{ki} = \lambda c_c^\top W_{QK} \bar{d}_k$. Thus, the query-key mechanism only tests whether a dormant key has the correct centroid component.

In summary, there are C context collectors that check whether past contexts have the correct semantic meaning (the average direction $\bar{d}_c$), and then they collect both this meaning as well as the nuanced low-level meaning (η_{ci}). As an example, tokens that are ends of some words can try to collect the previous tokens of that word by querying the "high-level" word membership feature and then copy-pasting the nuanced content of the tokens belonging to the same word. Note that heads of this type occur in pretrained transformers, as displayed in Figure 13 in Appendix E.3.

Theorem 7.2. *Let $2C\phi \leq \frac{1}{4}, \phi \leq \frac{1}{20}, \delta \leq \frac{1}{2}$ and $\frac{1}{20} \leq \epsilon \leq \frac{1}{4}$. If the following holds:*

$$12\,\frac{\kappa C}{\sqrt{d}} + 5(r_\eta + C) < \frac{\kappa}{\delta^2\sqrt{d}}\, r_{\text{eff}}(\Sigma_D), \tag{4}$$

then it is cheaper to represent this mapping with a sink rather than a diagonal attention.

This result (proved in Appendix D) provides a sufficient condition for a sink to incur lower cost than a diagonal attention, when the input-output distribution is described via a general copy-paste mechanism. The idea is to show that the RHS of (4) lower bounds the cost of diagonal attention, and the LHS upper bounds the sink cost. We now provide an interpretation of these two bounds. *(i)* As for the diagonal lower bound, when d grows, the relative cost of W_{QK} decreases and so does the RHS of (4). The crucial term is $r_{\text{eff}}(\Sigma_D)/\delta^2$: if the nuanced signal is small compared to the high-level content, then it is expensive for the diagonal pattern to separate the nuanced part, so dormant embeddings attend to themselves instead of other same-group embeddings. Similarly, the higher the effective rank $r_{\text{eff}}(\Sigma_D)$ of the nuanced part is, the more dimensions W_{QK} has to blow up to guarantee the separation of the same-group dormant tokens. *(ii)* As for the sink upper bound, the term $\kappa C/\sqrt{d}$ in the LHS of (4) ensures that copy-paste tokens attend according to the task description. The term $r_\eta + C$ upper bounds the number of dimensions that W_{VO} and the MLP need to operate on for dormant embeddings. This is larger than the equivalent cost of a diagonal solution, since copy-paste and dormant embeddings do not share the resources to perform the same update.

To obtain conditions under which diagonal attention is cheaper than sink, one needs to upper bound the cost of diagonal attention and lower bound the sink cost. Theorem D.1 in Appendix D contains *(i)* an upper bound on diagonal attention of order $\frac{\kappa C}{\sqrt{d}} + \frac{\kappa}{\sqrt{d}\delta^2} r_\eta + r_\eta + C$ (with smaller multiplicative constant on $r_\eta + C$ compared to the sink upper bound in the LHS of (4)), and *(ii)* a lower bound on sink of order $\kappa C/\sqrt{d}$. We conjecture that a lower bound on sink of order $r_\eta + C$ holds as well.

Figure 3 compares parameter costs of a single block trained to implement the task of Theorem 7.2 with $\phi = 0, C = 3, d = 100, T = 50$, as a function of the crucial quantity $r_{\text{eff}}(\Sigma_D)/\delta^2$ computed over a sweep of δ values and number of dormant tokens per group. The plot shows that, while the sink cost is independent of $r_{\text{eff}}(\Sigma_D)/\delta^2$, the cost of the diagonal solution grows log-linearly with it.

Extension to relative positional embeddings. Although our theory is stated for absolute positional embeddings, it extends naturally to RoPE-style relative positional encodings. For a fixed context length, RoPE corresponds to a fixed block-diagonal orthogonal rotation Φ applied after the query/key projections. Let us focus for simplicity on a single-head attention layer, and let $\tilde{W}_Q$ be such that $\tilde{W}_Q \Phi = \Phi W_Q$. Then, we can absorb the rotation matrix into the embedding geometry, $Z = \Phi X$, and define the new geometry on Z instead of X. As $\tilde{W}_Q = \Phi W_Q \Phi^T$ and rotation matrices do not change Frobenius norms, this problem is equivalent to the original one. Similar considerations hold for the key matrix and for multi-head attention.

Summary. Diagonal attention switches off token communi-

cation, thereby helping to prevent oversmoothing. However, its representation is often more expensive than that of the sink, which can explain the prevalence of the latter in pretrained transformers.

8. Concluding remarks

In this paper, we study attention sinks and compare them with alternative ways to shut down token communication. We close several gaps in the literature by showing that: *(i)* in order to represent sinks, transformers rotate the embeddings against that of the sink; *(ii)* sparse attention is needed to prevent oversmoothing despite existing theoretical mechanisms that could avoid it; *(iii)* sinks are theoretically and empirically necessary to switch attention off fully; *(iv)* there are other mechanisms to prevent token communication beyond sinks, such as diagonal attention; and, finally, *(v)* sinks are prevalent in practice since they have provably lower cost than diagonal attention for a wide range of data distributions.

Several limitations motivate open questions for future work: In which precise module is the alignment between sink and other embeddings established? Can we characterize the oversmoothing progression in multi-head regimes and with alternative oversmoothing measures? How do the relative positional embeddings change the geometry of heads? How to formally characterize the usefulness of diagonal attention from the next-token prediction viewpoint?

Acknowledgements

M. M. and P. S. are funded by the European Union (ERC, INF^2, project number 101161364). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This research was supported by the Scientific Service Units (SSU) of ISTA through resources provided by Scientific Computing (SciComp).

References

Barbero, F., Banino, A., Kapturowski, S., Kumaran, D., Madeira Araújo, J., Vitvitskyi, O., Pascanu, R., and Veličković, P. Transformers need glasses! information over-squashing in language tasks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Barbero, F., Arroyo, A., Gu, X., Perivolaropoulos, C., Veličković, P., Pascanu, R., and Bronstein, M. M. Why do LLMs attend to the first token? In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=tu4dFUsW5z.

Behrens, F., Biggio, L., and Zdeborová, L. Understanding counting in small transformers: The interplay between attention and feed-forward layers. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

Bengio, Y., Frasconi, P., and Gori, M. Recurrent neural networks for adaptive temporal processing. In *Proceedings of the 6th Italian Workshop on Parallel Architectures and Neural Networks WIRN93*, pp. 85–117, 1993.

Bondarenko, Y., Nagel, M., and Blankevoort, T. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Cai, Z., Zhang, Y., Gao, B., Liu, Y., Li, Y., Liu, T., Lu, K., Xiong, W., Dong, Y., Hu, J., and Xiao, W. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. In *First Conference on Language Modeling*, 2025.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Dong, Y., Cordonnier, J.-B., and Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning (ICML)*, 2021.

Dovonon, G. J.-S., Bronstein, M. M., and Kusner, M. Setting the record straight on transformer oversmoothing. In *Transactions on Machine Learning Research (TMLR)*, 2025.

Eldan, R. and Li, Y. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. The emergence of clusters in self-attention dynamics. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Giorlandino, A. and Goldt, S. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation. *arXiv preprint arXiv:2505.24333*, 2025.

Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., and Lin, M. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations (ICLR)*, 2025.

Guo, T., Pai, D., Bai, Y., Jiao, J., Jordan, M. I., and Mei, S. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs. In *The Second Conference on Parsimony and Learning (Recent*

Spotlight Track), 2025. URL https://openreview.net/forum?id=Zx6WUbE9J7.

Hankemeier, V. and Schilling, M. Stochastic parroting in temporal attention–regulating the diagonal sink. *arXiv preprint arXiv:2602.10956*, 2026.

Husain, H., Wu, H.-H., Gazit, T., Allamanis, M., and Brockschmidt, M. CodeSearchNet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*, 2019. doi: 10.48550/arXiv.1909.09436. URL https://arxiv.org/abs/1909.09436.

Ivanitskiy, M., Behn, C. D., and Fung, S. W. Motifs in attention patterns of large language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=ND2WsCwlDQ.

Kang, S., Kim, J., Kim, J., and Hwang, S. J. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations (ICLR)*, 2025.

Karagodin, N., Polyanskiy, Y., and Rigollet, P. Clustering in causal attention masking. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Karpathy, A. char-rnn. https://github.com/karpathy/char-rnn, 2015.

Kobayashi, S., Akram, Y., and Von Oswald, J. Weight decay induces low-rank attention layers. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Li, J.-N., Tu, Q., Mao, C., Yu, Z., Wen, J.-R., and Yan, R. Streamingdialogue: Prolonged dialogue learning via long context compression with minimal losses. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Luo, J., Fan, W.-C., Wang, L., He, X., Rahman, T., Abolmaesumi, P., and Sigal, L. To sink or not to sink: Visual information pathways in large vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017a. URL https://arxiv.org/abs/1609.07843.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017b.

Noci, L., Anagnostidis, S., Biggio, L., Orvieto, A., Singh, S. P., and Lucchi, A. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., Liu, D., Zhou, J., and Lin, J. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Qiu, Z., Huang, Z., Wen, K., Jin, P., Zheng, B., Zhou, Y., Huang, H., Wang, Z., Li, X., Zhang, H., et al. A unified view of attention and residual sinks: Outlier-driven rescaling is essential for transformer training. *arXiv preprint arXiv:2601.22966*, 2026.

Queipo-de Llano, E., Arroyo, Á., Barbero, F., Dong, X., Bronstein, M., LeCun, Y., and Shwartz-Ziv, R. Attention sinks and compression valleys in llms are two sides of the same coin. *arXiv preprint arXiv:2510.06477*, 2025.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Ran-Milo, Y. Attention sinks are provably necessary in softmax transformers: Evidence from trigger-conditional tasks. *arXiv preprint arXiv:2603.11487*, 2026.

Rulli, M. E., Petruzzi, S., Michielon, E., Silvestri, F., Scardapane, S., and Devoto, A. Attention sinks in diffusion language models. *arXiv preprint arXiv:2510.15731*, 2025.

Ruscio, V., Nanni, U., Silvestri, F., et al. What are you sinking? a geometric approach on attention sink. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Saada, T. N., Naderi, A., and Tanner, J. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. *arXiv preprint arXiv:2410.07799*, 2024.

Sanford, C., Hsu, D., and Telgarsky, M. Transformers, parallel computation, and logarithmic depth. In *International Conference on Machine Learning (ICML)*, 2024.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Shi, H., GAO, J., Xu, H., Liang, X., Li, Z., Kong, L., Lee, S. M. S., and Kwok, J. Revisiting over-smoothing in

BERT from the perspective of graph. In *International Conference on Learning Representations (ICLR)*, 2022.

Tu, C., Ye, P., Zhou, D., Bai, L., Yu, G., Chen, T., and Ouyang, W. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *International Journal of Computer Vision*, 134(1):22, 2026.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Wang, P., Zheng, W., Chen, T., and Wang, Z. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In *International Conference on Learning Representations (ICLR)*, 2022.

Wang, Y., Zhang, M., Sun, J., Wang, C., Yang, M., Xue, H., Tao, J., Duan, R., and Liu, J. Mirage in the eyes: Hallucination attack on multi-modal large language models with only attention sink. In *USENIX Security Symposium*, 2025.

Wen, Y., Wu, J., Jain, A., Goldstein, T., and Panda, A. Analysis of attention in video diffusion transformers. *arXiv preprint arXiv:2504.10317*, 2025.

Wu, X., Ajorlou, A., Wang, Y., Jegelka, S., and Jadbabaie, A. On the role of attention masks and layernorm in transformers. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Yu, Z., Wang, Z., Fu, Y., Shi, H., Shaikh, K., and Lin, Y. C. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In *International Conference on Machine Learning (ICML)*, 2024.

Zhai, S. Exclusive self attention. *arXiv preprint arXiv:2603.09078*, 2026.

Zhang, S., Khan, M., and Papyan, V. Attention sinks and outlier features: A "catch, tag, and release" mechanism for embeddings. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

A. Proofs for Section 4

Proposition 4.1. *Let $Z \in \mathbb{R}^{d\times(T+1)}$ be a matrix of input embeddings $z_0, z_1, \ldots, z_T$, and let $\mathcal{J} \subset [T]$ be a set of indices. Then, there exists $W \in \mathbb{R}^{d\times d}$ such that $z_j^T W z_0 > z_j^T W z_i$ for all $j \in \mathcal{J}$ and $i \in [T]$ if and only if the two sets of vectors $\{z_0 - z_i,\, i \in [T]\}$ and $\{z_j,\, j \in \mathcal{J}\}$ lie in half-spaces.*

Proof. We first rewrite the conditions:

$$\begin{aligned} & z_j^T W z_0 > z_j^T W z_i \\ \iff \quad & z_j^T W (z_0 - z_i) > 0 \\ \iff \quad & \mathrm{tr}(z_j^T W (z_0 - z_i)) > 0 \\ \iff \quad & \mathrm{tr}(W (z_0 - z_i) z_j^T) > 0 \\ \iff \quad & \langle W, z_j (z_0 - z_i)^T \rangle_F > 0. \end{aligned}$$

Thus, we have a system of strict linear inequalities. Using Farkas' lemma, we know that the existence of the solution (W) is equivalent to the non-existence of a non-zero vector $\alpha = (\alpha_{ij})_{i,j}$ such that $\alpha_{ij} \geq 0$ and $\sum_{i\in[T], j\in\mathcal{J}} \alpha_{ij} z_j (z_0 - z_i)^T = 0$. Geometrically, if we treat the matrices $z_j(z_0 - z_i)^T$ as vectors, this is equivalent to these vectors lying in a half space. We will show that a sufficient and necessary condition for this is what is claimed in the theorem statement.

The sufficiency. Assume there exists $\alpha = (\alpha_{ij})_{i,j}$ such that $\alpha_{ij} \geq 0$ and $\sum_{i\in[T], j\in\mathcal{J}} \alpha_{ij} z_j (z_0 - z_i)^T = 0$. And assume by contradiction that both $(z_0 - z_i)$ and z_j lie in a half-space. Then there exist vectors u, v such that $u^T z_j > 0\ \forall j \in \mathcal{J}$ and $(z_0 - z_i)^T v > 0\ \forall i \in [T]$. Multiply the involved equation from left by u^T and from right by v. We get: $\sum_{i\in[T], j\in\mathcal{J}} \alpha_{ij} u^T z_j (z_0 - z_i)^T v = 0$. Because all the involved terms except α_{ij} are strictly positive, all α_{ij} must be zero, which is a contradiction.

The necessity. Assume now that the set of vectors $z_j(z_0 - z_i)^T$ does lie in a half-space and assume by contradiction that either $(z_0 - z_i)$ or z_j^T do not lie in a half-space. W.l.o.g. let that be the z_j's. Then, there exists a non-zero, non-negative vector α_j such that $\sum_j \alpha_j z_j = 0$. Hence, we can build α by simply repeating this sub-vector along the i index and readily get that $\sum_{i\in[T], j\in\mathcal{J}} \alpha_{ij} z_j (z_0 - z_i)^T = 0$ by considering the inner sum over j.

□

B. Proofs for Section 5

B.1. Existence of $W_{VO}^{(l)}$ preventing oversmoothing

We provide examples showing that, even under uniform attention, when skip connections and layer normalization are present, for every full-column-rank token matrix, there exist configurations of matrices $W_{VO}^{(l)}$ that can counteract the homogenizing effect of attention. To that end, we study the evolution as the number of layers increases of

$$X^{(l+1)} = X^{(l)} + W_{VO}^{(l)} \,\mathrm{RMS}(X^{(l)})(A^{(l)})^\top = X^{(l)} + W_{VO}^{(l)} X^{(l)} D^{(l)} (A^{(l)})^\top, \tag{5}$$

where $D^{(l)} = \mathrm{diag}(d_1^{(l)}, \ldots, d_T^{(l)})$, with $d_i^{(l)} = \frac{\sqrt{d}}{\left\|X_i^{(l)}\right\|_2}$ for all $i \in [T]$.

For simplicity, we consider causal uniform attention at every layer, i.e., for every $l \in \mathbb{N}_0$, $A^{(l)} \equiv A$, $A_{ij} = 1/i$ for $j \leq i$ and $A_{ij} = 0$ otherwise. We adopt this choice because it represents a regime in which token representations are mixed as strongly as possible. However, as will become clear from the proofs, similar results can be easily obtained for different causal attention schemes.

