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

Peter Súkeník
Institute of Science and Technology (ISTA)
Austria
peter.sukenik@ista.ac.at

Cristina López Amado
Institute of Science and Technology (ISTA)
Austria
cristina.lopezamado@ista.ac.at

Christoph H. Lampert*
Institute of Science and Technology (ISTA)
Austria
chl@ist.ac.at

Marco Mondelli*
Institute of Science and Technology (ISTA)
Austria
marco.mondelli@ista.ac.at

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 is expected to smooth 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 [39] that enables communication between tokens and, hence, makes it possible for non-recurrent architectures to solve sequential modeling tasks. Unlike in GNNs [36], 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 [4] in language modeling. Thus, it came as a surprise when [44] revealed that many tokens in many attention heads do not communicate at all. More precisely, [44] 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 [13], vision transformers [7], vision-language models [23], multi-modal models [18] and diffusion-based models [42, 32].

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. [44] used sinks explicitly to generate long-form content without any fine-tuning and suggested to

*Equal contribution

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

assign a dedicated sink token during pre-training. Similar concepts were discussed in the follow-ups [22, 6]. [45] 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 [18], hallucination [38, 41], quantization issues [5], and training stability [28]. Thus, some works [28, 13] 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 [2, 1] claimed that sinks help transformers reduce oversmoothing and oversquashing. Others [14, 31] 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, [14] neither proves that sinks are the only mechanism for attention switch, nor it 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 [14, 31] 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 [14], 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. [14] argues that sinks implement the active-dormant mechanism and justifies the emergence of sinks in a near-convergence regime on a bigram-backcopy task. [47] 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, [3] shows that sinks are sometimes necessary for token counting and empirical histogram tasks. [2, 1] study sinks from the representational viewpoint and show that they help prevent oversmoothing and oversquashing, while [29] connects sinks with compression valleys arguing they are both manifestations of massive activations in the residual stream. [33] proposes a geometric interpretation in which sinks act as reference-frame anchors, simplifying the geometry and subsequent attention allocation. [27] argues that attention and residual sinks rescale and refine non-outlier features and attention weights. The contemporaneouswork [31] proves that sinks are necessary in an adaptation of the bigram-backcopy task from [14]. 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 [31] can be recovered from the more general version of our Proposition 6.1 presented in Appendix C as we prove in Appendix E. Finally, while diagonal patterns have been discussed e.g. in [15, 46], to the best of our knowledge we are the first to explicitly consider them in the context of oversmoothing and sinks.

Oversmoothing. [8] 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, [8] does not analyze their joint effect. [11] 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. [9] 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 [26, 12]. [43] shows that oversmoothing occurs in single-head self-attention without skip connections regardless of the attention mask, and layer normalization may prevent it. [19] extends the clustering results of [11] 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): [34] proves rank collapse at initialization for single-head attention without skip connections, layer normalization, or MLPs; [2, 1] 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 [40, 43] and rank collapse [37, 26, 12] (see [9] 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).

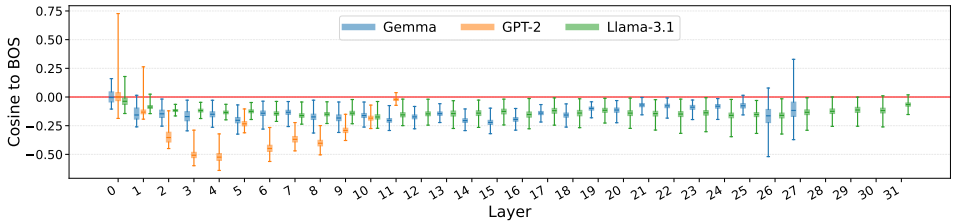

Figure 1: Cosine similarity between BOS and all other 256 token embeddings at the input to the attention layer, pooled across heads and 50 sequences across 4 language datasets (`TinyStories` [10], `tinyshakespeare` [20], `WikiText` [24], and `CodeSearchNet-Python` [16]). The whiskers of the box plot point to absolute minimum and maximum.

4 Warm-up: How transformers represent sinks

While sinks frequently appear in many heads per block [17, 13], 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 $z_0, z_1, \ldots, z_T$, 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.*

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 $z_0^T z_i \leq 0$ 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 G.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 deliberately choose diagonal attention rather than sink as the sparse endpoint: assuming a zero sink value, interpolating sinks with uniform attention only rescales the uniform attention contribution. Diagonal attention instead prevents inter-token communication while retaining nonzero updates. 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*

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

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 [19], 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 [8, 9, 26]. 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.

While Theorem 5.1 characterizes a single attention step, if its sufficient conditions hold consistently across layers of a transformer without MLP blocks, the layer-wise effects on the surrogate would accumulate, leading to more or less oversmoothing with depth as attention becomes denser.

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 sequences. We report results for the C4 dataset [30], and show in Figure 5 in Appendix G.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 G.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 (62) in Appendix G.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

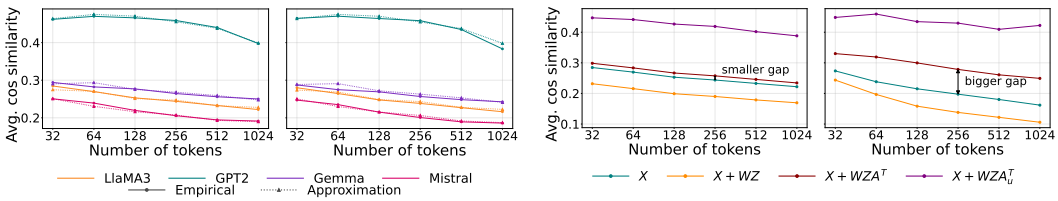

Figure 2: **Left:** empirical average cosine similarity (solid) and theoretical approximation (2) (dashed) for X (first plot) and $X + WZ$ (second plot), across models and context lengths. **Right:** empirical average cosine similarity in LLaMA3-8B for different components of the attention step, averaging over all heads (first plot) and over heads with uniformity coefficient (see (63) in Appendix G.1) larger than 0.6 (second plot). Cosine similarity generally increases as the attention matrix becomes denser.

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 G.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 [14]: 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 [14] which also supports it empirically (see Figure 1(b) therein). While [14] 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. 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 G.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 [14], 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 [14, 31]).