First, we provide examples where the direction of each token is preserved at every attention step. Consequently, if the initial token representations have distinct directions, oversmoothing is avoided.

Proposition B.1. *Let $d \geq T$ and $X^{(0)}$ a full-column-rank matrix. Let us denote $Z_i^{(l)} = \mathrm{RMS}(X_i^{(l)})$, for $i \in [T]$. For a*

fixed $\lambda > 0$, consider a matrix $W_{VO}^{(l)} = W_{VO}$, such that

$$\begin{aligned} W_{VO}Z_1^{(0)} &= \lambda Z_1^{(0)}, \\ W_{VO}Z_i^{(0)} &= \lambda(Z_i^{(0)} - Z_{i-1}^{(0)}), \qquad i \in \{2, \dots, T\}. \end{aligned}$$

Then, with the dynamics defined in (5), for every token $i \in [T]$ and every layer $l \in \mathbb{N}_0$, $Z_i^{(l)} = Z_i^{(0)}$.

Proof. Let us denote $v_i := Z_i^{(0)}$, for $i =\in [T]$. We will prove that $Z_i^{(l)} = v_i$ for all $l \in \mathbb{N}_0$ by induction on i.

(i) Base case. Let us prove $Z_1^{(l)} = v_1$ for all $l \in \mathbb{N}_0$. We have

$$X_1^{(1)} = W_{VO}Z_1^{(0)} + X_1^{(0)} = \lambda v_1 + \frac{\left\|X_1^{(0)}\right\|}{\sqrt{d}} v_1 = \left(\lambda + \frac{\left\|X_1^{(0)}\right\|}{\sqrt{d}}\right) v_1.$$

Therefore, $Z_1^{(1)} = v_1$.

For a fixed l, assume $Z_1^{(l)} = v_1$. Then,

$$X_1^{(l+1)} = W_{VO}Z_1^{(l)} + X_1^{(l)} = \lambda v_1 + \frac{\left\|X_1^{(l)}\right\|}{\sqrt{d}} v_1 = \left(\lambda + \frac{\left\|X_1^{(l)}\right\|}{\sqrt{d}}\right) v_1.$$

Therefore, $Z_1^{(l)} = v_1$ for all $l \in \mathbb{N}_0$.

(ii) Induction. Assume $Z_k^{(l)} = v_k$ for all $l \in \mathbb{N}_0$ and for all $k < i$. Then,

$$\begin{aligned} X_i^{(1)} &= \frac{1}{i}W_{VO}Z_1^{(0)} + \frac{1}{i}W_{VO}Z_2^{(0)} + \cdots + \frac{1}{i}W_{VO}Z_{i-1}^{(0)} + \frac{1}{i}W_{VO}Z_i^{(0)} + X_i^{(0)} \\ &= \frac{1}{i}\lambda v_1 + \frac{1}{i}\lambda(v_2 - v_1) + \cdots + \frac{1}{i}\lambda(v_{i-1} - v_{i-2}) + \frac{1}{i}\lambda(v_i - v_{i-1}) + \frac{\left\|X_i^{(0)}\right\|}{\sqrt{d}} v_i \\ &= \left(\frac{\lambda}{i} + \frac{\left\|X_i^{(0)}\right\|}{\sqrt{d}}\right) v_i. \end{aligned}$$

Therefore, $Z_i^{(1)} = v_i$.

For a fixed l, assume $Z_i^{(l)} = v_i$. Then, since we also have $Z_k^{(l)} = v_k$ for all $k < i$ (induction hypothesis),

$$\begin{aligned} X_i^{(l+1)} &= \frac{1}{i}W_{VO}Z_1^{(l)} + \frac{1}{i}W_{VO}Z_2^{(l)} + \cdots + \frac{1}{i}W_{VO}Z_{i-1}^{(l)} + \frac{1}{i}W_{VO}Z_i^{(l)} + X_i^{(l)} \\ &= \frac{1}{i}W_{VO}v_1 + \frac{1}{i}W_{VO}v_2 + \cdots + \frac{1}{i}W_{VO}v_{i-1} + \frac{1}{i}W_{VO}v_i + X_i^{(l)} \\ &= \frac{1}{i}\lambda v_1 + \frac{1}{i}\lambda(v_2 - v_1) + \cdots + \frac{1}{i}\lambda(v_{i-1} - v_{i-2}) + \frac{1}{i}\lambda(v_i - v_{i-1}) + \frac{\left\|X_i^{(l)}\right\|}{\sqrt{d}} v_i \\ &= \left(\frac{\lambda}{i} + \frac{\left\|X_i^{(l)}\right\|}{\sqrt{d}}\right) v_i. \end{aligned}$$

Therefore, we have proven $Z_i^{(l)} = v_i$ for all $l \in \mathbb{N}_0$. □

Remark B.2. Note that the matrix W_{VO} defined in Proposition B.1 always exists. First, let us denote $v_i := Z_i^{(0)}$, for $i \in [T]$ and let $V = \mathrm{RMS}(X^{(0)}) \in \mathbb{R}^{d \times T}$. We can extend $\{v_1, \dots, v_T\}$ to a basis of $\mathbb{R}^d$, $\mathcal{B} = \{v_1, \dots, v_T, u_{T+1}, \dots, u_d\}$. The

conditions in Proposition B.1 determine the image of the first T elements of the basis, and the image of the remaining $d-T$ basis vectors can be chosen arbitrarily. W_{VO} is a matrix such that

$$W_{VO}V = \lambda \begin{pmatrix} | & | & & | \\ v_1 & v_2 - v_1 & \dots & v_T - v_{T-1} \\ | & | & & | \end{pmatrix} = \lambda V L, \quad L = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ & & & \dots & & \\ & & & \dots & & \\ 0 & 0 & 0 & \dots & 1 & -1 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

Since V has full column rank, it has a left inverse $V^+ = (V^\top V)^{-1}V^\top$, so one possible expression for W_{VO} is

$$W_{VO} = \lambda V L (V^\top V)^{-1} V^\top.$$

Although such a configuration is possible, it is not realistic to expect a trained transformer to converge to such a configuration, since this would mean that the attention step reduces to a rescaling of each token. The following result establishes a much stronger statement: for any full-column-rank token matrix, one can choose value matrices so that each layer applies any prescribed scaled orthogonal transformation that preserves the current token span.

Proposition B.3. *Let $d \geq T$ and let $X^{(0)} \in \mathbb{R}^{d\times T}$ have full column rank. For each layer $l \in \mathbb{N}_0$, let $U^{(l)} = [U_1^{(l)}\ U_2^{(l)}] \in \mathbb{R}^{d\times d}$ be an orthogonal matrix such that the columns of $U_1^{(l)}$ span $\mathrm{col}(X^{(l)})$ and the columns of $U_2^{(l)}$ span $\mathrm{col}(X^{(l)})^\perp$. Let $Q_1^{(l)} \in \mathbb{R}^{T\times T}$ and $Q_2^{(l)} \in \mathbb{R}^{(d-T)\times(d-T)}$ be arbitrary orthogonal matrices and let $\sigma^{(l)} > 0$. Define*

$$Q^{(l)} = U^{(l)} \begin{pmatrix} Q_1^{(l)} & 0 \\ 0 & Q_2^{(l)} \end{pmatrix} (U^{(l)})^\top.$$

Then it is possible to choose matrices $W_{VO}^{(l)}$ such that, with the dynamics defined in (5),

$$X^{(l+1)} = \sigma^{(l)} Q^{(l)} X^{(l)}.$$

Consequently, $X^{(l+1)}$ has full column rank, and the pairwise cosine similarities between tokens are preserved across layers.

Proof. For any fixed matrix $B^{(l)} \in \mathbb{R}^{T\times T}$, we can set

$$W_{VO}^{(l)} = X^{(l)}(B^{(l)} - I)((A^{(l)})^\top)^{-1}(D^{(l)})^{-1}(X^{(l)})^+,$$

where $(X^{(l)})^+ = ((X^{(l)})^\top X^{(l)})^{-1}(X^{(l)})^\top$ is the Moore–Penrose left inverse of $X^{(l)}$, so that

$$X^{(l+1)} = X^{(l)} B^{(l)}.$$

Note that $W_{VO}^{(l)}$ is well defined because $A^{(l)}$ is lower triangular with nonzero diagonal and $d_i^{(l)} \neq 0$ for all i.

Let us define $B^{(l)} = \sigma^{(l)}(X^{(l)})^+ Q^{(l)} X^{(l)}$, where $Q^{(l)} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. For $\sigma^{(l)} Q^{(l)} X^{(l)} = X^{(l)} B^{(l)}$ to hold, it is enough that:

$$Q^{(l)} X^{(l)} = P^{(l)} Q^{(l)} X^{(l)},$$

where $P^{(l)} = (X^{(l)})(X^{(l)})^+$ is the projection matrix onto the column space of $X^{(l)}$. This means that rotating a column of $X^{(l)}$ by $Q^{(l)}$ must lie in the subspace spanned by the columns of $X^{(l)}$.

We can construct such a $Q^{(l)}$ by decomposing the feature space $\mathbb{R}^d$ into two orthogonal subspaces: $\mathbb{R}^d = \mathrm{col}(X^{(l)}) \oplus \mathrm{col}(X^{(l)})^\perp$, with dimensions T and $d-T$ respectively. We then define $Q^{(l)}$ as a block-diagonal orthogonal matrix with respect to this decomposition:

$$Q^{(l)} = U^{(l)} \begin{pmatrix} Q_1^{(l)} & 0 \\ 0 & Q_2^{(l)} \end{pmatrix} (U^{(l)})^\top,$$

where $Q_1^{(l)} \in \mathbb{R}^{T\times T}$ and $Q_2^{(l)} \in \mathbb{R}^{(d-T)\times(d-T)}$ are arbitrary orthogonal matrices, $U^{(l)} = [U_1^{(l)}, U_2^{(l)}]$ is orthogonal, with the columns of $U_1^{(l)}$ spanning $\mathrm{col}(X^{(l)})$ and the columns of $U_2^{(l)}$ spanning $\mathrm{col}(X^{(l)})^\perp$.

By construction, $Q^{(l)}$ preserves $\mathrm{col}(X^{(l)})$. Thus, with the above choice of $B^{(l)}$,

$$X^{(l+1)} = X^{(l)}B^{(l)} = \sigma^{(l)}Q^{(l)}X^{(l)}.$$

Since $\sigma^{(l)} > 0$ and $Q^{(l)}$ is orthogonal, the pairwise inner products of the normalized tokens remain constant across layers. □

B.2. Proof of Theorem 5.1 and discussion of its assumptions

Theorem 5.1. *Assume $\mathbb{E}[\epsilon_i\epsilon_i^\top] = \Sigma_V$ for all $i \in [T]$ and $\mathbb{E}[\epsilon_j\epsilon_i^\top] = \Sigma_C$ for all $i \neq j$. Let $B := \Sigma_V - \Sigma_C$ and $C = \bar{z}\bar{z}^\top + \Sigma_C$. Then, for all $\lambda \in [0,1]$ and $i < j$, $k \in [T]$, we have*

$$\begin{aligned}
&\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)] = \beta^2\,\mathrm{tr}(C) + 2\beta\,\mathrm{tr}(CW) + \mathrm{tr}(CW^\top W)\\
&+\frac{\lambda}{j}\left(\beta\,\mathrm{tr}(BW) + \mathrm{tr}(BW^\top W)\right),\\
&\mathbb{E}[\|Y_k(\lambda)\|^2] = \beta^2\,\mathrm{tr}(B+C) + 2\beta\,\mathrm{tr}((B+C)W)\\
&+\mathrm{tr}((B+C)W^\top W) - 2\left(1-\frac{1}{k}\right)(\beta\,\mathrm{tr}(BW)\\
&+\mathrm{tr}(BW^\top W))\lambda + \left(1-\frac{1}{k}\right)\mathrm{tr}(BW^\top W)\lambda^2.
\end{aligned} \tag{2}$$

Thus, assuming $\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)] > 0\ \forall i,j \in [T]$, we have that (i) *if $\beta\,\mathrm{tr}(BW) > 0$, $\hat{\rho}(Y(\lambda))$ is strictly increasing in λ; and* (ii) *if $\beta\,\mathrm{tr}(BW) + \mathrm{tr}(BW^\top W) < 0$, $\hat{\rho}(Y(\lambda))$ is strictly decreasing.*

Proof. First, let us compute the exact value of $\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)]$. Let $A_\lambda = (1-\lambda)I + \lambda A_u$ with coordinates $A_\lambda = (a_{ij}^\lambda)$, where $\sum_{j=1}^T a_{ij}^\lambda = 1$. Then, the expression of token i after the head update is:

$$Y_i(\lambda) = \beta Z_i + (WZA_\lambda^\top)_i = \beta Z_i + \sum_{k=1}^T a_{ik}^\lambda WZ_k. \tag{6}$$

Let us compute $\mathbb{E}[Z_k^\top MZ_l]$ for a fixed matrix $M \in \mathbb{R}^{d\times d}$ and arbitrary indices $k,l \in [T]$. Since $\mathbb{E}[\epsilon_i] = 0$,

$$\mathbb{E}[Z_k^\top MZ_l] = \bar{z}^\top M\bar{z} + \mathbb{E}[\epsilon_k^\top M\epsilon_l] = \bar{z}^\top M\bar{z} + \mathrm{tr}(\mathbb{E}[\epsilon_l\epsilon_k^\top]M).$$

Then, since $\mathbb{E}[\epsilon_l\epsilon_l^\top] = \Sigma_V$ and $\mathbb{E}[\epsilon_l\epsilon_k^\top] = \Sigma_C$ for $l \neq k$,

$$\mathbb{E}[Z_k^\top MZ_l] = \bar{z}^\top M\bar{z} + \mathrm{tr}(\Sigma_C M) + \delta_{kl}\,\mathrm{tr}(BM) = \mathrm{tr}(CM) + \delta_{kl}\,\mathrm{tr}(BM).$$

From Equation (6), we have

$$Y_i(\lambda)^\top Y_j(\lambda) = \beta^2 Z_i^\top Z_j + \beta\sum_{k=1}^T a_{jk}^\lambda Z_i^\top WZ_k + \beta\sum_{k=1}^T a_{ik}^\lambda Z_k^\top W^\top Z_j + \sum_{k=1}^T\sum_{l=1}^T a_{ik}^\lambda a_{jl}^\lambda Z_k^\top W^\top WZ_l.$$

Combining both equations, we have

$$\begin{aligned}
\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)] &= \beta^2(\mathrm{tr}(C) + \delta_{ij}\,\mathrm{tr}(B))\\
&+\beta(\sum_{k=1}^T a_{jk}^\lambda(\mathrm{tr}(CW) + \delta_{ik}\,\mathrm{tr}(BW)) + \sum_{k=1}^T a_{ik}^\lambda(\mathrm{tr}(CW^\top) + \delta_{kj}\,\mathrm{tr}(BW^\top)))\\
&+\sum_{k=1}^T\sum_{l=1}^T a_{ik}^\lambda a_{jl}^\lambda(\mathrm{tr}(CW^\top W) + \delta_{kl}\,\mathrm{tr}(BW^\top W)).
\end{aligned}$$

Since $\sum_{k=1}^{T} a_{ik}^{\lambda} = 1$, we obtain

$$\begin{aligned}\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)] &= \beta^2(\operatorname{tr}(C) + \delta_{ij}\operatorname{tr}(B)) + \beta(\operatorname{tr}(CW) + a_{ji}^{\lambda}\operatorname{tr}(BW) + \operatorname{tr}(CW^\top) + a_{ij}^{\lambda}\operatorname{tr}(BW^\top)) \\ &\quad + \operatorname{tr}(CW^\top W) + \sum_{k=1}^{T} a_{ik}^{\lambda} a_{jk}^{\lambda}\operatorname{tr}(BW^\top W) \\ &= \beta^2(\operatorname{tr}(C) + \delta_{ij}\operatorname{tr}(B)) + \beta(2\operatorname{tr}(CW) + (a_{ji}^{\lambda} + a_{ij}^{\lambda})\operatorname{tr}(BW)) \\ &\quad + \operatorname{tr}(CW^\top W) + \sum_{k=1}^{T} a_{ik}^{\lambda} a_{jk}^{\lambda}\operatorname{tr}(BW^\top W),\end{aligned}$$

where the last equality holds because C and B are symmetric. Note that

$$a_{ij}^{\lambda} = \begin{cases} 1 - \lambda + \frac{\lambda}{i} & \text{if } j = i \\ \frac{\lambda}{i} & \text{if } j < i \\ 0 & \text{if } j > i \end{cases},$$

so

$$\sum_{k=1}^{T} a_{ik}^{\lambda} a_{jk}^{\lambda} = \begin{cases} \frac{\lambda}{j} & \text{if } i < j \\ 1 - \lambda(2-\lambda)(1 - \frac{1}{i}) & \text{if } i = j \end{cases}.$$

We now distinguish the two cases and conclude the first part of the proof.