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 [21], 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 l_2 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 to a single-block optimality analysis, keeping the backcopy part of the distribution and dropping the bigram part. The context embeddings are split into the dormant set $\mathcal{D}$ and the copy-paste set $\mathcal{C}$, together with the BOS, indexed by 0. Let the token directions $\{h_0\} \cup \{h_i : i \in \mathcal{C} \cup \mathcal{D}\}$ be orthonormal, and let $p_1, \ldots, p_T$ be unit positional vectors orthogonal to all token directions. We write $E = \operatorname{span}\{p_1, \ldots, p_T\}$ and $r = \dim E$, so that $r + |\mathcal{C}| + |\mathcal{D}| + 1 \leq d$, while T may exceed d. The input embeddings are $x_{00} = \sqrt{d}\, h_0$ and $x_{ti} = \sqrt{d/2}(h_i + p_t)$ for $t \geq 1$. The first token after BOS is dormant; subsequent tokens can belong to either group, and every sequence of this form has positive probability. The block leaves BOS unchanged, maps a dormant input x_{ti} to $2x_{ti}$, and maps a copy-paste input x_{ti} to $x_{ti} + x_{t-1,j}$, where j is the preceding token.

Unlike orthogonal token embeddings, the positional vectors can share directions. We assume that they remain spread across their span: the matrix P with rows $p_t^\top$ satisfies $\|P\|_{\mathrm{op}}^2 \leq BT/r$ for some $B \geq 1$. We also assume that one linear map $R : E \to E$ can distinguish each immediate predecessor. More precisely, for some $\gamma \in (0, 1]$ and $L \geq 1$, $p_{t-1}^\top R p_t = 1,\ p_s^\top R p_t \leq 1 - \gamma,\ (2 \leq t \leq T,\ 1 \leq$

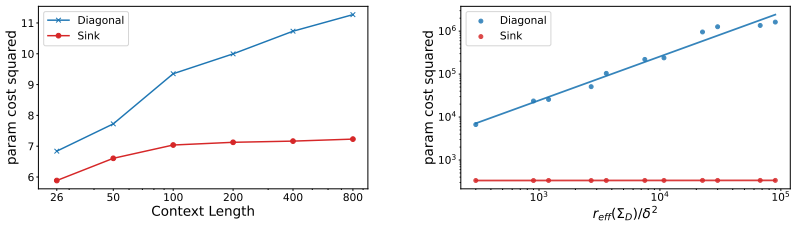

Figure 3: Frobenius squared cost of a single transformer block if the attention pattern is either sink or diagonal. **Left:** Backcopy task of Theorem 7.1 and cost as a function of the context length. **Right:** Generic task of Theorem 7.2 and cost as a function of $r_{\text{eff}}(\Sigma_D)/\delta^2$.

$s \le t,\ s \ne t-1$) while also $p_1^\top R p_1 \le 1,\ \|R\|_* \le Lr$. Thus, a shared query-key map can select the previous position without requiring orthogonality of position embeddings. Harmonic positional embeddings, closely related to RoPEs, satisfying these conditions are constructed in Appendix D.1.

Both attention patterns send copy-paste tokens to their immediate predecessor. They differ on dormant tokens: a sink sends them to BOS, whereas diagonal attention sends them to themselves. As above, we use the hard-attention approximation with logit margin $\kappa > 0$ and the standard $1/\sqrt{d}$ score normalization. Cost is measured as sum of squared Frobenius norms of matrix parameters.

Theorem 7.1. *Let $T \ge 3$, and let a single-head transformer block have an MLP hidden width of at least $r + |\mathcal{D}| + 1$. Under the distribution and positional assumptions above, put $C = |\mathcal{C}|$, $D = |\mathcal{D}|$, and $m = \min\{C, D\}$. There is a sink representation of cost at most*

$$U_{\text{sink}} = \frac{2\kappa}{\sqrt{d}}\left[\frac{2Lr}{\gamma} + \sqrt{2D}\left(1 + \frac{1}{\gamma}\right)\right] + (5 + 25C)(r + D) + 27C, \tag{3}$$

whereas every diagonal representation has cost at least

$$L_{\text{diag}} = 2\sqrt{\frac{2}{3}}\,\kappa\sqrt{\frac{mr(T-1)(T-2)}{dB}}. \tag{4}$$

If $U_{\text{sink}} < L_{\text{diag}}$, it is cheaper to represent this mapping with a sink than with diagonal attention.

This result, proved in Appendix D.1, gives an analytic condition for a sink to incur lower cost. For fixed token-set sizes, the value/output projections and MLP in our sink construction have cost $\mathcal{O}(r)$, and the total sink upper bound is $\mathcal{O}(r + \kappa Lr/(\sqrt{d}\,\gamma))$. The diagonal lower bound scales as $\kappa T\sqrt{mr/(dB)}$. Thus, the comparison depends on both the positional rank and the separation achieved by the shared map; their dependence on T must be accounted for. Appendix D.1.2 shows that harmonic frames can be chosen with $r, d = \Theta(\log T)$ and bounded B, L, γ^{-1}. For fixed κ, this gives a sink cost of $\mathcal{O}(\log T)$ and an optimal diagonal cost of $\Theta(T)$. There are also fixed-dimensional families with sink cost $\mathcal{O}(T^\varepsilon)$ and optimal diagonal cost $\Theta(T)$ for every prescribed $\varepsilon > 0$, with the fixed dimension chosen depending on ε; see Appendix D.1.1. Finally, at fixed positional geometry and token-set sizes, increasing d makes the sufficient inequality harder to satisfy: it reduces the attention-cost terms by $1/\sqrt{d}$ while leaving the value/output and MLP costs unchanged.

Figure 3 compares the total cost of a single transformer block exhibiting either diagonal or sink attention patterns. The positional embedding rank first tracks dimension until $d = 100$, then remains fixed ($D = C = 5$). The estimated polynomial growth rates are: 0.85 for sink solutions with $T < 100$, 0.09 for sink solutions with $T > 100$. For diagonal, it is 1.87 for $T < 100$ and 0.94 for $T > 100$. These results are agree with the rates presented in Theorem 7.1. The slight deviations from precise numbers (0.94 vs 1, 0.09 vs. 0 and 0.85 vs. 1) are best explained by finiteness of the parameters and thus the asymptotics being only approximate, as well as experimental confounds (training with a fixed number of steps and learning rate might affect different T values differently).

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 $C \ge 1$ 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, 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 G.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:*

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

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

This result (proved in Appendix D.2) 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 (5) 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 (5). 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 (5) 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.4 in Appendix D.2 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 (5)), 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 (RoPE). Our analysis is formulated for absolute positional embeddings, but a similar geometric perspective can be developed for RoPE. Under RoPE, queries and keys are rotated according to their positions, $q_i = \Phi_i W_Q x_i$ and $k_j = \Phi_j W_K x_j$, yielding attention scores $S_{ij} = \frac{1}{\sqrt{d}} x_i^\top W_Q^\top \Phi_i^\top \Phi_j W_K x_j$. For a fixed context length, we can interpret this through a fixed position-indexed geometry by defining $z_{a,i} = \Phi_i x_a$ for an embedding x_a occurring at position i. Their inner products satisfy $z_{a,i}^\top z_{b,j} = x_a^\top \Phi_i^\top \Phi_j x_b$, so similarity depends on both the embeddings and their positions.

Importantly, this geometry can be related directly to the query-key computation. Defining $\widetilde{W}_{Q,i} = \Phi_i W_Q \Phi_i^\top$ and $\widetilde{W}_{K,i} = \Phi_i W_K \Phi_i^\top$, we obtain $q_i = \widetilde{W}_{Q,i} z_{a,i}$ and $k_j = \widetilde{W}_{K,j} z_{b,j}$. Moreover, orthogonality of the rotations implies $\|\widetilde{W}_{Q,i}\|_F = \|W_Q\|_F$ and $\|\widetilde{W}_{K,i}\|_F = \|W_K\|_F$. Thus, the RoPE scores can be expressed exactly using position-indexed embeddings and transformed query-key matrices, without changing their individual Frobenius costs. However, unlike in the original setting, these transformed matrices depend on position and must all originate from the same shared W_Q, W_K.

This motivates defining dormant groups using the position-indexed query-key geometry: the same embedding at different positions may belong to different groups, while different embeddings may belong to the same group if their position-indexed representations are sufficiently close. The query-key and residual-stream parts of our analysis are then affected differently. The sink construction becomes a problem of satisfying attention-score constraints over position-indexed dormant groups. Its cost must additionally account for the requirement that all position-dependent query-key matrices arise from a single shared pair W_Q, W_K. The value-output and MLP constructions, however, remain unchanged since they do not need to distinguish individual dormant groups.

Similarly, the diagonal lower bound depends only on query-key score constraints, which can be aggregated into a position-indexed generalization of Σ_D. We discuss these extensions in Appendix F. A fully rigorous RoPE cost comparison, including the shared-parameter constraints and generalized effective-rank bounds, remains an interesting direction for future work.

Summary. Diagonal attention switches off token communication, 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 depth-wise progression of oversmoothing in multi-head transformers and under 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?

Acknowledgments and Disclosure of Funding

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

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

[2] Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João Madeira Araújo, Oleksandr Vitvitskyi, Razvan Pascanu, and Petar Veličković. Transformers need glasses! information over-squashing in language tasks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[3] Freya Behrens, Luca Biggio, and Lenka Zdeborová. Understanding counting in small transformers: The interplay between attention and feed-forward layers. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

[4] Yoshua Bengio, Paolo Frasconi, and Marco Gori. Recurrent neural networks for adaptive temporal processing. In *Proceedings of the 6th Italian Workshop on Parallel Architectures and Neural Networks WIRN93*, pages 85–117, 1993.

[5] Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[6] Zefan Cai, Yichi Zhang, Bofei Gao, Yuliang Liu, Yucheng Li, Tianyu Liu, Keming Lu, Wayne Xiong, Yue Dong, Junjie Hu, and Wen Xiao. Pyramidkv: Dynamic kv cache compression based on pyramidal information funneling. In *First Conference on Language Modeling*, 2025.

[7] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

[8] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning (ICML)*, 2021.

[9] Gbetondji Jean-Sebastien Dovonon, Michael M. Bronstein, and Matt Kusner. Setting the record straight on transformer oversmoothing. In *Transactions on Machine Learning Research (TMLR)*, 2025.

[10] Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023.

[11] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[12] Alessio Giorlandino and Sebastian Goldt. 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.

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

[14] Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I. Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in LLMs. In *The Second Conference on Parsimony and Learning (Recent Spotlight Track)*, 2025.

[15] Victoria Hankemeier and Malte Schilling. Stochastic parroting in temporal attention–regulating the diagonal sink. *arXiv preprint arXiv:2602.10956*, 2026.

[16] Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. CodeSearchNet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436*, 2019.

[17] Michael Ivanitskiy, Cecilia Diniz Behn, and Samy Wu Fung. Motifs in attention patterns of large language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025.

[18] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations (ICLR)*, 2025.

[19] Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[20] Andrej Karpathy. char-rnn. `https://github.com/karpathy/char-rnn`, 2015.

[21] Seijin Kobayashi, Yassir Akram, and Johannes Von Oswald. Weight decay induces low-rank attention layers. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[22] Jia-Nan Li, Quan Tu, Cunli Mao, Zhengtao Yu, Ji-Rong Wen, and Rui Yan. Streamingdialogue: Prolonged dialogue learning via long context compression with minimal losses. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[23] Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models. In *International Conference on Learning Representations (ICLR)*, 2026.

[24] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

[25] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017.

[26] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[27] Zihan Qiu, Zeyu Huang, Kaiyue Wen, Peng Jin, Bo Zheng, Yuxin Zhou, Haofeng Huang, Zekun Wang, Xiao Li, Huaqing Zhang, et al. A unified view of attention and residual sinks: Outlier-driven rescaling is essential for transformer training. *arXiv preprint arXiv:2601.22966*, 2026.

[28] Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[29] Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin. *arXiv preprint arXiv:2510.06477*, 2025.

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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

[32] Maximo Eduardo Rulli, Simone Petruzzi, Edoardo Michielon, Fabrizio Silvestri, Simone Scardapane, and Alessio Devoto. Attention sinks in diffusion language models. *arXiv preprint arXiv:2510.15731*, 2025.

[33] Valeria Ruscio, Umberto Nanni, Fabrizio Silvestri, et al. What are you sinking? a geometric approach on attention sink. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[34] Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. *arXiv preprint arXiv:2410.07799*, 2024.