(i) <u>Case $i < j$:</u>

$$\begin{aligned}\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)] &= \beta^2\operatorname{tr}(C) + \beta(2\operatorname{tr}(CW) + \frac{\lambda}{j}\operatorname{tr}(BW)) + \operatorname{tr}(CW^\top W) + \frac{\lambda}{j}\operatorname{tr}(BW^\top W) \\ &= \beta^2\operatorname{tr}(C) + 2\beta\operatorname{tr}(CW) + \operatorname{tr}(CW^\top W) + \frac{\lambda}{j}\left(\beta\operatorname{tr}(BW) + \operatorname{tr}(BW^\top W)\right).\end{aligned}$$

(ii) <u>Case $i = j$:</u>

$$\begin{aligned}\mathbb{E}[\|Y_i(\lambda)\|^2] &= \beta^2\operatorname{tr}(B+C) + 2\beta\operatorname{tr}(CW) + 2\beta\left(1 - \lambda + \frac{\lambda}{i}\right)\operatorname{tr}(BW) \\ &\quad + \operatorname{tr}(CW^\top W) + \left(1 - \lambda(2-\lambda)\left(1 - \frac{1}{i}\right)\right)\operatorname{tr}(BW^\top W) \\ &= \beta^2\operatorname{tr}(B+C) + 2\beta\operatorname{tr}((B+C)W) + \operatorname{tr}((B+C)W^\top W) \\ &\quad - 2\left(1 - \frac{1}{i}\right)(\beta\operatorname{tr}(BW) + \operatorname{tr}(BW^\top W))\lambda + \left(1 - \frac{1}{i}\right)\operatorname{tr}(BW^\top W)\lambda^2.\end{aligned}$$

We now analyze the monotonicity of $\hat{\rho}(Y(\lambda))$ and provide sufficient conditions for the numerator and denominator to be strictly increasing or decreasing. Let $D_k(\lambda) := \mathbb{E}[\|Y_k(\lambda)\|^2]$. Then,

$$D_k'(\lambda) = -2\left(1 - \frac{1}{k}\right)(\beta\operatorname{tr}(BW) + (1-\lambda)\operatorname{tr}(BW^\top W)).$$

First, note that for $i \neq j$, $\mathbb{E}[(\epsilon_i - \epsilon_j)(\epsilon_i - \epsilon_j)^\top] = 2(\Sigma_V - \Sigma_C)$, so $B = \Sigma_V - \Sigma_C$ is positive semidefinite, which implies $\operatorname{tr}(BW^\top W) \geq 0$. Therefore, D_k' increases with λ. Then, $D_k'(\lambda) < 0\ \forall\lambda \in [0,1]$ is equivalent to $D_k'(1) < 0$ and $D_k'(\lambda) > 0\ \forall\lambda \in [0,1]$ is equivalent to $D_k'(0) > 0$. Since $D_k'(1) = -2\beta\left(1 - \frac{1}{k}\right)\operatorname{tr}(BW)$ and $D_k'(0) = -2\left(1 - \frac{1}{k}\right)(\beta\operatorname{tr}(BW) + \operatorname{tr}(BW^\top W))$, $D_k(\lambda)$ strictly decreases iff $\beta\operatorname{tr}(BW) > 0$ and strictly increases iff $\beta\operatorname{tr}(BW) + \operatorname{tr}(BW^\top W) < 0$. We distinguish two cases:

(i) $\beta\operatorname{tr}(BW) > 0$. Then, $\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)]$ strictly increases with λ and $D_k(\lambda)$ strictly decreases. Since both numerator and denominator are assumed to be positive, we conclude that $\hat{\rho}(Y(\lambda))$ strictly increases with λ.

(ii) $\beta\operatorname{tr}(BW) + \operatorname{tr}(BW^\top W) < 0$. Then, $\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)]$ strictly decreases with λ and $D_k(\lambda)$ strictly increases. Since both numerator and denominator are assumed to be positive, we conclude that $\hat{\rho}(Y(\lambda))$ strictly decreases with λ. □

Remark B.4 (Interpretation of the conditions of Theorem 5.1). First, note that $B = \frac{1}{2}\mathbb{E}[(Z_i - Z_j)(Z_i - Z_j)^\top]$, so B captures the covariance structure of token-to-token differences. Therefore,

$$\operatorname{tr}(BW) = \frac{1}{2}\mathbb{E}[(Z_i - Z_j)^\top W(Z_i - Z_j)].$$

Then, $\operatorname{tr}(BW) > 0$ means that, on average over token pairs, the value-output map W maps discrepancy directions between tokens to directions positively aligned with the original differences. Similarly,

$$\operatorname{tr}(BW^\top W) = \frac{1}{2}\mathbb{E}[\|W(Z_i - Z_j)\|^2],$$

so this term measures how much token-to-token variability is retained by the attention branch after applying W.

Assuming $\beta > 0$, the first condition, $\beta \operatorname{tr}(BW) > 0$ means that the attention branch is positively aligned with the residual token differences. Intuitively, in this regime, sparse attention preserves or reinforces token-specific components. Making attention more uniform averages out these components, which increases cross-token alignment and thus leads to an increase of $\hat{\rho}(Y(\lambda))$. By contrast, $\beta \operatorname{tr}(BW) < 0$ means that the attention branch is anti-aligned with the residual stream, which can potentially counteract the mixing effect of the attention.

The second condition, $\beta \operatorname{tr}(BW) + \operatorname{tr}(BW^\top W) < 0$ corresponds to the case where this anti-alignment is strong enough to dominate the nonnegative term $\operatorname{tr}(BW^\top W)$. In this regime, sparse attention cancels token differences rather than preserving them. Making attention more uniform weakens this cancellation, so $\hat{\rho}(Y(\lambda))$ decreases.

The intermediate case $\beta \operatorname{tr}(BW) < 0$, but $\beta \operatorname{tr}(BW) + \operatorname{tr}(BW^\top W) > 0$ has a mixed interpretation. The attention branch is anti-aligned with the residual token differences, but this anti-alignment is not strong enough to dominate the nonnegative term $\operatorname{tr}(BW^\top W)$. In this case, $\mathbb{E}[Y_i(\lambda)^\top Y_j(\lambda)]$ still increases with λ, since its slope is proportional to $\beta \operatorname{tr}(BW) + \operatorname{tr}(BW^\top W)$. However, by analyzing $D'_k(\lambda)$, we see that the token norms are not monotone: they initially decrease and then increase, attaining their minimum at

$$\lambda_* = \frac{\beta \operatorname{tr}(BW) + \operatorname{tr}(BW^\top W)}{\operatorname{tr}(BW^\top W)} \in (0, 1).$$

Thus, in this regime, $\hat{\rho}(Y(\lambda))$ is increasing in λ for $\lambda \in [0, \lambda_*)$, but for $\lambda \in [\lambda_*, 1]$ there is a competition between two effects: increasing uniformity increases both cross-token alignment and the token norms. Consequently, Theorem 5.1 does not imply a general monotonicity statement for $\hat{\rho}(Y(\lambda))$ in this intermediate regime.

C. Proofs for Section 6

Both Proposition 6.1 and Corollary C.3 directly follow from the more general result below.

Theorem C.1. *Consider any head of a globally optimal solution of the regularized objective. Assume that $W_{VO}x_0 = 0$. If this head is implementing the attention sink mechanism, then it must implement the hard attention switch mechanism. Moreover, let a_i be a collection of all input embeddings across all training data points with the positions ranging from 1 to $T - 1$. Let b_i be the collection of the outputs of the attention layer of all these embeddings. If we further assume rank$(\{a_i\})$ = rank$(\{b_i\})$ and that, for all input embedding sequences to the head, either the sequence is full column rank or the sequence lies in a strict half-space, then if the head is implementing the hard attention switch mechanism, it must implement it solely via the attention sink, i.e., all context embeddings a_i for which $b_i = 0$ attend fully to the BOS token.*

Proof. We start with the first implication. If the head is implementing the attention sink mechanism, there must exist at least one context in one sample that attends solely to the BOS token. This means that the query and key matrices are non-zero. From this, it follows that some context in some sequence has to have non-zero output of the attention layer. Otherwise, we could set all the matrices in the attention layer to 0 and achieve the same mapping with better regularization loss. This implies the head is implementing the hard attention switch (considering this non-zero mapping as the witness of the condition in the definition of the hard attention map).

For the second implication, we just need to realize that b_i can be written as a convex combination of some value embeddings of some a_js, i.e., $b_i = \sum_j \alpha_j W_{VO} a_j$. This means that $b_i \in$ Span$(\{W_{VO}a_k\})$. As this holds for every b_i, we get Span$(\{b_i\}) \subset$ Span$(\{W_{VO}a_i\})$. As rank$(\{a_i\})$ = rank$(\{b_i\})$, W_{VO} has full rank on Span$(\{a_i\})$, otherwise the rank inequality could not be satisfied. This means that the kernel of W_{VO} on Span$(\{a_i\})$ is empty except for 0. Now assume by

contradiction that some a_i for which $b_i = 0$ does not fully attend to sink but rather uses a non-trivial combination of the previous value embeddings, i.e., $0 = b_i = \sum_j \alpha_j W_{VO} a_j = W_{VO} \sum_j \alpha_j a_j$. If the embeddings a_j lie in a strict half-space and neither of the $W_{VO} a_j$ is 0 (as guaranteed by the kernel computation), then so must $W_{VO} a_j$ embeddings. As we are taking a convex combination, it is impossible to mix the embeddings up to 0, which is a contradiction. Furthermore, if the embeddings a_i are full rank, then $0 \neq \sum_j \alpha_j a_j \in \text{Span}(\{a_i\})$ and, as such, cannot belong to the kernel of W_{VO}, which is again a contradiction. Therefore, such a_i can only attend to sink with an attention weight of roughly 1. □

Remark C.2. Let us discuss the expected minimal size of the training dataset to allow for the crux condition of Proposition 6.1 (for the simplified setting in which the copy-paste tokens copy-paste immediate past) and Corollary C.3 (each independent dimension being followed by a copy-paste token at least once or each token being copy-pasted at least once, respectively) to be satisfied with high probability. If the dimensions are roughly equally likely to be sampled, then this is an example of the coupon collector problem: denoting the dimension of the span of the input space as R, the expected size of the dataset is roughly $R \log R$ if there is no bias concerning the position at which any of the independent dimensions occur. Even if certain dimensions are only sampled at certain positions within the sequence, the bound would worsen only mildly to $R(\log R + \log T)$, where T is the context length, using the union bound over context position in the standard coupon collector bound.

Corollary C.3. *Consider the input-output distribution of Proposition 6.1 assuming there is a fixed set of dormant tokens $\mathcal{D}$ and a fixed set of copy-paste tokens $\mathcal{C}$. Let tokens be sampled from a first-order Markov chain in which BOS is only followed by dormant tokens and any token can follow any other token with non-zero probability. Let inputs be computed as a sum of token embedding and positional embedding and every input sequence be full rank. Then, if every token gets copy-pasted at least once (except the BOS), all dormant tokens have to attend fully to the BOS token.*

Remark C.4. Heads implementing the behavior of Corollary C.3 as well as the one from Proposition 6.1 are ubiquitous in pretrained transformers. Empirically, $16\%, 13\%$ and 21% of the heads of gpt2, Llama 3.1-8B and Gemma 7B, pooled across 50 sequences per each of 4 language datasets (`TinyStories`, `tinyshakespeare`, `WikiText`, and `CodeSearchNet-Python`), have at least 60% of the attention mass on the sink and the lower-1 diagonal, plus at least 10% on each of the two. It is also well known that such heads are the essential first step for implementing the induction head, a useful tool for in-context learning (Sanford et al., 2024). The heads that show the generalized copy-paste pattern in which the copy-paste tokens copy also from more distant past are even more prevalent.

D. Proofs for Section 7

Theorem 7.1. *Let a single-head transformer block represent exactly the input-output distribution above. Let c_1 be such that $c_1 \geq d/(T + |\mathcal{C}| + |\mathcal{D}| + 2)$. We have that, if*

$$\begin{aligned} &\frac{2\kappa}{\sqrt{d}}\left(\sqrt{2|\mathcal{D}|} + T - 1\right) + 3T + 3|\mathcal{D}| + |\mathcal{C}| \\ &+ 4c_1(T + |\mathcal{C}| + |\mathcal{D}| + 2)|\mathcal{C}| - 2 \\ &< \frac{\kappa}{\sqrt{3d}} \min\{|\mathcal{C}|, |\mathcal{D}|\}^{1/2} T^{3/2}, \end{aligned} \tag{3}$$

then it is cheaper to represent this mapping with a sink rather than a diagonal attention.

Proof. **Upper-bound on sink representation cost.** We can construct the parameters as follows:

$$W_K^T W_Q = \frac{2\kappa}{\sqrt{d}}\left((h_0 + p_0)\sum_{k\in\mathcal{D}} h_k^T + \sum_{t=1}^{T-1} p_t p_{t+1}^T\right), \tag{7}$$

$$W_O W_V = I - h_0 h_0^T - p_0 p_0^T, \tag{8}$$

$$W_1 = \sum_{i\in\mathcal{D}} h_i h_i^T + \sum_{t=1}^{T} p_t p_t^T - 2\mathbf{1}\sum_{i\in\mathcal{C}} h_i^T, \tag{9}$$

$$W_2 = \sum_{i\in\mathcal{D}} h_i h_i^T + \sum_{t=1}^{T} p_t p_t^T, \tag{10}$$

where we recall that W_1, W_2 are the MLP matrices. First we argue that this representation solves the task, using the sink as the attention pattern. Since every input embedding has norm $\sqrt{d}$, RMSNorm acts as the identity before the attention layer.

Note that the BOS token fits the label, as it belongs to the null space of all the weight matrices and does not receive attention from any other tokens. Next, take any dormant token $x_{tk} = \sqrt{d/2}(h_k + p_t)$ with $k \in \mathcal{D}$. The attention logit with the BOS, including the standard division by $\sqrt{d}$, is

$$\frac{1}{\sqrt{d}} x_{00}^T W_K^T W_Q x_{tk} = 2\kappa,$$

while the attention logit with the preceding token, if $t > 1$, is only

$$\frac{1}{\sqrt{d}} x_{t-1,i}^T W_K^T W_Q x_{tk} = \kappa.$$

The attention logit with all other tokens is 0. Therefore, under the hard-attention approximation, the BOS receives the full attention mass. Since $W_O W_V$ annihilates the BOS embedding, the dormant token does not receive any update in the attention layer. Then, since the post-attention residual stream is still x_{tk} and hence still has norm $\sqrt{d}$, the second RMSNorm also acts as the identity. The MLP acts as an identity on this vector, adding exactly x_{tk} to the residual stream and thus exactly representing the task.

Finally, take any copy-paste token $x_{tj} = \sqrt{d/2}(h_j + p_t)$ with $j \in \mathcal{C}$. Its token embedding has zero attention contribution with the sink term, as it is in the kernel of the corresponding query-key matrix. Therefore, the only non-zero attention logit is the one with the preceding token, thanks to the positional embeddings:

$$\frac{1}{\sqrt{d}} x_{t-1,i}^T W_K^T W_Q x_{tj} = \kappa.$$

Thus, the copy-paste token precisely copies the content of the previous token, since the value and output matrices act as an identity on everything but the BOS. After this attention update, the second RMSNorm may rescale the residual stream by a positive scalar, but this does not affect the sign pattern used by the ReLU argument. Finally, the MLP layer makes sure that any vector containing the h_j token embedding of a copy-paste token does not survive the rectification by the ReLU owing to the strong negative all-ones bias. This is, in fact, one of the most expensive components in this representation.

Next, we compute the cost of such representation. Because both $W_K^T W_Q$ and $W_O W_V$ are products of two matrices, they will contribute with the nuclear norm of the product, owing to the variational form of nuclear norm. The contribution of $W_K^T W_Q$ is therefore

$$\frac{2\kappa}{\sqrt{d}} \left(\sqrt{2|\mathcal{D}|} + T - 1 \right).$$

The contribution of $W_O W_V$ is simply $T + |\mathcal{D}| + |\mathcal{C}| - 2$. The contribution of W_1 is

$$4d|\mathcal{C}| + T + |\mathcal{D}|.$$

Finally, the contribution of W_2 is $|\mathcal{D}| + T$. Using $d \leq c_1(T + |\mathcal{C}| + |\mathcal{D}| + 2)$ and adding these together, we get that the cost of the sink is upper bounded by

$$\frac{2\kappa}{\sqrt{d}} \left(\sqrt{2|\mathcal{D}|} + T - 1 \right) + 3T - 2 + 3|\mathcal{D}| + |\mathcal{C}| + 4c_1(T + |\mathcal{C}| + |\mathcal{D}| + 2)|\mathcal{C}|.$$

Lower-bound on the representation cost of diagonal attention. To represent the diagonal attention and the copy-paste attention pattern, the following two series of inequalities must be satisfied, now with the standard $1/\sqrt{d}$ attention scaling:

$$\begin{aligned}
&\frac{1}{\sqrt{d}} x_{t-1,i}^T W_K^T W_Q x_{tj} \geq \frac{1}{\sqrt{d}} x_{sk}^T W_K^T W_Q x_{tj} + \kappa \\
&\forall (i,k) \in \mathcal{V}^2, \quad \forall j \in \mathcal{C}, \quad \forall t \in \{2, \ldots, T\}, \quad \forall s \in [t-2], \\
&\frac{1}{\sqrt{d}} x_{t-1,i}^T W_K^T W_Q x_{tj} \geq \frac{1}{\sqrt{d}} x_{tj}^T W_K^T W_Q x_{tj} + \kappa \\
&\forall i \in \mathcal{V}, \quad \forall j \in \mathcal{C}, \quad \forall t \in \{2, \ldots, T\}, \\
&\frac{1}{\sqrt{d}} x_{ti}^T W_K^T W_Q x_{ti} \geq \frac{1}{\sqrt{d}} x_{sj}^T W_K^T W_Q x_{ti} + \kappa \\
&\forall i \in \mathcal{D}, \quad \forall j \in \mathcal{V}, \quad \forall t \in \{1, \ldots, T\}, \quad \forall s \in [t-1],
\end{aligned}$$

with $\mathcal{V} = \mathcal{C} \cup \mathcal{D}$. Let d_0, a be arbitrary but fixed token indices of a dormant and a copy-paste token, respectively, and let t be an arbitrary but fixed time index. Choosing $i \equiv a, j \equiv a$ in the second inequality and $i \equiv d_0, j \equiv d_0, s \equiv t-1$ in the third inequality, and multiplying both inequalities by $\sqrt{d}$, we get the following simplified couple:

$$x_{t-1,a}^T W_K^T W_Q x_{ta} \geq x_{ta}^T W_K^T W_Q x_{ta} + \kappa\sqrt{d},$$
$$x_{td_0}^T W_K^T W_Q x_{td_0} \geq x_{t-1,d_0}^T W_K^T W_Q x_{td_0} + \kappa\sqrt{d}.$$

Adding them up and using

$$x_{ti} = \sqrt{\frac{d}{2}}(h_i + p_t),$$

we get

$$(p_t - p_{t-1})^T W_K^T W_Q (h_{d_0} - h_a) \geq \frac{4\kappa}{\sqrt{d}}.$$

Because d_0, a, t are arbitrary, this must hold for all indices from their respective sets. Thus, $p_t^T W_K^T W_Q (h_{d_0} - h_a)$ is increasing in t. From this, it follows that $p_t^T W_K^T W_Q (h_{d_0} - h_a)$ must be either positive for $t \in \{T/2, \dots, T\}$ or negative for $t \in \{1, \dots, T/2\}$. We omit caring about the integerness of $T/2$ as we can without loss of generality choose the longer half, losing only a negligible part of the lower bound. Without loss of generality, let it be the second half and re-index the time so that we have the following implication:

$$p_i^T W_K^T W_Q (h_{d_0} - h_a) \geq \frac{4\kappa}{\sqrt{d}} i \qquad \forall i \in [T/2].$$

However, this must hold for all possible choices of $a \in \mathcal{C}$ and $d_0 \in \mathcal{D}$. Take $\min\{|\mathcal{C}|, |\mathcal{D}|\}$ distinct pairs of embeddings and stack them, divided by $\sqrt{2}$ to make them unit-norm, into a matrix H. Let $U\Sigma V^T$ be the full SVD of $W_K^T W_Q$. Consider a matrix P whose rows are $\{p_i\}_{i=1}^{T/2}$. Then, we have

$$PU\Sigma V^T H \geq \frac{4\kappa}{\sqrt{2d}}(1, \dots, T/2)^T \mathbf{1}^T.$$

We can rewrite this as

$$\hat{P}\Sigma\hat{H} \geq \frac{2\sqrt{2}\kappa}{\sqrt{d}}(1, \dots, T/2)^T \mathbf{1}^T,$$

where $\hat{P}, \hat{H}$ are again orthogonal matrices. Then we have

$$\|\Sigma\|_*^2 \geq \|\Sigma\|_F^2 \geq \left\|\hat{P}\Sigma\hat{H}\right\|_F^2 \geq \frac{1}{3d}\min\{|\mathcal{C}|, |\mathcal{D}|\}\kappa^2 T^3,$$

and thus

$$\left\|W_K^T W_Q\right\|_* = \|\Sigma\|_* \geq \frac{\kappa}{\sqrt{3d}}\min\{|\mathcal{C}|, |\mathcal{D}|\}^{1/2} T^{3/2}.$$

Finally, we lower-bound both the cost of the value and output matrices as well as the cost of the MLP block by zero. Therefore, if

$$\frac{2\kappa}{\sqrt{d}}\left(\sqrt{2|\mathcal{D}|} + T - 1\right) + 3T + 3|\mathcal{D}| + |\mathcal{C}| + 4c_1(T + |\mathcal{C}| + |\mathcal{D}| + 2)|\mathcal{C}| - 2 < \frac{\kappa}{\sqrt{3d}}\min\{|\mathcal{C}|, |\mathcal{D}|\}^{1/2} T^{3/2},$$

then the sink representation is cheaper than every diagonal-attention representation. □

Theorem 7.2. *Let $2C\phi \leq \frac{1}{4}, \phi \leq \frac{1}{20}, \delta \leq \frac{1}{2}$ and $\frac{1}{20} \leq \epsilon \leq \frac{1}{4}$. If the following holds:*

$$12\frac{\kappa C}{\sqrt{d}} + 5(r_\eta + C) < \frac{\kappa}{\delta^2\sqrt{d}} r_{\text{eff}}(\Sigma_D), \tag{4}$$

then it is cheaper to represent this mapping with a sink rather than a diagonal attention.

Before we start the proof, we recall some definitions and introduce additional quantities. Define the normalized vectors

$$u_0 := \frac{b_{\mathrm{BOS}}}{\sqrt{d}}, \qquad u_c := \frac{\bar{d}_c}{\sqrt{d}}, \qquad u_{C+c} := \frac{c_c}{\sqrt{d}}.$$

Let $m := 1 + 2C$, $U \in \mathbb{R}^{d\times m}$ have columns $u_0, u_1, \ldots, u_{2C}$ and $G := U^\top U$. Since $G_{ii} = 1$ and $|G_{ij}| \le \phi$ for $i \ne j$, Gershgorin's circle theorem gives

$$\lambda_{\min}(G) \ge 1 - 2C\phi.$$

Assume $2C\phi < 1$, and define

$$\gamma := \frac{1}{1-2C\phi}, \qquad \beta := \sqrt{\frac{1+2C\phi}{1-2C\phi}}, \qquad \Pi := (U^\top U)^{-1}U^\top.$$

Then, $\Pi u_j = e_j$ for all j, $\|\Pi\|_2^2 \le \gamma$, and $\|U\|_2 \le \sqrt{1+2C\phi}$. Recall $r_\eta = \dim \operatorname{span}\{\eta_{ci} : \text{all } c, i\}$. For each group c, recall that $\mathcal{P}_c$ is the set of unordered pairs $\{i, j\}$ such that both relative orders of d_{ci} and d_{cj} occur in the data. Recall also

$$\Sigma_D := \sum_{c=1}^{C} \sum_{\{i,j\}\in\mathcal{P}_c} (d_{ci} - d_{cj})(d_{ci} - d_{cj})^\top, \qquad r_{\mathrm{eff}}(\Sigma_D) := \frac{\mathrm{tr}(\Sigma_D)}{\|\Sigma_D\|_2}.$$

Define

$$s_\lambda := \max\{\lambda(C+1), \sqrt{C+\lambda^2}\}, \qquad \tilde{s}_\lambda := \sqrt{1+\lambda^2(C+1)^2}.$$

and

$$\Delta_{\mathrm{diag}} := \min\{\|d_{ci} - d_{cj}\|_2^2 : \ c \in [C],\ i \ne j,\ d_{cj} \text{ may appear before } d_{ci} \text{ in the data}\},$$

with the convention that $\Delta_{\mathrm{diag}} = +\infty$ if no such pair exists.

We assume

$$2C\phi \le \frac{1}{4}, \qquad \phi \le \frac{1}{20}, \qquad \delta \le \frac{1}{2}, \qquad \frac{1}{20} \le \epsilon \le \frac{1}{4}.$$

Since $\lambda = \sqrt{1-\delta^2}$, this implies $\lambda \ge \sqrt{3}/2$. Moreover,

$$\gamma \le \frac{4}{3}, \qquad \beta \le \frac{4}{3}.$$

Now, Theorem 7.2 is implied by the following result.

Theorem D.1. *Under the setup above, the following four bounds hold.*

1. Sink upper bound. *There exists a sink solution with total cost at most*

$$\boxed{U_{\mathrm{sink}} \le 12\,\frac{\kappa C}{\sqrt{d}} + 5(r_\eta + C).}$$

2. Diagonal lower bound. *Every diagonal solution has total cost at least*

$$\boxed{L_{\mathrm{diag}} \ge \frac{\kappa}{\delta^2\sqrt{d}}\, r_{\mathrm{eff}}(\Sigma_D) + \frac{\kappa C}{2\sqrt{d}}.}$$

3. Diagonal upper bound. *There exists a diagonal solution with total cost at most*

$$\boxed{U_{\mathrm{diag}} \le 13\,\frac{\kappa C}{\sqrt{d}} + \frac{4\kappa\sqrt{d}}{\Delta_{\mathrm{diag}}}\, r_\eta + 3(r_\eta + C).}$$

4. Sink lower bound. *Every sink solution has total cost at least*

$$\boxed{L_{\mathrm{sink}} \ge \frac{\kappa C}{2\sqrt{d}}.}$$

Proof. We prove the four bounds one by one.

1. Sink upper bound. We construct an explicit sink solution.

Step 1: query-key construction. Because every nuisance vector η_{ci} is orthogonal to $\mathcal{S}$, we have $U^\top \eta_{ci} = 0$ and therefore $\Pi\eta_{ci} = 0$. Hence

$$\Pi d_{ci} = \Pi(\lambda \bar{d}_c + \eta_{ci}) = \lambda\sqrt{d}\, e_c.$$

Similarly,

$$\Pi b_{\mathrm{BOS}} = \sqrt{d}\, e_0, \qquad \Pi c_c = \sqrt{d}\, e_{C+c}.$$

Let $r := (0, 1, \ldots, 1)^\top \in \mathbb{R}^m$. Define

$$M := a\, r e_0^\top + b \sum_{c=1}^{C} e_{C+c} e_c^\top, \qquad W_{QK} := \Pi^\top M \Pi.$$

We choose

$$a := \frac{\kappa}{\lambda\sqrt{d}}, \qquad b := \frac{\kappa(1+\lambda)}{\lambda^2\sqrt{d}}.$$

Let us check the attention scores. For a dormant query d_{ci}, the BOS score is

$$s(d_{ci}, b_{\mathrm{BOS}}) = \frac{d_{ci}^\top W_{QK} b_{\mathrm{BOS}}}{\sqrt{d}} = \frac{(\lambda\sqrt{d}\, e_c)^\top M(\sqrt{d}\, e_0)}{\sqrt{d}} = \lambda a \sqrt{d} = \kappa.$$

Every non-BOS score for a dormant query is 0, so dormant tokens attend to BOS with margin at least κ.

For a copy-paste query c_c, the BOS score is

$$s(c_c, b_{\mathrm{BOS}}) = \frac{c_c^\top W_{QK} b_{\mathrm{BOS}}}{\sqrt{d}} = a\sqrt{d} = \frac{\kappa}{\lambda} \geq \kappa.$$

The score of a dormant key from the correct group is

$$s(c_c, d_{ci}) = \frac{c_c^\top W_{QK} d_{ci}}{\sqrt{d}} = \lambda b \sqrt{d} = \kappa\left(1 + \frac{1}{\lambda}\right).$$

Thus, if the correct group is present, its score is larger than that of BOS by exactly κ, and it is larger than that of wrong groups by at least κ. If the correct group is absent, the score of BOS is larger than that of all wrong groups by at least κ. Since Πd_{ci} depends only on c, all correct-group dormant keys receive the same score, and the copy-paste attention is uniform over them.

Let us now bound the cost. Since the first term in M is rank one and the second term has singular value b with multiplicity C,

$$\|M\|_* \leq a\sqrt{2C} + bC = \frac{\kappa}{\sqrt{d}}\left(\frac{\sqrt{2C}}{\lambda} + \frac{C(1+\lambda)}{\lambda^2}\right).$$

Therefore

$$\|W_{QK}\|_* \leq \|\Pi\|_2^2 \|M\|_* \leq \gamma\, \frac{\kappa}{\sqrt{d}}\left(\frac{\sqrt{2C}}{\lambda} + \frac{C(1+\lambda)}{\lambda^2}\right).$$

The variational form of nuclear norm then gives

$$\|W_Q\|_F^2 + \|W_K\|_F^2 \leq 2\|W_{QK}\|_*.$$

Using $\gamma \leq 4/3$ and $\lambda \geq \sqrt{3}/2$, we get the simplified estimate

$$\|W_Q\|_F^2 + \|W_K\|_F^2 \leq 12\, \frac{\kappa C}{\sqrt{d}}.$$

Step 2: the base dormant-content map. Let P_η be the orthogonal projector onto the nuisance span $\mathrm{span}\{\eta_{ci}\}$. Define

$$D_{\mathrm{cent}} := \mathrm{diag}(0, \underbrace{1, \dots, 1}_{C}, \underbrace{0, \dots, 0}_{C}), \qquad Q_{\mathrm{cent}} := U D_{\mathrm{cent}} \Pi.$$

Thus Q_{cent} acts as the identity on each centroid $\bar{d}_c$, and kills b_{BOS}, every copy-paste vector c_c, and every nuisance vector η_{ci}. Define

$$T_{\mathrm{base}} := P_\eta + Q_{\mathrm{cent}}.$$

Then

$$T_{\mathrm{base}} d_{ci} = d_{ci}, \qquad T_{\mathrm{base}} b_{\mathrm{BOS}} = 0, \qquad T_{\mathrm{base}} c_c = 0.$$

Moreover,

$$\|T_{\mathrm{base}}\|_* \le \|P_\eta\|_* + \|Q_{\mathrm{cent}}\|_*.$$

Here $\|P_\eta\|_* = r_\eta$, while

$$\|Q_{\mathrm{cent}}\|_* \le \|U\|_2 \, \|D_{\mathrm{cent}}\|_* \, \|\Pi\|_2 \le C\beta.$$

Therefore

$$\|T_{\mathrm{base}}\|_* \le r_\eta + C\beta.$$

Step 3: scaled value-output and MLP maps. Let

$$W_O W_V := W_{VO} := \frac{1}{2} T_{\mathrm{base}}.$$

The variational form of nuclear norm then gives

$$\|W_V\|_F^2 + \|W_O\|_F^2 \le 2\|W_{VO}\|_* = \|T_{\mathrm{base}}\|_*.$$

Let the MLP implement the linear map

$$T_{\mathrm{MLP}} := (1+\epsilon) T_{\mathrm{base}}.$$

Since biases are unregularized, a linear map can be realized on the bounded set of relevant normalized post-attention states by placing all ReLU preactivations in one positive region and using an output bias to subtract the constant offset. Hence

$$\|W_1\|_F^2 + \|W_2\|_F^2 \le 2(1+\epsilon)\|T_{\mathrm{base}}\|_*.$$

The combined value-output and MLP cost is therefore at most

$$(3+2\epsilon)\|T_{\mathrm{base}}\|_*.$$

Since $\epsilon \le 1/4$ and $\beta \le 4/3$,

$$(3+2\epsilon)(r_\eta + C\beta) \le 5(r_\eta + C).$$

Step 4: verifying the relaxed labels. Dormant tokens attend to BOS, and the BOS value is 0. Hence the post-attention state of a dormant token is d_{ci}. The MLP contributes $(1+\epsilon)d_{ci}$. Therefore the final dormant output is

$$d_{ci} + (1+\epsilon)d_{ci} = (2+\epsilon)d_{ci},$$

which is allowed because it corresponds to $\phi_{ci} = 1+\epsilon$.