[35] Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. In *International Conference on Machine Learning (ICML)*, 2024.

[36] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[37] Han Shi, JIAHUI GAO, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen M. S. Lee, and James Kwok. Revisiting over-smoothing in BERT from the perspective of graph. In *International Conference on Learning Representations (ICLR)*, 2022.

[38] Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *International Journal of Computer Vision*, 134(1):22, 2026.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[40] Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In *International Conference on Learning Representations (ICLR)*, 2022.

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

[42] Yuxin Wen, Jim Wu, Ajay Jain, Tom Goldstein, and Ashwinee Panda. Analysis of attention in video diffusion transformers. *arXiv preprint arXiv:2504.10317*, 2025.

[43] Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the role of attention masks and layernorm in transformers. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

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

[45] Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In *International Conference on Machine Learning (ICML)*, 2024.

[46] Shuangfei Zhai. Exclusive self attention. *arXiv preprint arXiv:2603.09078*, 2026.

[47] Stephen Zhang, Mustafa Khan, and Vardan Papyan. 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:

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

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

$$X^{(l+1)} = X^{(l)} + W_{VO}^{(l)} \operatorname{RMS}(X^{(l)})(A^{(l)})^\top = X^{(l)} + W_{VO}^{(l)} X^{(l)} D^{(l)} (A^{(l)})^\top, \tag{6}$$

where $D^{(l)} = \operatorname{diag}(d_1^{(l)}, \ldots, d_T^{(l)})$, 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 (6), 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

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

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

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

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 (6),

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

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*

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

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.*

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:

$$Y_i(\lambda) = \beta Z_i + (WZA_\lambda^\top)_i = \beta Z_i + \sum_{k=1}^T a_{ik}^\lambda W Z_k. \tag{7}$$

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

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

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

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

From Equation (7), we have

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

Combining both equations, we have

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

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) Case $i < j$:

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

(ii) Case $i = j$:

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

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 $b_i \in \text{Span}(\{W_{VO} a_k\})$. As this holds for every b_i, we get $\text{Span}(\{b_i\}) \subset \text{Span}(\{W_{VO} a_i\})$. As $\text{rank}(\{a_i\}) = \text{rank}(\{b_i\})$, W_{VO} has full rank on $\text{Span}(\{a_i\})$, otherwise the rank inequality could not be satisfied. This means that the kernel of W_{VO} on $\text{Span}(\{a_i\})$ 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 [35]. 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

D.1 Backcopy with a low-dimensional positional frame

We give the complete statement and proof of Theorem 7.1. We then give both fixed-dimensional and logarithmic-dimensional families for which sinks are asymptotically cheaper.

Data and positional assumptions. Let $\mathcal{C}$ and $\mathcal{D}$ be disjoint, nonempty sets of copy-paste and dormant token indices, respectively, and put $C = |\mathcal{C}|$, $D = |\mathcal{D}|$, and $m = \min\{C, D\}$. Let $\{h_0\} \cup \{h_i : i \in \mathcal{C} \cup \mathcal{D}\}$ be an orthonormal family in $\mathbb{R}^d$. Let $p_1, \ldots, p_T$ be unit vectors orthogonal to every token direction, and define

$$E = \text{span}\{p_1, \ldots, p_T\}, \qquad r = \dim E, \qquad P = \begin{bmatrix} p_1^\top \\ \vdots \\ p_T^\top \end{bmatrix}.$$

In particular, $1 \leq r \leq T$ and $r + C + D + 1 \leq d$. Assume, for some $B \geq 1$, that

$$\|P\|_{\text{op}}^2 \leq \frac{BT}{r}. \tag{8}$$

Assume also that there are a linear map $R : E \to E$, a gap $\gamma \in (0, 1]$, and $L \geq 1$ such that

$$\begin{aligned} p_{t-1}^\top R p_t &= 1 && (2 \leq t \leq T), \\ p_s^\top R p_t &\leq 1 - \gamma && (2 \leq t \leq T,\ 1 \leq s \leq t,\ s \neq t-1), \\ p_1^\top R p_1 &\leq 1, \qquad \|R\|_* \leq Lr. \end{aligned} \tag{9}$$

We extend R by zero on $E^\perp$.

Set $s_0 = \sqrt{d/2}$ and define

$$x_{00} = \sqrt{d}\, h_0, \qquad x_{ti} = s_0(h_i + p_t) \quad (1 \leq t \leq T,\ i \in \mathcal{C} \cup \mathcal{D}). \tag{10}$$

An input sequence has BOS at position 0, a dormant token at position 1, and an arbitrary ordinary token at each position $2, \ldots, T$. Every such sequence is in the support of the distribution. For an input sequence $x_0 = x_{00}$, $x_t = x_{t,i_t}$, its target is

$$f^*(x)_t = \begin{cases} x_0, & t = 0, \\ 2x_t, & i_t \in \mathcal{D}, \\ x_t + x_{t-1}, & i_t \in \mathcal{C}. \end{cases} \tag{11}$$

The last case occurs only for $t \geq 2$.

Block, routing, and cost. We use ideal RMSNorm, $\mathrm{RMS}(y) = \sqrt{d}\, y/\|y\|_2$ for $y \neq 0$, with no learned gain. All inputs in (10) have norm $\sqrt{d}$, so the first RMSNorm leaves them unchanged. Consistently with the paper's convention, put $W_{QK} = W_Q^\top W_K$. For convenience in the proof, write

$$M = W_{QK}^\top = W_K^\top W_Q, \qquad S_{ts} = \frac{x_t^\top W_{QK} x_s}{\sqrt{d}} = \frac{x_s^\top M x_t}{\sqrt{d}}, \quad 0 \leq s \leq t. \tag{12}$$

A prescribed key $\pi(t)$ must satisfy $S_{t,\pi(t)} \geq S_{ts} + \kappa$ for every competing causal key $s \neq \pi(t)$, and is assigned attention weight one in the hard-attention model. Both patterns have $\pi(0) = 0$ and $\pi(t) = t - 1$ for copy-paste tokens. For dormant tokens, the sink pattern has $\pi(t) = 0$, whereas the diagonal pattern has $\pi(t) = t$.

The block is

$$y_t = x_t + W_O W_V x_{\pi(t)}, \qquad f_\Theta(x)_t = y_t + W_2 \sigma\big(W_1\, \mathrm{RMS}(y_t) + b_1\big) + b_2, \tag{13}$$

where σ is coordinatewise ReLU. Query, key, value, and output matrices are $d \times d$, and the MLP has hidden width w. Its biases are unregularized, and its weight cost is

$$\mathcal{R}(\Theta) = \|W_Q\|_F^2 + \|W_K\|_F^2 + \|W_V\|_F^2 + \|W_O\|_F^2 + \|W_1\|_F^2 + \|W_2\|_F^2. \tag{14}$$

An admissible representation satisfies both the prescribed routing margins and $f_\Theta(x) = f^*(x)$ on every sequence in the support.

Theorem D.1 (Full version of Theorem 7.1)**.** *Assume $T \geq 3$, the data and positional conditions above, and $w \geq r + D + 1$. Define*

$$U_{\mathrm{sink}} = \frac{2\kappa}{\sqrt{d}} \left[\frac{2Lr}{\gamma} + \sqrt{2D} \left(1 + \frac{1}{\gamma} \right) \right] + (5 + 25C)(r + D) + 27C, \tag{15}$$

$$L_{\mathrm{diag}} = 2\sqrt{\frac{2}{3}}\, \kappa \sqrt{\frac{mr(T-1)(T-2)}{dB}}. \tag{16}$$

There exists an admissible sink representation of cost at most U_{sink}. Every admissible diagonal representation has cost at least L_{diag}. Consequently, $U_{\mathrm{sink}} < L_{\mathrm{diag}}$ implies that a sink representation is strictly cheaper than every admissible diagonal representation.

Proof. We first recall the factorization identity

$$\inf_{A^\top B = N} \big(\|A\|_F^2 + \|B\|_F^2 \big) = 2\|N\|_*. \tag{17}$$

Indeed, $\|A^\top B\|_* \leq \|A\|_F \|B\|_F \leq (\|A\|_F^2 + \|B\|_F^2)/2$. For a singular value decomposition $N = U\Sigma V^\top$, equality is attained by $A = \Sigma^{1/2} U^\top$ and $B = \Sigma^{1/2} V^\top$, padding with zero rows if necessary.

Upper bound on the sink representation cost. Let $h_{\mathcal{D}} = \sum_{b\in\mathcal{D}} h_b$ and set

$$\beta = \frac{2\kappa}{\sqrt{d}\,\gamma}, \qquad \alpha = \frac{1}{\sqrt{2}}\left(\beta + \frac{2\kappa}{\sqrt{d}}\right), \qquad M_{\text{sink}} = \beta R + \alpha h_0 h_{\mathcal{D}}^\top. \tag{18}$$

For an ordinary query at position t and an ordinary key at position s, orthogonality of the token and positional subspaces gives

$$S_{ts} = \frac{\beta\sqrt{d}}{2}\, p_s^\top R p_t. \tag{19}$$

For a copy-paste query, the BOS score is zero. Its predecessor has score $\beta\sqrt{d}/2 = \kappa/\gamma \geq \kappa$, and its score exceeds every competing ordinary-key score by at least $\beta\sqrt{d}\,\gamma/2 = \kappa$. Thus every copy-paste query selects its immediate predecessor.

For a dormant query, the BOS score is $\alpha\sqrt{d/2}$. Every ordinary-key score is at most $\beta\sqrt{d}/2$, including at $t = 1$ by the final positional condition. Therefore

$$S_{t0} - S_{ts} \geq \alpha\sqrt{d/2} - \frac{\beta\sqrt{d}}{2} = \kappa.$$

Every dormant query selects BOS. At position 0 the only causal key is BOS itself. By (17), the query and key matrices can be chosen with total squared cost

$$2\|M_{\text{sink}}\|_* \leq 2\beta\|R\|_* + 2\alpha\|h_0 h_{\mathcal{D}}^\top\|_* \leq 2\beta L r + 2\alpha\sqrt{D}. \tag{20}$$

Let Π be the orthogonal projector onto $E \oplus \operatorname{span}\{h_i : i \in \mathcal{C}\cup\mathcal{D}\}$, and choose $W_V = W_O = \Pi$. This projector annihilates BOS and preserves every ordinary input. Its cost is

$$\|W_V\|_F^2 + \|W_O\|_F^2 = 2(r + C + D). \tag{21}$$

The post-attention residual is consequently $y_0 = x_0$, $y_t = x_t$ on dormant tokens, and $y_t = x_t + x_{t-1}$ on copy-paste tokens. The MLP must therefore add x_t only on dormant tokens.

To construct this MLP, let $F = E \oplus \operatorname{span}\{h_b : b \in \mathcal{D}\}$ and $n = \dim F = r + D$. Choose an orthonormal basis $u_1, \ldots, u_n$ consisting of a basis of E and the dormant token directions. Define $h_{\mathcal{C}} = \sum_{a\in\mathcal{C}} h_a$ and, for an MLP input z,

$$\begin{aligned} A(z) &= h_{\mathcal{C}}^\top z, \\ H_j(z) &= \mathrm{ReLU}(u_j^\top z - 5A(z) + s_0), \quad 1 \leq j \leq n, \\ H_0(z) &= \mathrm{ReLU}(-5A(z) + s_0), \\ g(z) &= \sum_{j=1}^{n} u_j\big(H_j(z) - H_0(z)\big). \end{aligned} \tag{22}$$

This is a two-layer ReLU MLP of width $n + 1$, with every hidden bias equal to s_0 and output bias zero.

For a dormant token, $z = \mathrm{RMS}(y_t) = x_t \in F$, $A(z) = 0$, and $|u_j^\top z| \leq s_0$. It follows that $H_j(z) = u_j^\top z + s_0$, $H_0(z) = s_0$, and $g(z) = \sum_j u_j u_j^\top z = z = x_t$. For BOS, all inner products with $h_{\mathcal{C}}$ and u_j vanish, so $H_j(z) = H_0(z) = s_0$ and $g(z) = 0$.