If a copy-paste token has no correct-group predecessor, it attends to BOS and remains c_c. Since $T_{\mathrm{base}} c_c = 0$, the MLP contributes 0, so the final output is exactly c_c.

Suppose a copy-paste token has correct-group mean μ. Its attention contribution is $\frac{1}{2}\mu$, so the post-attention state is

$$c_c + \frac{1}{2}\mu.$$

The MLP sees

$$\mathrm{RMS}\left(c_c + \frac{1}{2}\mu\right) = \frac{c_c + \frac{1}{2}\mu}{R_\mu}, \qquad R_\mu := \frac{\|c_c + \frac{1}{2}\mu\|_2}{\sqrt{d}}.$$

Since $T_{\mathrm{base}} c_c = 0$ and $T_{\mathrm{base}}\mu = \mu$, the MLP contributes

$$(1+\epsilon)\frac{1}{2R_\mu}\mu.$$

Therefore the final coefficient of μ is

$$\theta_\mu = \frac{1}{2}\left(1 + \frac{1+\epsilon}{R_\mu}\right).$$

We now bound R_μ. Since μ is an average of dormant tokens of norm $\sqrt{d}$, we have $\|\mu\|_2 \le \sqrt{d}$. Furthermore, we have

$$\langle c_c, \mu\rangle = \lambda\langle c_c, \bar{d}_c\rangle,$$

because the nuisance component of μ is orthogonal to $\mathcal{S}$. Hence $|\langle c_c, \mu\rangle| \le \lambda\phi d \le \phi d$. Therefore

$$R_\mu^2 = 1 + \frac{\|\mu\|_2^2}{4d} + \frac{\langle c_c, \mu\rangle}{d} \le 1 + \frac{1}{4} + \phi \le \frac{13}{10}.$$

Thus $R_\mu \le \sqrt{13/10} < 7/6$.

For the lower bound,

$$\|\mu\|_2^2 \ge \lambda^2 d,$$

and so

$$R_\mu^2 \ge 1 + \frac{\lambda^2}{4} - \lambda\phi \ge 1,$$

using $\lambda \ge \sqrt{3}/2$ and $\phi \le 1/20$. Hence $1 \le R_\mu \le 7/6$.

Therefore

$$\theta_\mu \le \frac{1}{2}(1 + 1 + \epsilon) \le 1 + \epsilon,$$

and

$$\theta_\mu \ge \frac{1}{2}\left(1 + \frac{6}{7}(1+\epsilon)\right) = \frac{13}{14} + \frac{3\epsilon}{7} \ge 1 - \epsilon,$$

where the last inequality uses $\epsilon \ge 1/20$. Thus the relaxed copy-paste label is satisfied.

Combining the query-key cost with the value-output and MLP cost proves

$$U_{\mathrm{sink}} \le 12\,\frac{\kappa C}{\sqrt{d}} + 5(r_\eta + C).$$

2. Diagonal lower bound. This lower bound only uses attention-map constraints, so the relaxed value labels do not affect it.

Step 1: dormant self-attention cost. Fix a group c and a pair $\{i, j\} \in \mathcal{P}_c$. Since the diagonal pattern must work for both relative orders, we have

$$\frac{d_{ci}^\top W_{QK} d_{ci} - d_{ci}^\top W_{QK} d_{cj}}{\sqrt{d}} \ge \kappa$$

and

$$\frac{d_{cj}^\top W_{QK} d_{cj} - d_{cj}^\top W_{QK} d_{ci}}{\sqrt{d}} \ge \kappa.$$

Adding the two inequalities gives

$$(d_{ci} - d_{cj})^\top W_{QK}(d_{ci} - d_{cj}) \ge 2\kappa\sqrt{d}.$$

Summing over all admissible pairs yields

$$\langle W_{QK}, \Sigma_D \rangle \geq 2\kappa\sqrt{d}\, N_D,$$

where $N_D := \sum_c |\mathcal{P}_c|$. Hence,

$$\|W_{QK}\|_* \geq \frac{2\kappa\sqrt{d}\, N_D}{\|\Sigma_D\|_2}.$$

Since $d_{ci} - d_{cj} = \eta_{ci} - \eta_{cj}$ and $\|\eta_{ci}\|_2 = \delta\sqrt{d}$, we have

$$\mathrm{tr}(\Sigma_D) \leq 4\delta^2 d\, N_D.$$

Thus

$$\|W_{QK}\|_* \geq \frac{\kappa}{2\delta^2\sqrt{d}}\, r_{\mathrm{eff}}(\Sigma_D).$$

Step 2: copy-paste cost. In the diagonal pattern, copy-paste tokens attend to themselves when the correct group is absent. Hence for every c,

$$\frac{c_c^\top W_{QK}(\lambda \bar{d}_c - c_c)}{\sqrt{d}} \geq \kappa,$$

and for every $c' \neq c$,

$$\frac{c_c^\top W_{QK}(\lambda \bar{d}_c - \lambda \bar{d}_{c'})}{\sqrt{d}} \geq \kappa.$$

Summing all these inequalities gives

$$\langle W_{QK}, B_{\mathrm{copy}} \rangle \geq C^2\kappa\sqrt{d},$$

with

$$B_{\mathrm{copy}} = d\, U N_\lambda U^\top, \quad N_\lambda = \sum_{c=1}^{C} e_{C+c}\big(\lambda((C+1)e_c - s) - e_{C+c}\big)^\top, \quad s := \sum_{k=1}^{C} e_k.$$

Note that

$$\|N_\lambda\|_2 = \tilde{s}_\lambda,$$

implying that

$$\|B_{\mathrm{copy}}\|_2 \leq d(1 + 2C\phi)\tilde{s}_\lambda.$$

Therefore

$$\|W_{QK}\|_* \geq \frac{C^2\kappa}{\sqrt{d}\,(1 + 2C\phi)\tilde{s}_\lambda}.$$

Note that B_{copy} and Σ_D act on orthogonal subspaces, so the costs in the two steps add:

$$\|W_{QK}\|_* \geq \frac{\kappa}{2\delta^2\sqrt{d}}\, r_{\mathrm{eff}}(\Sigma_D) + \frac{C^2\kappa}{\sqrt{d}\,(1 + 2C\phi)\tilde{s}_\lambda}.$$

The variational form of nuclear norm then gives

$$\|W_Q\|_F^2 + \|W_K\|_F^2 \geq \frac{\kappa}{\delta^2\sqrt{d}}\, r_{\mathrm{eff}}(\Sigma_D) + \frac{2C^2\kappa}{\sqrt{d}\,(1 + 2C\phi)\tilde{s}_\lambda}.$$

Under the assumptions, $1 + 2C\phi \leq 5/4$ and $\tilde{s}_\lambda \leq C + 2$, implying that

$$\frac{2C^2}{(1 + 2C\phi)\tilde{s}_\lambda} \geq \frac{2C^2}{(5/4)(C+2)} \geq \frac{C}{2}.$$

Thus, we have the desired lower bound:

$$L_{\mathrm{diag}} \geq \frac{\kappa}{\delta^2\sqrt{d}}\, r_{\mathrm{eff}}(\Sigma_D) + \frac{\kappa C}{2\sqrt{d}}.$$

3. Diagonal upper bound. We construct an explicit diagonal solution.

Let P_η be the orthogonal projector onto the nuisance span. Define

$$W_{QK}^{\text{diag}} = \Pi^\top M_{\text{diag}} \Pi + \alpha_{\text{diag}} P_\eta,$$

where

$$M_{\text{diag}} = s \sum_{c=1}^{C} e_c e_c^\top + a \sum_{c=1}^{C} e_{C+c} e_{C+c}^\top + b \sum_{c=1}^{C} e_{C+c} e_c^\top,$$

with

$$s := \frac{\kappa}{\lambda^2 \sqrt{d}}, \qquad a := \frac{\kappa}{\sqrt{d}}, \qquad b := \frac{2\kappa}{\lambda\sqrt{d}}, \qquad \alpha_{\text{diag}} := \frac{2\kappa\sqrt{d}}{\Delta_{\text{diag}}}.$$

We now verify the attention scores induced by this construction. First consider a dormant query $d_{ci} = \lambda \bar{d}_c + \eta_{ci}$. Since $\Pi d_{ci} = \lambda\sqrt{d}\, e_c$ and $\Pi \eta_{ci} = 0$, the special-span part $\Pi^\top M_{\text{diag}} \Pi$ gives

$$\frac{d_{ci}^\top \Pi^\top M_{\text{diag}} \Pi d_{c'j}}{\sqrt{d}} = \begin{cases} \lambda^2 s \sqrt{d} = \kappa, & c' = c, \\ 0, & c' \neq c. \end{cases}$$

It also gives score 0 to BOS and to all copy-paste keys. Thus, before adding the nuisance projector, a dormant token gives score κ to every dormant key from its own group and score 0 to all dormant keys from other groups, BOS, and copy-paste keys. This already separates its group from all other groups, but it does not yet distinguish the token from the other dormant tokens in the same group.

The nuisance term $\alpha_{\text{diag}} P_\eta$ provides this within-group diagonal separation. Indeed, for another dormant key $d_{c'j}$, the additional score is

$$\frac{\alpha_{\text{diag}}\, \eta_{ci}^\top \eta_{c'j}}{\sqrt{d}}.$$

Therefore the nuisance contribution to the score gap between the self key d_{ci} and another dormant key $d_{c'j}$ is

$$\frac{\alpha_{\text{diag}}}{\sqrt{d}} \left(\|\eta_{ci}\|_2^2 - \eta_{ci}^\top \eta_{c'j} \right) = \frac{\alpha_{\text{diag}}}{2\sqrt{d}} \|\eta_{ci} - \eta_{c'j}\|_2^2,$$

where we used that all nuisance vectors have the same norm. In particular, for $c' = c$ and $j \neq i$, the special-span scores are tied at κ, and the nuisance term breaks the tie in favor of the self key by at least

$$\frac{\alpha_{\text{diag}} \Delta_{\text{diag}}}{2\sqrt{d}} = \kappa.$$

For $c' \neq c$, the special-span part already gives the self key an additional margin κ over the wrong group, and the nuisance term only increases the self-versus-other gap. Thus every dormant query attends to itself with margin at least κ.

Next consider a copy-paste query c_c. Since $\Pi c_c = \sqrt{d}\, e_{C+c}$, the special-span part gives

$$s(c_c, c_c) = a\sqrt{d} = \kappa,$$

and for a dormant key $d_{c'i}$,

$$s(c_c, d_{c'i}) = \begin{cases} \lambda b \sqrt{d} = 2\kappa, & c' = c, \\ 0, & c' \neq c. \end{cases}$$

It gives score 0 to BOS and to copy-paste keys other than c_c. The nuisance projector does not affect copy-paste queries, because copy-paste vectors lie in the special span and are orthogonal to the nuisance span. Therefore, if a correct-group dormant key is present, it receives score 2κ, while the self key receives score κ and all wrong-group dormant keys receive score 0. Hence the correct group wins by margin at least κ. If no correct-group dormant key is present, the self key receives score κ and all remaining admissible keys receive score 0, so the copy-paste token attends to itself by margin κ.

The nuclear norm is bounded by

$$\|W_{QK}^{\text{diag}}\|_* \leq \gamma \|M_{\text{diag}}\|_* + \alpha_{\text{diag}} \|P_\eta\|_*.$$

Now

$$\|M_{\text{diag}}\|_* \le Cs + Ca + Cb = \frac{\kappa C}{\sqrt{d}}\left(1+\frac{1}{\lambda}\right)^2,$$

and $\|P_\eta\|_* = r_\eta$. Hence

$$\|W_{QK}^{\text{diag}}\|_* \le \gamma\,\frac{\kappa C}{\sqrt{d}}\left(1+\frac{1}{\lambda}\right)^2 + \frac{2\kappa\sqrt{d}}{\Delta_{\text{diag}}} r_\eta.$$

The variational form of nuclear norm then gives

$$\|W_Q\|_F^2 + \|W_K\|_F^2 \le 2\gamma\,\frac{\kappa C}{\sqrt{d}}\left(1+\frac{1}{\lambda}\right)^2 + \frac{4\kappa\sqrt{d}}{\Delta_{\text{diag}}} r_\eta.$$

Using $\gamma \le 4/3$ and $\lambda \ge \sqrt{3}/2$ gives

$$2\gamma\left(1+\frac{1}{\lambda}\right)^2 \le 13.$$

Thus

$$\|W_Q\|_F^2 + \|W_K\|_F^2 \le 13\,\frac{\kappa C}{\sqrt{d}} + \frac{4\kappa\sqrt{d}}{\Delta_{\text{diag}}} r_\eta.$$

For the value-output map, use

$$W_{VO}^{\text{diag}} := (1-\epsilon)T_{\text{base}}.$$

This makes dormant self-attention contribute $(1-\epsilon)d_{ci}$, so the final dormant coefficient is $2-\epsilon$, which is allowed. It also makes copy-paste tokens receive $(1-\epsilon)\mu_c$ when the correct group is present, which is allowed, and gives zero value to copy-paste self-attention when no correct group is present.