For a copy-paste token, write $y = x_{t,a} + x_{t-1,i}$ with $a \in \mathcal{C}$. We have $\|y\|_2 \leq 2\sqrt{d}$ and $h_{\mathcal{C}}^\top y \geq s_0$. Thus, for $z = \mathrm{RMS}(y)$, $A(z) \geq s_0/2$ and $|u_j^\top z| \leq \sqrt{d} = \sqrt{2}\, s_0$. Consequently,

$$u_j^\top z - 5A(z) + s_0 \leq \left(\sqrt{2} - \frac{3}{2}\right) s_0 < 0, \qquad -5A(z) + s_0 \leq -\frac{3}{2} s_0 < 0.$$

All hidden units vanish, so $g(z) = 0$. The block therefore gives exactly the targets in (11).

The first-layer rows in (22) are $u_j^\top - 5h_{\mathcal{C}}^\top$ and $-5h_{\mathcal{C}}^\top$; the second-layer columns are u_j and $-\sum_j u_j$. Since $h_{\mathcal{C}} \perp F$ and $\|h_{\mathcal{C}}\|_2^2 = C$,

$$\|W_1\|_F^2 = n(1 + 25C) + 25C, \qquad \|W_2\|_F^2 = 2n. \tag{23}$$

If $w > n + 1$, the remaining units are set to zero. Adding (20), (21), and (23) gives

$$\mathcal{R}(\Theta) \leq 2\beta L r + 2\alpha\sqrt{D} + 2(r + C + D) + (3 + 25C)(r + D) + 25C = U_{\text{sink}}.$$

Lower bound on the diagonal representation cost. Consider any admissible diagonal representation, and write $M = W_K^\top W_Q$. Choose m disjoint pairs $(a_j, b_j) \in \mathcal{C} \times \mathcal{D}$ and put

$$v_j = \frac{h_{b_j} - h_{a_j}}{\sqrt{2}}, \qquad V = [v_1 \ \cdots \ v_m].$$

The columns of V are orthonormal. For every $t = 3, \dots, T$, the support contains a sequence with a_j at both positions $t-1, t$ and a sequence with b_j at both positions. The former query must select its predecessor, whereas the latter must select itself. Comparing these two keys gives

$$\frac{1}{\sqrt{d}}(x_{t-1,a_j} - x_{t,a_j})^\top M x_{t,a_j} \geq \kappa, \tag{24}$$

$$\frac{1}{\sqrt{d}}(x_{t,b_j} - x_{t-1,b_j})^\top M x_{t,b_j} \geq \kappa. \tag{25}$$

Adding the inequalities cancels the query positional component. With $\Delta p_t = p_t - p_{t-1}$, we obtain

$$2\kappa \leq \frac{s_0^2}{\sqrt{d}} \Delta p_t^\top M (h_{b_j} - h_{a_j}) = \sqrt{\frac{d}{2}} \Delta p_t^\top M v_j.$$

Therefore

$$(p_t - p_{t-1})^\top M v_j \geq a, \qquad a = \frac{2\sqrt{2}\,\kappa}{\sqrt{d}}. \tag{26}$$

We use the following elementary sequence estimate. If $z_{i+1} - z_i \geq a > 0$ for $i = 1, \dots, N-1$, then, with $\bar{z} = N^{-1} \sum_i z_i$,

$$\begin{aligned} \sum_{i=1}^N z_i^2 &\geq \sum_{i=1}^N (z_i - \bar{z})^2 = \frac{1}{N} \sum_{i<j} (z_j - z_i)^2 \\ &\geq \frac{a^2}{N} \sum_{i<j} (j-i)^2 = \frac{a^2}{12} N(N^2-1). \end{aligned} \tag{27}$$

Apply this with $N = T-1$ to the scores $p_2^\top M v_j, \dots, p_T^\top M v_j$. Writing P' for the submatrix of P with rows $2, \dots, T$ and summing over j yields

$$\|P' M V\|_F^2 \geq \frac{m a^2}{12} (T-1)\big((T-1)^2 - 1\big). \tag{28}$$

On the other hand, (8) and $\|V\|_{\mathrm{op}} = 1$ imply

$$\|P' M V\|_F \leq \|P\|_{\mathrm{op}} \|M\|_F \leq \sqrt{\frac{BT}{r}}\, \|M\|_*. \tag{29}$$

Combining these inequalities with $\mathcal{R}(\Theta) \geq \|W_Q\|_F^2 + \|W_K\|_F^2 \geq 2\|M\|_*$ gives

$$\mathcal{R}(\Theta) \geq \frac{2a\sqrt{m}}{\sqrt{12}} \sqrt{\frac{r(T-1)((T-1)^2-1)}{BT}} = 2\sqrt{\frac{2}{3}}\, \kappa \sqrt{\frac{mr(T-1)(T-2)}{dB}}.$$

Here we used $((T-1)^2 - 1) = T(T-2)$. This proves the lower bound and, together with the sink construction, the claimed sufficient comparison. □

Support and cost conventions. The diagonal proof only requires the repeated copy-paste and repeated dormant pairs used above; full support is a simple sufficient assumption. Starting the increment argument at $t = 3$ is necessary because position 1 is dormant. The lower bound concerns the prescribed diagonal routing class, even if other attention patterns can compute the same output mapping. If hidden biases are included in the penalty, the displayed sink construction incurs the additional cost $(r + D + 1)d/2$. All factorized matrix costs above use the literal sum of squared Frobenius norms, and therefore include the factor 2 in (17).

Dependence on context length. Because $r = \operatorname{rank} P \le T$, the frame factor $\max\{1, T/r\}$ equals T/r. Thus (16) is exactly the form obtained from

$$\frac{2\sqrt{2/3}\,\kappa}{\sqrt{dB}}\sqrt{\frac{m(T-1)((T-1)^2-1)}{\max\{1,T/r\}}}.$$

For orthonormal positions, $r = T$ and $B = 1$, recovering the $T^{3/2}/\sqrt{d}$ dependence. For $T \gg r$, the lower bound is of order $\kappa T\sqrt{mr/(dB)}$. At fixed token-set sizes, the upper bound is $O(r + \kappa Lr/(\sqrt{d}\,\gamma))$.