The value-output cost is at most

$$2(1-\epsilon)\|T_{\text{base}}\|_* \le 2(r_\eta + C\beta) \le 3(r_\eta + C).$$

Set the MLP to zero. Therefore

$$U_{\text{diag}} \le 13\,\frac{\kappa C}{\sqrt{d}} + \frac{4\kappa\sqrt{d}}{\Delta_{\text{diag}}} r_\eta + 3(r_\eta + C).$$

4. Sink lower bound. This lower bound only uses copy-routing constraints in the sink pattern. The copy-paste query c_c must make the correct group have score larger than BOS by κ:

$$\frac{c_c^\top W_{QK}(\lambda \bar{d}_c - b_{\text{BOS}})}{\sqrt{d}} \ge \kappa,$$

and it must make the correct group have score larger than any wrong dormant group by κ:

$$\frac{c_c^\top W_{QK}(\lambda \bar{d}_c - \lambda \bar{d}_{c'})}{\sqrt{d}} \ge \kappa \qquad (c' \ne c).$$

For a fixed c, summing the first inequality and all $C-1$ wrong-group inequalities gives

$$\frac{c_c^\top W_{QK}\left(\lambda(C+1)\bar{d}_c - \lambda\sum_{k=1}^{C}\bar{d}_k - b_{\text{BOS}}\right)}{\sqrt{d}} \ge C\kappa.$$

Summing this inequality over all $c = 1, \dots, C$ gives

$$\frac{\langle W_{QK}, B_{\text{sink}}\rangle}{\sqrt{d}} \ge C^2\kappa,$$

where

$$B_{\text{sink}} := \sum_{c=1}^{C} c_c\left(\lambda(C+1)\bar{d}_c - \lambda\sum_{k=1}^{C}\bar{d}_k - b_{\text{BOS}}\right)^\top.$$

We now upper-bound the operator norm of B_{sink}. Let $s := \sum_{k=1}^{C} e_k$. Using the matrix U of normalized special vectors, we can write

$$B_{\text{sink}} = d\, U N_{\text{sink}} U^{\top},$$

where

$$N_{\text{sink}} := \sum_{c=1}^{C} e_{C+c} \left(\lambda((C+1)e_c - s) - e_0\right)^{\top}.$$

Therefore,

$$\|B_{\text{sink}}\|_2 \le d\, \|U\|_2^2\, \|N_{\text{sink}}\|_2 \le d(1+2C\phi)\|N_{\text{sink}}\|_2.$$

It remains to compute $\|N_{\text{sink}}\|_2$. The nonzero rows of N_{sink} are

$$r_c := \lambda((C+1)e_c - s) - e_0, \qquad c = 1, \dots, C.$$

A direct calculation gives

$$\langle r_c, r_c \rangle = 1 + \lambda^2(C^2 + C - 1),$$

and, for $c \neq c'$,

$$\langle r_c, r_{c'} \rangle = 1 - \lambda^2(C+2).$$

Thus $N_{\text{sink}} N_{\text{sink}}^{\top}$ is a $C \times C$ matrix with constant diagonal and constant off-diagonal entries. Its eigenvalue on the span of $\mathbf{1}_C$ is

$$C + \lambda^2,$$

and its eigenvalue on the orthogonal complement of $\mathbf{1}_C$ is

$$\lambda^2 (C+1)^2.$$

Hence

$$\|N_{\text{sink}}\|_2 = s_\lambda := \max\{\lambda(C+1), \sqrt{C+\lambda^2}\}.$$

Therefore,

$$\|B_{\text{sink}}\|_2 \le d(1+2C\phi) s_\lambda.$$

By nuclear/operator norm duality,

$$C^2 \kappa \sqrt{d} \le \langle W_{QK}, B_{\text{sink}} \rangle \le \|W_{QK}\|_* \|B_{\text{sink}}\|_2.$$

Using the operator-norm bound above, we obtain

$$\|W_{QK}\|_* \ge \frac{C^2 \kappa}{\sqrt{d}\,(1+2C\phi) s_\lambda}.$$

The variational form of the nuclear norm, equivalently the balanced factorization bound for $W_{QK} = W_Q^{\top} W_K$, gives

$$\|W_Q\|_F^2 + \|W_K\|_F^2 \ge 2\|W_{QK}\|_*.$$

Therefore

$$L_{\text{sink}} \ge \frac{2C^2 \kappa}{\sqrt{d}\,(1+2C\phi) s_\lambda}.$$

Finally, under the small-geometry assumptions, $s_\lambda \le C+1$ and $1 + 2C\phi \le 5/4$. Hence

$$\frac{2C^2}{(1+2C\phi) s_\lambda} \ge \frac{2C^2}{(5/4)(C+1)} \ge \frac{C}{2}.$$

Therefore

$$L_{\text{sink}} \ge \frac{\kappa C}{2\sqrt{d}},$$

which completes the proof. □

E. Further experimental evidence

E.1. Oversmoothing

Experimental details. We consider the same experimental setting as in Section 5: several trained transformer models (LLaMA3-8B, Gemma-7B, GPT2-XL, and Mistral-7B) and different context lengths. For each model, context length and dataset, we average results across all heads and over a batch of 50 sequences. To compute the theoretical quantities, we estimate the parameters of the token distribution for each context length from a batch of 50 sequences. For each attention head, W corresponds to the composition of the learnable normalization scaling, the value matrix, and the output projection matrix corresponding to the head.

Consider a layer l with n_l heads. We denote by $\mathrm{ATTN}_l^h(Z^{(l)})$ the attention update corresponding to head h, and by $\mathrm{ATTN}_l(Z^{(l)})$ the full attention update of layer l, obtained by concatenating the contributions of all heads.

In our experiments, we use two slightly different quantities depending on the goal. When comparing the empirical cosine similarity with our theoretical approximation for a single head h, we use the token matrix $X + \mathrm{ATTN}_l^h(Z^{(l)})$, since this is precisely the quantity captured by the theory.

However, when empirically assessing the effect of attention on oversmoothing after the skip connection, a single-head update must be rescaled to be comparable to the full layer update; otherwise, its contribution would be too small relative to the residual stream. For this reason, we use

$$X + s_l\,\mathrm{ATTN}_l^h(Z^{(l)})\ \text{, where } s_l = \frac{\frac{1}{T}\sum_{i=1}^{T}\left\|\mathrm{ATTN}_l(Z_i^{(l)})\right\|}{\frac{1}{n_l}\sum_{h=1}^{n_l}\frac{1}{T}\sum_{i=1}^{T}\left\|\mathrm{ATTN}_l^h(Z_i^{(l)})\right\|}. \tag{11}$$

Thus, s_l rescales each head update by the ratio between the average norm of the full layer update and the average norm of a single-head update in that layer.

In several experiments, we also need a measure of how dense an attention matrix is. To this end, we define the *uniformity coefficient* of a matrix $A \in \mathbb{R}^{T\times T}$ as

$$u(A) = \frac{1}{T}(1 + \sum_{i=2}^{T}(-\sum_{j=1}^{i} A_{ij}\log A_{ij})/\log i) \in [0,1], \tag{12}$$

that is, the average normalized row-wise entropy of A. Values of $u(A)$ close to 0 correspond to sparse attention, while values close to 1 indicate attention closer to uniform.

Token representations tend to exhibit positive cosine similarity across layers. Our theoretical assumptions on the token representation distribution are motivated by the empirical observation that token representations tend to have positive cosine similarity across layers. Figure 4 provides evidence for this behavior. For each layer, it reports the average cosine similarity of the token matrix both before entering the layer and after adding the corresponding attention update to the residual stream. In all cases, the average cosine similarity remains positive across layers.

Robustness across models and datasets. Here, we verify that the conclusions of Section 5 hold consistently across both models and datasets. To this end, we repeat the same experiments on LLaMA3-8B, Gemma-7B, GPT2-XL, and Mistral-7B using data from different domains, namely Wikipedia text (WikiText-103 (Merity et al., 2017b)) and code (CodeParrot[1]). First, Figure 5 shows that the approximation provided in Theorem 5.1 remains close to the empirical average cosine similarity across all models and datasets. Second, Figure 6 shows that we observe the same qualitative behavior described in Section 5 across different models and datasets: *(i)* attention slightly increases cosine similarity on average, and this increase becomes substantially larger when the average is restricted to heads with dense attention patterns; *(ii)* if each head's attention matrix is replaced by causal uniform attention, then the increase in cosine similarity is significantly bigger. GPT2-XL is the only model where the gap with uniform attention is not that big. Note that this is the only model where normalization changes the cosine similarity. We leave for further work to analyze how changing the normalization method affects our analysis.

[1] https://huggingface.co/datasets/codeparrot/codeparrot-clean-train

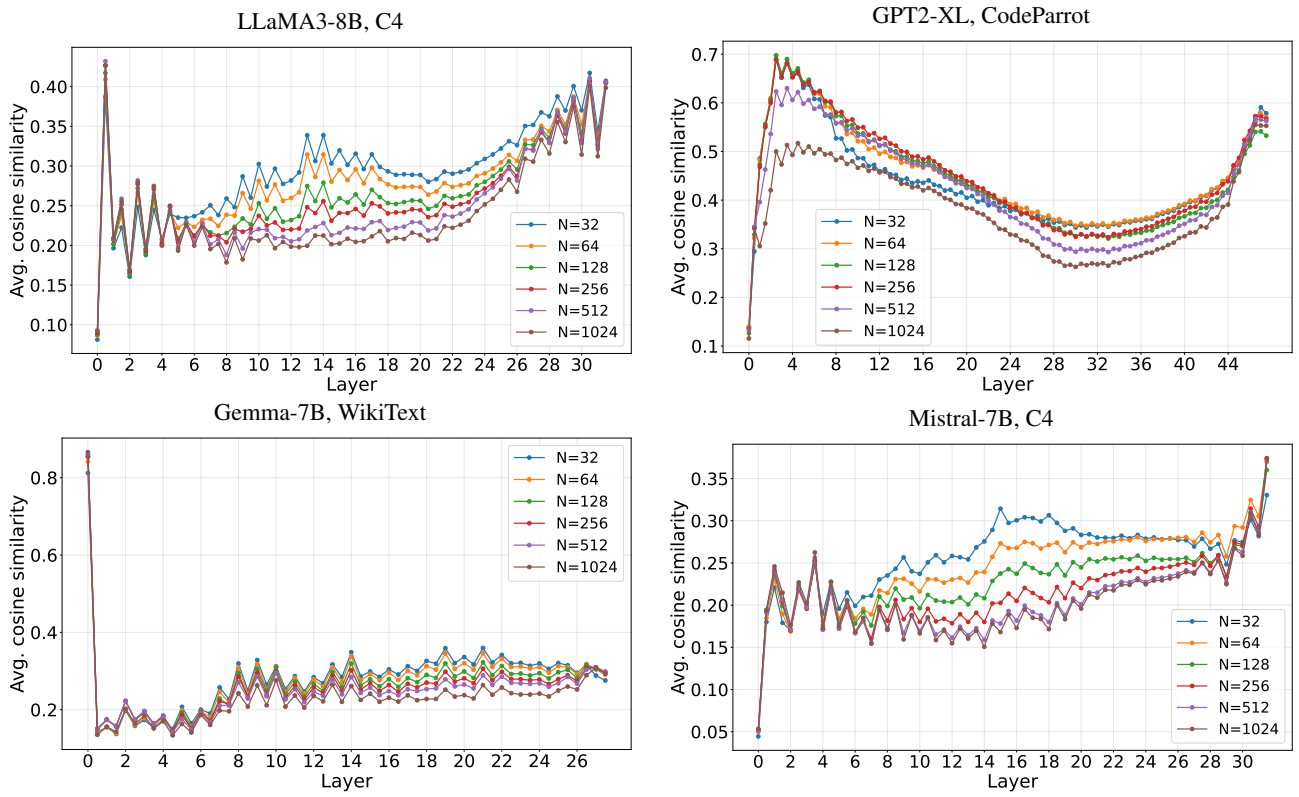

Figure 4. Evolution of average cosine similarity across layers for different models, datasets, and context lengths. For each layer, the first value corresponds to the token matrix before the layer, and the second to the token matrix after adding the attention update to the residual stream. In all cases, average cosine similarity remains positive across layers.

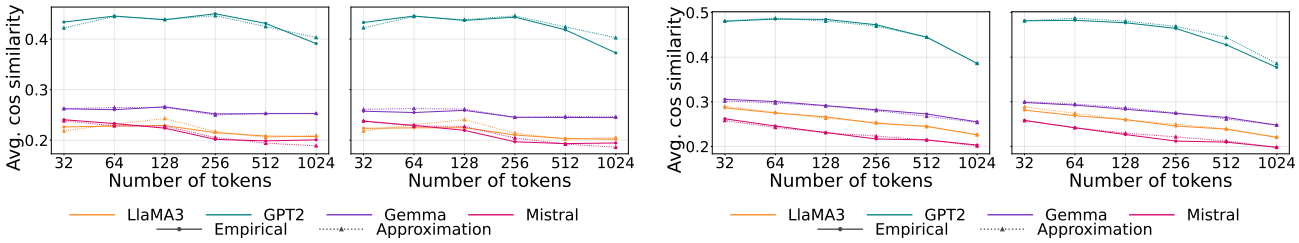

Figure 5. Empirical average cosine similarity (solid) and theoretical approximation (2) (dashed) for X (first plot) and $X + WZ$ (second plot), across models and context lengths, on the CodeParrot dataset (left) and WikiText (right).

Approximation is accurate for individual heads. We verify that the approximation provided by (1) remains close to the empirical average cosine similarity not only after averaging across heads, but also for most individual heads. Figure 7 shows, for different models and datasets and a context length of 512, the absolute difference between the empirical value and its theoretical approximation for each head. Overall, this difference stays close to 0 for many heads and never exceeds 0.1, indicating that the approximation is accurate for most heads.

How often do the conditions of Theorem 5.1 hold? Here, we examine to what extent Theorem 5.1 explains the oversmoothing behavior observed in trained transformers. The theorem provides sufficient conditions under which the approximation of the average cosine similarity after adding a head output to the residual stream, $\hat{\rho}(Y(\lambda))$, is strictly increasing or decreasing in λ. We study how often these conditions are satisfied in practice, and whether they are consistent with the token mixing observed empirically.

To this end, we consider the changes in average cosine similarity between different attention steps: (i) $\text{avg. cossim}(X + WZA^T) - \text{avg. cossim}(X + WZ)$ (Figure 8); and (ii) $\text{avg. cossim}(X + WZA_u^\top) - \text{avg. cossim}(X + WZA^\top)$ (Figure 9).

To isolate the effect of sparsity, in our theoretical model, the attention matrix is parametrized as $(1-\lambda)I + \lambda A_u$ where A_u is the $T \times T$ uniform causal attention matrix and $\lambda \in [0, 1]$. Since a general attention matrix need not lie on this interpolation path, it is not immediate that the theorem should explain the empirical behavior. Nevertheless, one may heuristically interpret $\text{avg. cossim}(X + WZA^T) - \text{avg. cossim}(X + WZ)$ as the change in cosine similarity between $\lambda = 0$ and some $\lambda_A \in [0, 1]$, and $\text{avg. cossim}(X + WZA_u^\top) - \text{avg. cossim}(X + WZA^\top)$ as the difference between $\lambda = \lambda_A \in [0, 1]$ and $\lambda = 1$. In this interpretation, attention matrices with larger uniformity scores should correspond to larger values of λ_A. Since sink attention patterns are far from this interpolation, we exclude them from this analysis.

Under this interpretation, Theorem 5.1 predicts the following behavior:

- Both quantities should be positive when $\beta \operatorname{tr}(BW) > 0$ ($\hat{\rho}(Y(\lambda))$ strictly increases with λ).
- Both quantities should be negative when $\beta \operatorname{tr}(BW) + \operatorname{tr}(BW^\top W) < 0$ ($\hat{\rho}(Y(\lambda))$ strictly decreases with λ).

The figures show that this prediction is consistent with the behavior of most heads in trained transformers. In particular, most heads satisfying $\beta \operatorname{tr}(BW) > 0$ (red) exhibit an increase in cosine similarity after applying attention, and this increase remains smaller than under uniform attention. Likewise, for most heads satisfying $\beta \operatorname{tr}(BW) + \operatorname{tr}(BW^\top W) < 0$ (blue), uniform attention induces less token mixing than the head's actual attention matrix, which in turn induces less mixing than no attention. For heads that satisfy neither condition (gray), both behaviors are observed, as expected since the theorem makes no prediction in that case. The main exception is Figure 9 for GPT2-XL on CodeParrot, where many red heads exhibit less token mixing under uniform attention than under the learned attention matrix; however, these heads are already highly uniform, with uniformity coefficient above 0.8.

Finally, heads satisfying the second condition are rare in all models except GPT2-XL, and even there they remain a minority. This helps explain the dominant empirical trend: in trained transformers, denser attention patterns typically lead to more token mixing.

A closer look into uniformity and oversmoothing. Figure 10 shows the relative change in cosine similarity between token representations before each head update (X) and after it ($X + WZA^\top$), with heads ordered by uniformity coefficient. Overall, we observe a positive correlation between the increase in cosine similarity and the uniformity of the attention heads.