The dependence of L and γ on T cannot generally be omitted. Indeed, for $1 \le i < j \le T-1$, apply (9) at query position $j+1$ to obtain

$$(p_j - p_i)^\top R p_{j+1} \ge \gamma, \qquad \|p_j - p_i\|_2 \ge \frac{\gamma}{\|R\|_{\mathrm{op}}} \ge \frac{\gamma}{Lr}.$$

Disjoint balls of radius $\gamma/(2Lr)$ around these unit vectors lie inside a ball of radius $1 + \gamma/(2Lr)$ in E. Comparing volumes gives

$$T - 1 \le \left(1 + \frac{2Lr}{\gamma}\right)^r. \tag{30}$$

Hence bounded L and a uniformly positive γ cannot persist as $T \to \infty$ at fixed r. This does not prevent an asymptotic separation: at fixed r, d, C, D, κ and bounded B, the sufficient condition $L(T)/\gamma(T) = o(T)$ already makes the sink upper bound smaller than the diagonal lower bound for all sufficiently large admissible lengths.

Relation to finite-temperature softmax. The theorem uses hard routing with a margin, as in the main text. For ordinary softmax, a dominant logit with gap at least κ has weight at least $(1+Te^{-\kappa})^{-1}$. Thus a uniform attention-mass tolerance ε can be enforced by taking $\kappa \ge \log(T/\varepsilon)$; the displayed bounds retain their explicit dependence on this choice. Exact output equality is asserted for the hard-attention model.

D.1.1 An explicit family with fixed embedding dimension

The next construction shows that the context-length comparison is nonvacuous at fixed dimension. It includes a monotone clock so that both attention patterns are realizable.

Theorem D.2 (A fixed-dimensional asymptotic separation)**.** *Fix an integer $k \ge 3$, nonempty token sets $\mathcal{C}, \mathcal{D}$, a margin $\kappa > 0$, and an ambient dimension $d \ge 2k + C + D + 3$. Consider the block above with hidden width $w \ge 2k + D + 3$. For every integer $q \ge 3$, there is a backcopy distribution with $T = q^k$ and positional rank $r = 2k + 2$ satisfying the preceding theorem. If $\mathcal{R}^*_{\mathrm{sink}}$ and $\mathcal{R}^*_{\mathrm{diag}}$ denote the infimal costs in the two admissible routing classes, then, as $q \to \infty$,*

$$\mathcal{R}^*_{\mathrm{sink}} = O(T^{2/k}), \qquad \mathcal{R}^*_{\mathrm{diag}} = \Theta(T).$$

In particular, a sink is strictly cheaper for all sufficiently large q.

Proof. Set $f_\ell = q^\ell$ for $\ell = 0, \dots, k-1$ and define

$$u_t = \frac{1}{\sqrt{k}}\big(\cos(2\pi f_0 t/T), \sin(2\pi f_0 t/T), \dots, \cos(2\pi f_{k-1} t/T), \sin(2\pi f_{k-1} t/T)\big). \tag{31}$$

These are unit vectors. The frequencies are distinct and belong to $(0, T/2)$, so discrete Fourier orthogonality gives $U^\top U = (T/(2k))I_{2k}$ for the matrix with rows $u_t^\top$. Let R_h rotate each harmonic pair backward by its one-step angle. Then $R_h u_t = u_{t-1}$ and $\|R_h\|_* = 2k$.

For any nonzero cyclic difference $a \pmod T$, write $a = q^j b$ with $q \nmid b$. The frequency f_{k-j-1} produces phase $2\pi b/q$. Therefore

$$\frac{1}{k}\sum_{\ell=0}^{k-1}\cos(2\pi f_\ell a/T) \le 1 - \gamma_T, \qquad \gamma_T = \frac{1-\cos(2\pi/q)}{k} = \Theta(T^{-2/k}). \tag{32}$$

The value $a = q^{k-1}$ attains this bound. The predecessor has cyclic difference zero; all its competing causal positions have a nonzero cyclic difference. Thus (32) supplies the required predecessor gap.

Choose $\rho^2 = 1/(2k+1)$ and append a two-dimensional clock:

$$p_t = \sqrt{1-\rho^2}\, u_t \oplus \rho\big(\sin(t/T), \cos(t/T)\big), \qquad R = (1-\rho^2)^{-1} R_h \oplus 0. \tag{33}$$

All p_t have unit norm. Their rank is $2k+2$: their real sine and cosine columns are an invertible change of coordinates of complex exponential columns with distinct bases $e^{\pm 2\pi \mathrm{i} f_\ell/T}$ and $e^{\pm \mathrm{i}/T}$. The first $2k+2$ rows form a nonsingular Vandermonde matrix, up to nonzero column factors. Also,

$$\|P\|_{\mathrm{op}}^2 \le (1-\rho^2)\frac{T}{2k} + \rho^2 T = \frac{2T}{2k+1}, \qquad \|R\|_* = \frac{2k}{1-\rho^2} = 2k+1 < r.$$

Thus the frame and shift conditions hold with $B = 2r/(2k+1)$, $L = 1$, and $\gamma = \gamma_T$. The condition at position 1 follows from the fact that every harmonic inner product is at most one. The sink upper bound is consequently $O(\gamma_T^{-1}) = O(T^{2/k})$, and the diagonal lower bound is $\Omega(T)$.

For completeness, we construct a diagonal representation with cost $O(T)$. Let e be the first clock-coordinate vector and set

$$\xi = p_t^\top R p_t = \frac{1}{k}\sum_{j=1}^{k} \cos(2\pi/q^j), \qquad \delta = 1 - \xi.$$

These quantities are independent of t, and $\delta/\gamma_T \le k$. Define

$$\beta = \frac{2\kappa}{\sqrt{d}\,\gamma_T}, \qquad \eta = \frac{T}{\rho\cos(1)}\left(\beta\delta + \frac{2\kappa}{\sqrt{d}}\right), \qquad M_{\mathrm{diag}} = \beta R + \eta e h_{\mathcal{D}}^\top. \tag{34}$$

For a copy-paste query, the clock term vanishes, so its predecessor routing is unchanged. For a dormant query at t, the difference between the current-key score and any earlier ordinary-key score satisfies

$$S_{tt} - S_{ts} \ge \frac{\sqrt{d}}{2}\left[-\beta\delta + \eta\rho\big(\sin(t/T) - \sin(s/T)\big)\right] \ge \kappa,$$

because $\sin(t/T) - \sin(s/T) \ge \cos(1)(t-s)/T$ for $1 \le s < t \le T$. The BOS score is zero, while

$$S_{tt} \ge \frac{\sqrt{d}}{2}\left[\beta\xi + \frac{\eta\rho\cos(1)}{T}\right] = \frac{\sqrt{d}}{2}\left(\beta + \frac{2\kappa}{\sqrt{d}}\right) \ge \kappa.$$