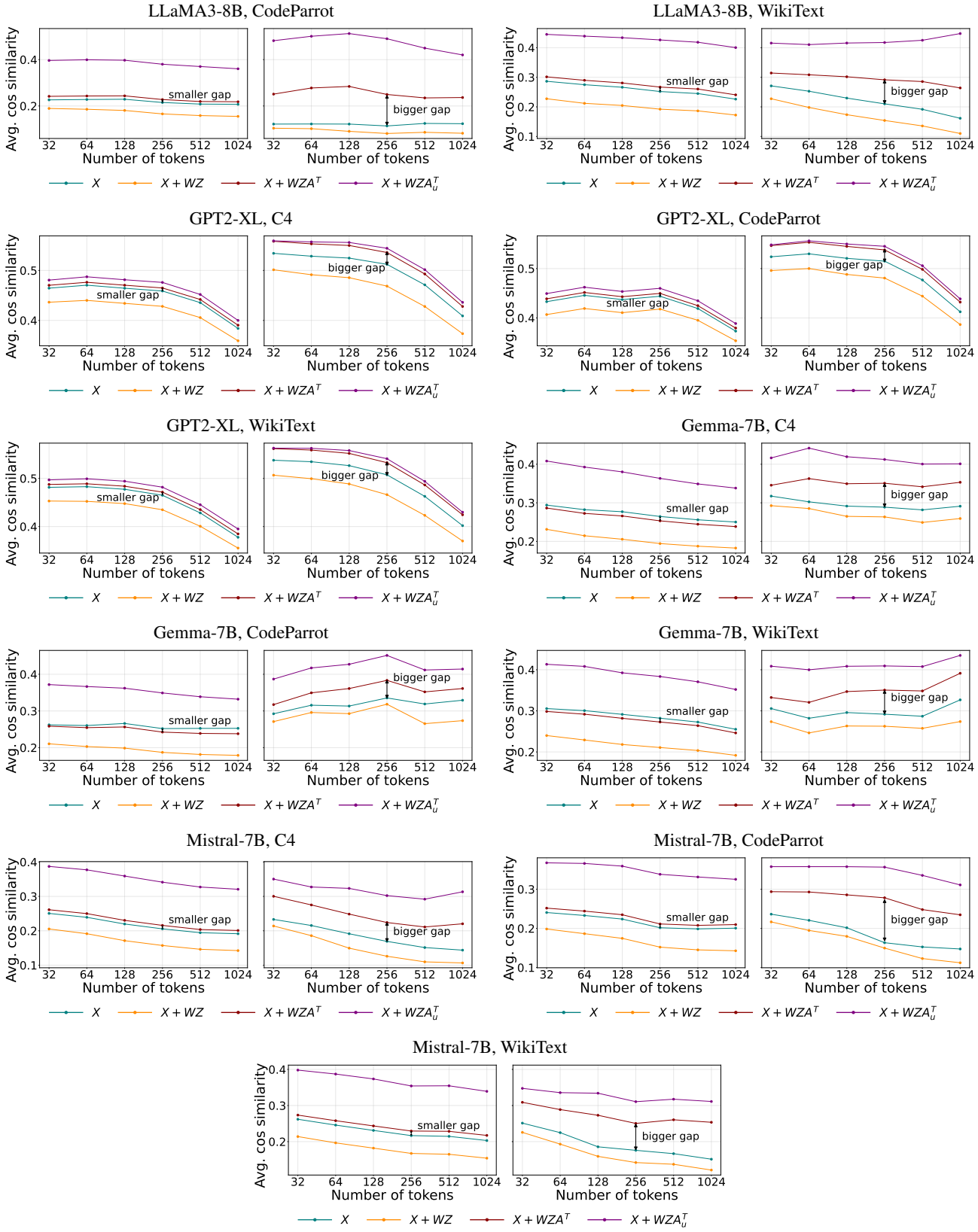

Figure 6. Empirical average cosine similarity for different models, datasets and context lengths for different components of the attention step. Results averaged over all heads (first plot) and over heads with uniformity coefficient larger than 0.6 (second plot). Cosine similarity generally increases as the attention matrix becomes denser.

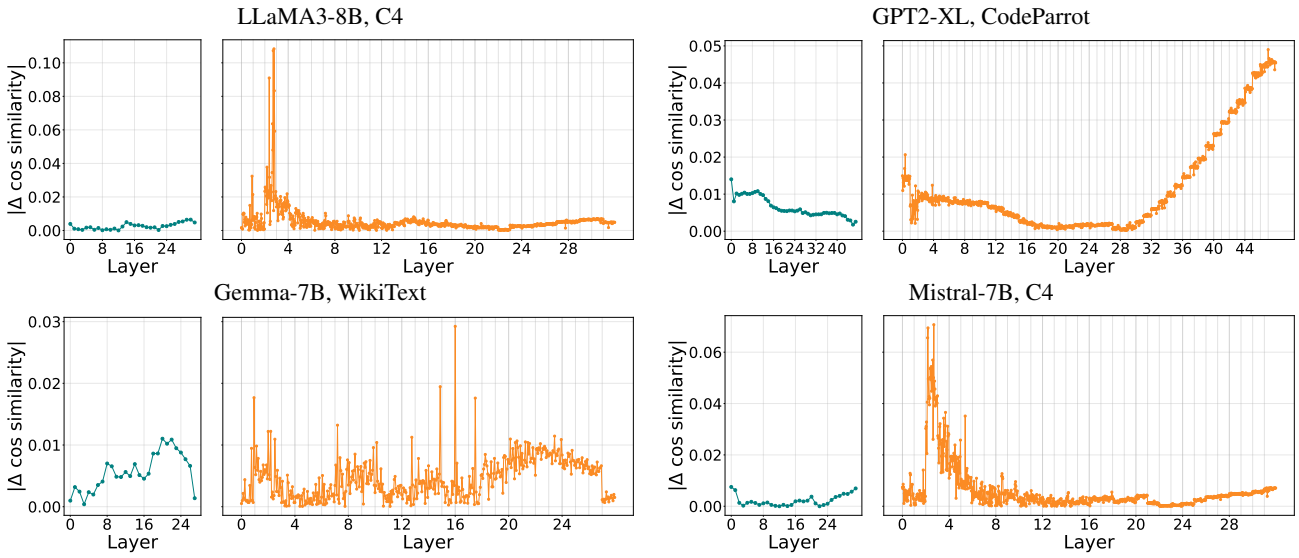

Figure 7. Absolute difference between the empirical average cosine similarity and the theoretical approximation (2) for individual heads, for X (blue plot) and $X + WZ$ (orange plot), across different models and datasets at context length 512. The x-axis indexes the heads in order, from the first to the last layer.

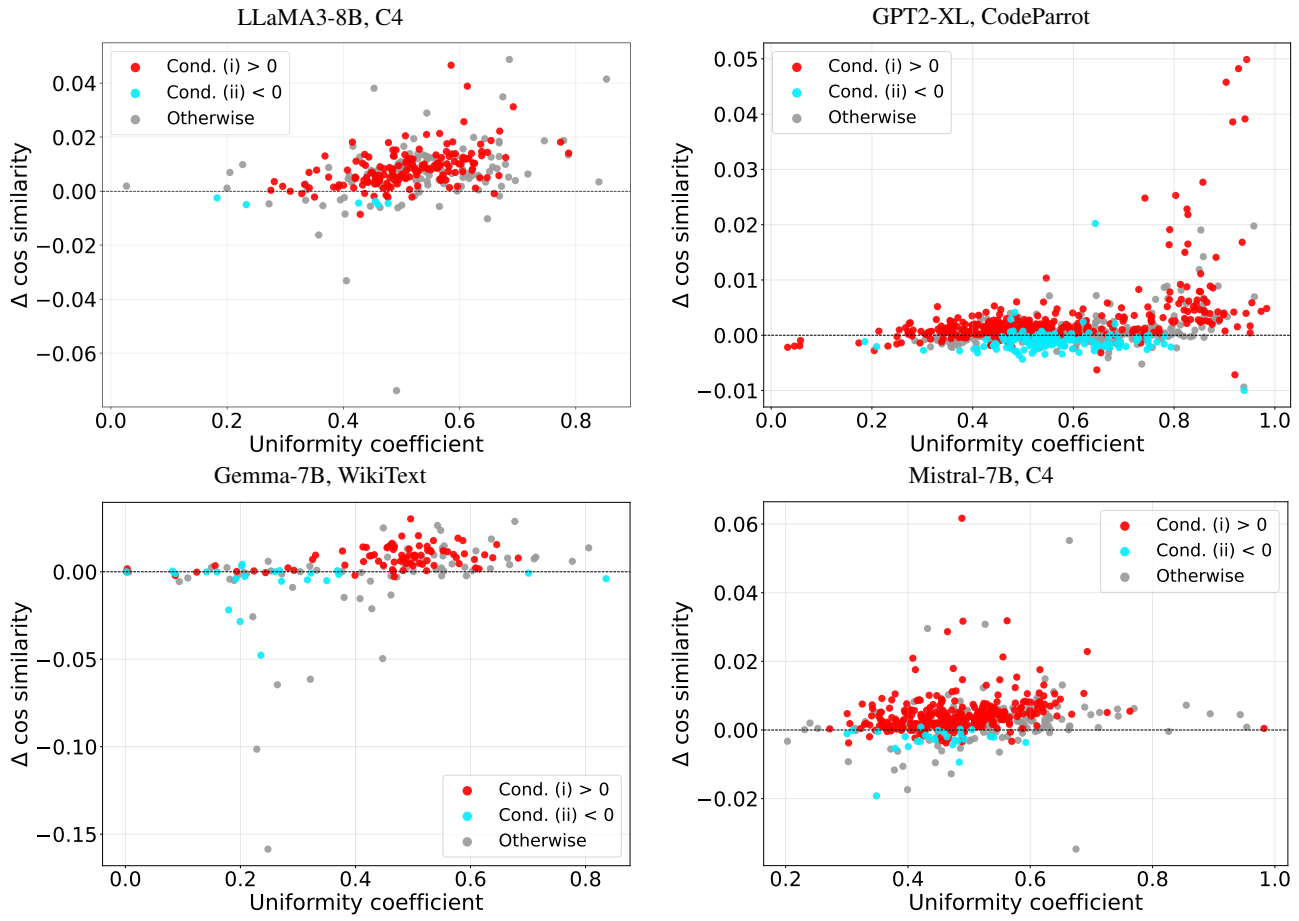

Figure 8. avg. $\cos\mathrm{sim}(X + WZA^\top) -$ avg. $\cos\mathrm{sim}(X + WZ)$ for each head of different models and datasets at context length 512. Heads satisfying $\beta\,\mathrm{tr}(BW) > 0$ (red) tend to yield positive values, whereas heads satisfying $\beta\,\mathrm{tr}(BW) + \mathrm{tr}(BW^\top W) < 0$ (blue) tend to have negative values. Heads satisfying neither condition are shown in gray.

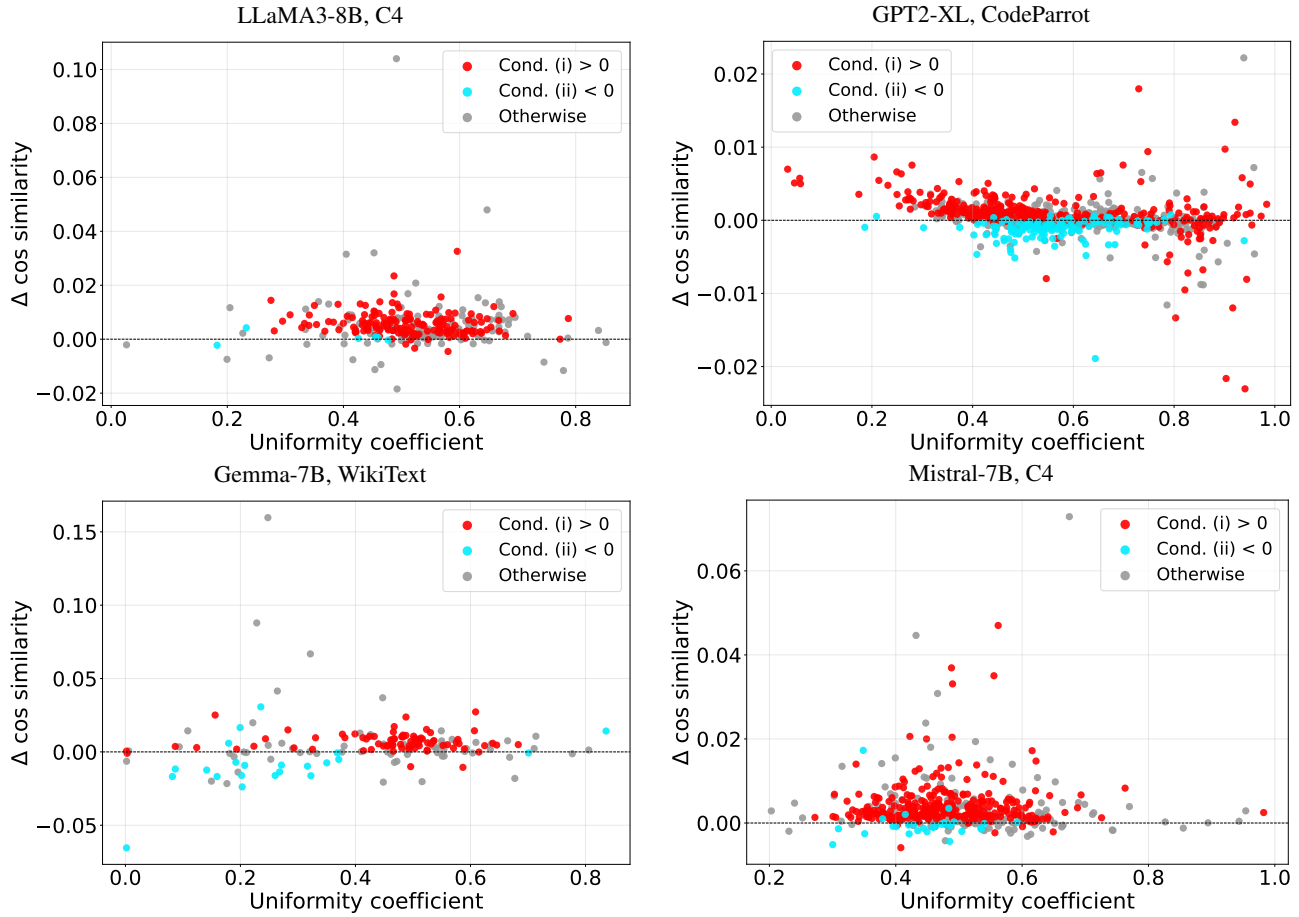

Figure 9. avg. $\cos\mathrm{sim}(X + WZA_u^\top) -$ avg. $\cos\mathrm{sim}(X + WZA^\top)$ for each head of different models and datasets at context length 512. Heads satisfying $\beta\,\mathrm{tr}(BW) > 0$ (red) tend to yield positive values, whereas heads satisfying $\beta\,\mathrm{tr}(BW) + \mathrm{tr}(BW^\top W) < 0$ (blue) tend to have negative values. Heads satisfying neither condition are shown in gray.

Across all models, this relative change becomes noticeably more positive once the uniformity coefficient exceeds a threshold of roughly 0.6. For some models, however, heads with very high uniformity exhibit smaller changes than heads with lower, though still high, uniformity. Since there are only few such heads, it is difficult to draw conclusions about this behavior.

Heads where greater uniformity does not increase oversmoothing. Finally, Figure 11 shows several examples of heads that do not follow the typical behavior described in this work. In particular, it includes cases where uniform attention yields less or equal token mixing than the learned head attention, even at relatively small uniformity coefficients (plots 1, 2, and 4), as well as a case where uniform attention produces less token mixing than applying no attention at all (plot 3).

E.2. Value mapping ranks

Figure 12 shows the distributions of ranks of the matrices of value vectors per layer, pooled across heads, 50 random sequences and four datasets (`TinyStories` (Eldan & Li, 2023), `tinyshakespeare` (Karpathy, 2015), `WikiText` (Merity et al., 2017a), and `CodeSearchNet-Python` (Husain et al., 2019)). Instead of raw ranks, we plot the rank of the matrix divided by $\min\{m, n\}$, where m, n are the dimensions of the matrix. For each model, we choose the context length to be as large as the inner dimension of the value mappings. As can be seen from the plots, models exhibit very high ranks in the middle layers. The only dataset on which outliers exist is `CodeSearchNet-Python` on gpt2, because the samples contain a lot of repeated tokens. Note that Llama and Gemma models are immune to repeated tokens and the ranks do not suffer deficiency even on those samples.

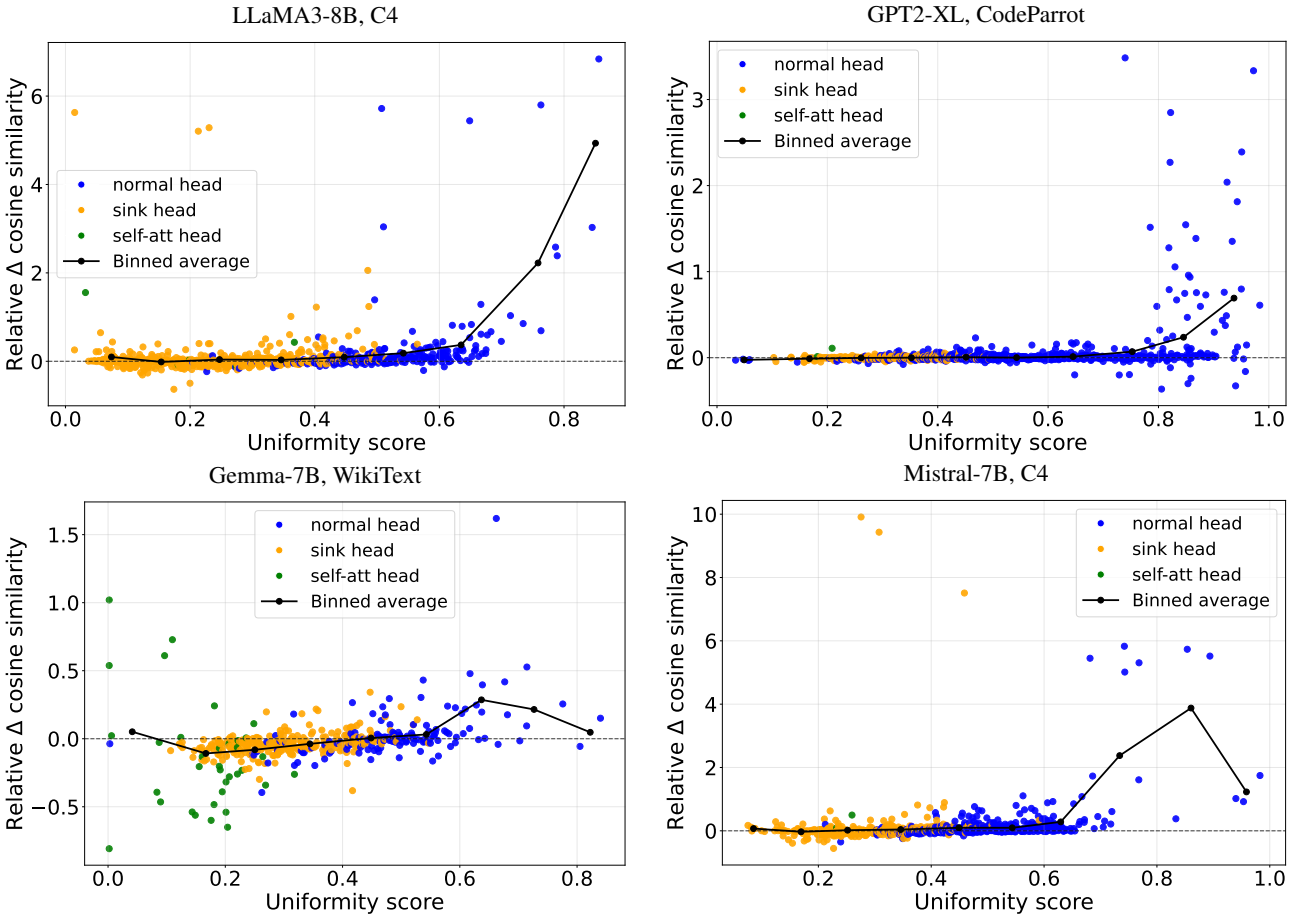

Figure 10. Relative change in average cosine similarity, $(\text{avg. } \cos\text{sim}(X + WZA^\top) - \text{avg. } \cos\text{sim}(X))/\,|\text{avg. } \cos\text{sim}(X)|$, for each head of different models and datasets at context length 512, plotted against the head's uniformity coefficient. Heads are classified as sink heads (yellow), self-attention heads (green), or neither (blue), using a threshold of 0.5 for both categories. The black line shows the average value across heads within bins of width 0.1.

E.3. Attention heads following the patterns from Section 7

Figure 13 displays heads whose patterns follow the assumptions in Theorems 7.1 and 7.2. As a source, we use the online tool `https://attention-motifs.github.io/v2/index.html` developed in (Ivanitskiy et al., 2025). The sequence we tested this on is:

love planning. It is one of my obsessions. After making numerous planning mistakes, however, I would like to share with you some tips I've learned along the way. If you are planning your child's. One of my favorite drinks when I return home is plantation iced tea. Last year, when I spent a week back home in Hawaii, I ended up drinking it as often as I found it on the menu. Now that the heat of summer is here, I'm dreaming of returning home to visit again. I really [. . .]

We note that the variance of the type of pattern as a function of the sequence is practically non-existent.

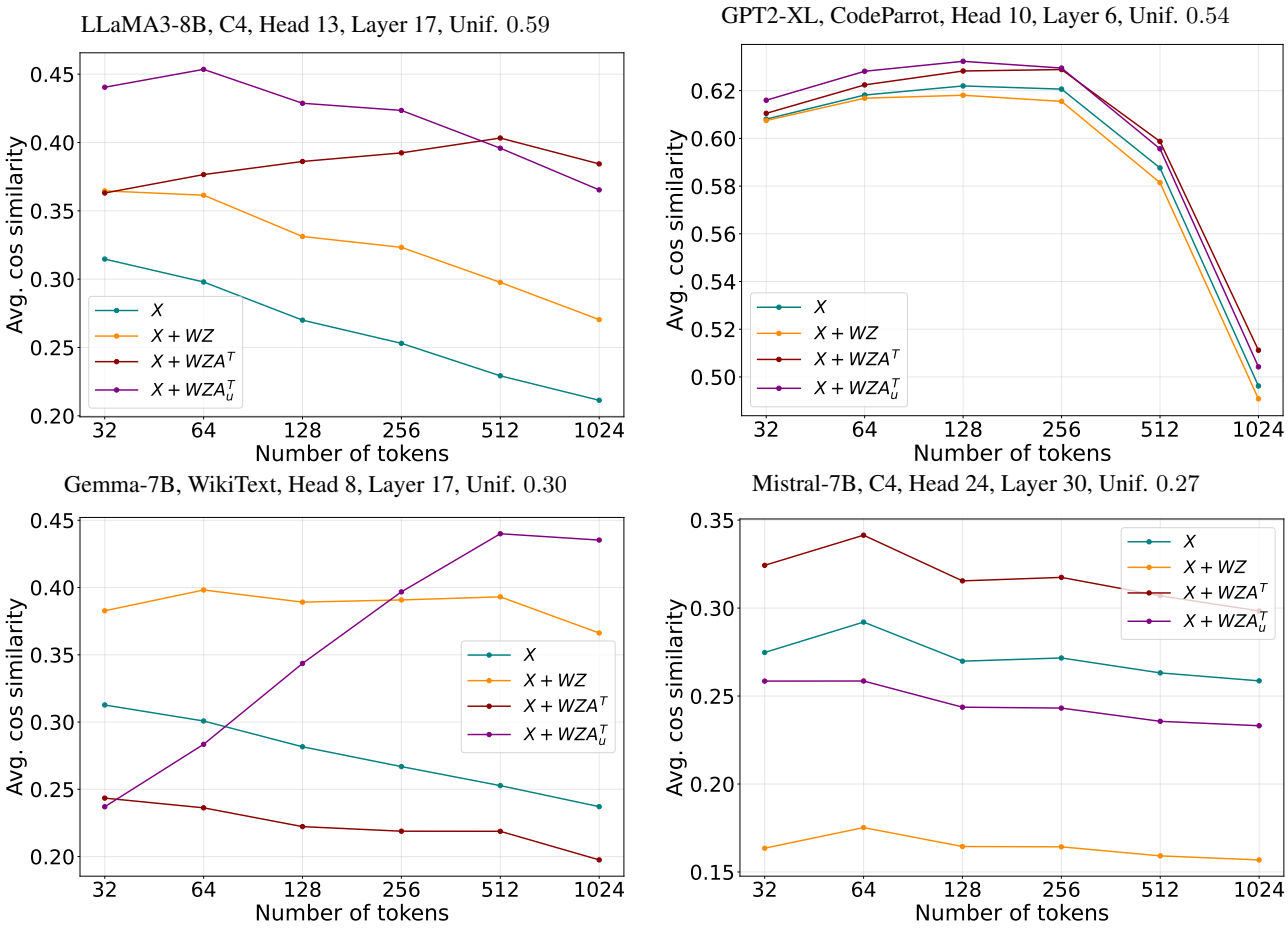

Figure 11. Empirical average cosine similarity for selected heads from different models and datasets, shown for different components of the attention step. These panels illustrate heads that deviate from the typical behavior described in the main text. The reported uniformity coefficient is averaged across context lengths.

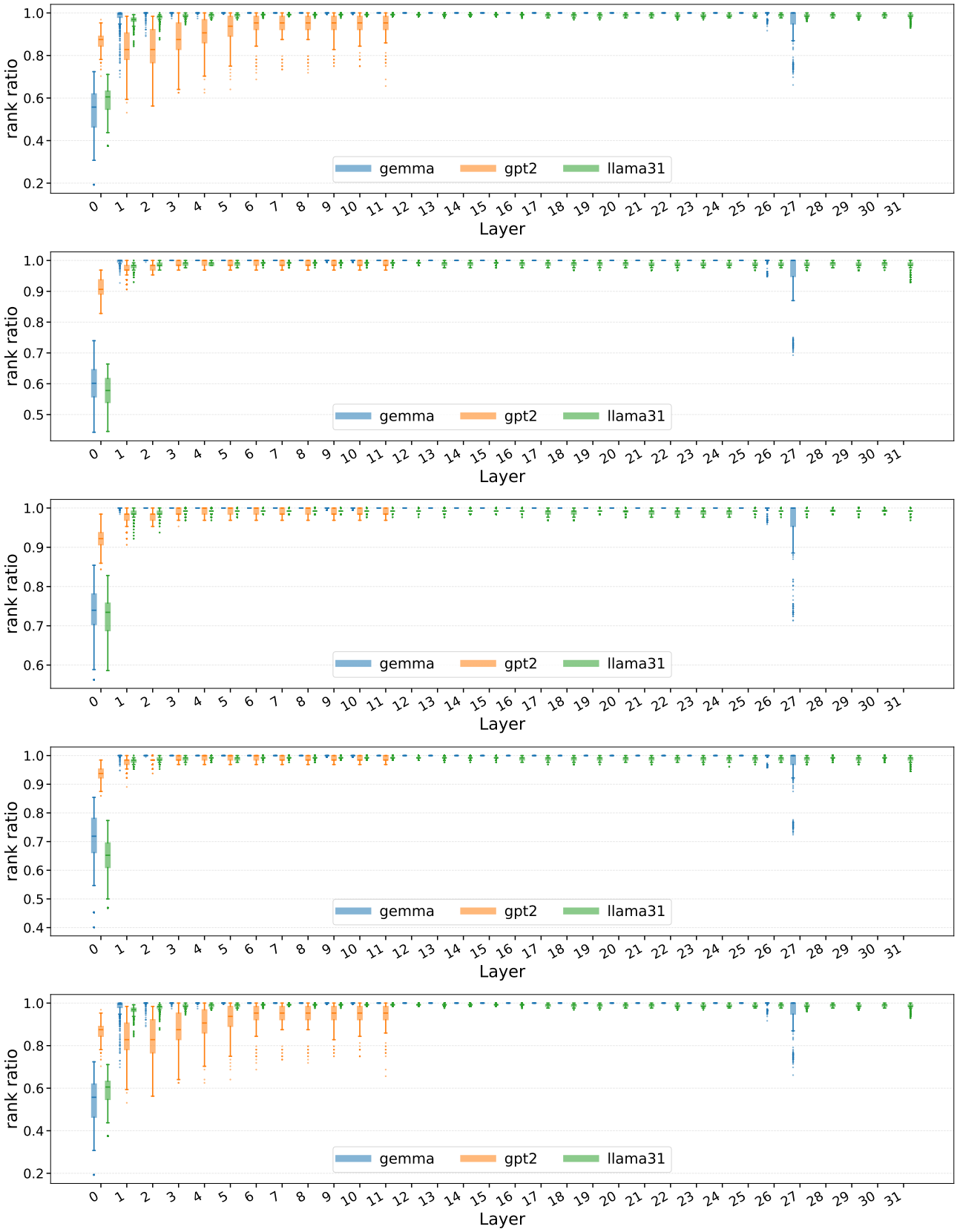

Figure 12. Rank distributions across layers, pooled over heads, 50 sequences and 4 datasets. The whiskers of the box plot point to absolute minimum and maximum. The relative rank tolerance is 0.001. **Upper:** Average for all four datasets. **Lower four:** Individual results for `TinyStories`, `tinyshakespeare`, `WikiText` and `CodeSearchNet-Python` (in this order).

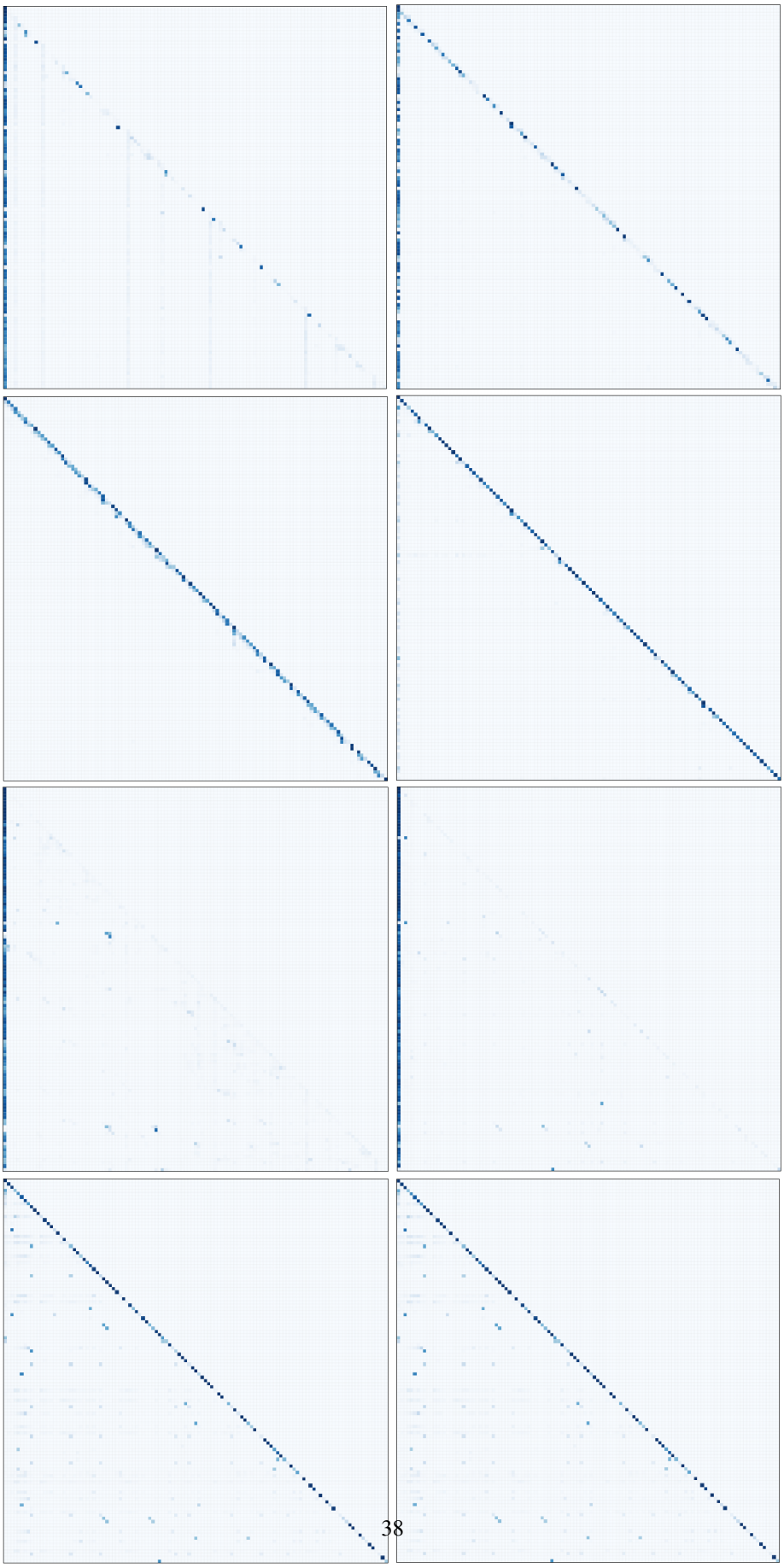

Figure 13. Attention patterns in pretrained transformers (taken randomly) from a pool of models available at `https://attention-motifs.github.io/v2/index.html`. The patterns satisfy the description of our distributions as copy-paste distribution sink (first row), copy-paste distribution diagonal (second row), copy clusters sink (Theorem 7.2) (third row) and