Thus all dormant queries select themselves. Taking $W_V = W_O = \Pi$ as above and setting the MLP identically to zero gives exactly the target mapping for diagonal routing, including at BOS. Finally, $\beta = O(T^{2/k})$ and $\beta\delta = O(1)$, so $\eta = O(T)$ and

$$\mathcal{R}(\Theta) \le 2\beta\|R\|_* + 2\eta\sqrt{D} + 2(r + C + D) = O(T).$$

Together with the lower bound, this proves $\mathcal{R}^*_{\mathrm{diag}} = \Theta(T)$. Since $2/k < 1$, the sink upper bound is eventually strictly smaller. □

The asymptotic statement is along the sequence lengths $T = q^k$, with k fixed and $q \to \infty$. The ambient dimension and token directions are fixed in this family; the positional embeddings and their shared shift depend on T. Thus the result concerns a family of finite-length backcopy distributions, rather than prefixes of one fixed positional-embedding sequence. Taking $k = 3$ gives the rank-eight example stated in the main text.

Uniform dependence on the harmonic count. The implicit constant in the preceding fixed-k bound is not uniform when k grows. We record the dependence explicitly. The elementary inequality $1 - \cos(2\pi/q) \ge 8/q^2$ for $q \ge 3$ gives $\gamma_T^{-1} \le kq^2/8$. Choosing the minimal ambient dimension $d = 2k + C + D + 3$ and recalling $r = 2k+2$, $L = 1$ and fixed C, D, (15) yields, for a constant $K_{C,D}$ independent of k, q, T, κ,

$$\mathcal{R}^*_{\mathrm{sink}} \le K_{C,D}\big(k + \kappa k^{3/2} q^2\big) = K_{C,D}\big(k + \kappa k^{3/2} T^{2/k}\big), \qquad T = q^k, \quad q \ge 3, \quad k \ge 3. \tag{35}$$

Taking $q = 3$ and $k \to \infty$ gives $d, r = \Theta(\log T)$ and, for fixed κ, $\mathcal{R}^*_{\text{sink}} = O((\log T)^{3/2})$. The diagonal lower bound remains $\Omega(T)$, while the diagonal construction in the proof above has cost $O(T)$ uniformly in k: indeed, $\beta = O(\kappa k^{1/2} q^2)$ in its worst-case form, but in the diagonal witness $\beta\delta = O(\kappa\sqrt{k})$ would be an overly crude estimate. More precisely, using the expression for ξ and δ in that construction, one has

$$\delta = \frac{1}{k}\sum_{j=1}^{k}\bigl(1 - \cos(2\pi/q^j)\bigr) \le \frac{2\pi^2}{k(q^2-1)},$$

so $\beta\delta = O(\kappa/\sqrt{k})$ for $d = \Theta(k)$. Consequently $\eta = O(\kappa T)$ and the diagonal cost is $O(k + \kappa T)$ for this choice of d. This already gives a subpolynomial sink upper bound. The following construction improves it to logarithmic cost and works for every sufficiently large T.

D.1.2 A logarithmic-cost family with logarithmic dimension

Theorem D.3 (Logarithmic sink cost versus linear diagonal cost)**.** *Fix nonempty token sets $\mathcal{C}, \mathcal{D}$ and a hard-attention margin $\kappa > 0$. For every sufficiently large integer T, there exist unit positional embeddings of rank $r = \Theta(\log T)$ in an ambient dimension $d = \Theta(\log T)$, satisfying (8)–(9) with $B \le 3$, $L = 1$, and $\gamma = 1/2$. With hidden width $w \ge r + D + 1$, the optimum costs among admissible sink and diagonal representations satisfy*

$$\mathcal{R}^*_{\text{sink}} = O(\log T), \qquad \mathcal{R}^*_{\text{diag}} = \Theta(T),$$

where the implicit constants may depend on C, D, κ but not on T. In particular, sinks are strictly cheaper for all sufficiently large T.

Proof. For an integer T, put $N = \lfloor (T-1)/2 \rfloor$ and $k = \lceil 32 \log T \rceil$ (natural logarithm). For large enough T, $k \le N$. We first show that there exist distinct frequencies $f_1, \dots, f_k \in \{1, \dots, N\}$ such that

$$\frac{1}{k}\sum_{j=1}^{k} \cos\left(\frac{2\pi f_j a}{T}\right) \le \frac{1}{2} \qquad \text{for all } a \in \{1, \dots, T-1\}. \tag{36}$$

Sample the frequencies independently and uniformly from $\{1, \dots, N\}$. For every nonzero residue $a \pmod T$, roots-of-unity summation gives

$$\mathbb{E}\cos\left(\frac{2\pi f_j a}{T}\right) = \begin{cases} -1/(T-1), & T \text{ odd}, \\ -(1+(-1)^a)/(T-2), & T \text{ even}, \end{cases} \le 0.$$

Since every summand lies in $[-1, 1]$, Hoeffding's inequality implies

$$\mathbb{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \cos\left(\frac{2\pi f_j a}{T}\right) > \frac{1}{2}\right\} \le e^{-k/8}.$$

A union bound over the $T-1$ nonzero offsets gives failure probability at most $(T-1)e^{-k/8} \le T^{-3}$. The probability of any repeated frequency is at most $k(k-1)/(2N) = O((\log T)^2/T)$. Thus, for sufficiently large T, there is a distinct-frequency realization satisfying (36). Fix such a realization.

Define the $2k$-dimensional harmonic vectors

$$u_t = \frac{1}{\sqrt{k}}\bigl(\cos(2\pi f_1 t/T), \sin(2\pi f_1 t/T), \dots, \cos(2\pi f_k t/T), \sin(2\pi f_k t/T)\bigr), \quad 1 \le t \le T. \tag{37}$$

They are unit vectors and, because the frequencies are distinct and none is a Nyquist frequency, their row matrix U satisfies $U^\top U = (T/(2k))I_{2k}$. Let R_h be the orthogonal, block-diagonal rotation for which $R_h u_t = u_{t-1}$. Then $u_{t-1}^\top R_h u_t = 1$ and $u_s^\top R_h u_t \le 1/2$ whenever $s \ne t - 1$ modulo T.

Set $\rho^2 = 1/(2k+1)$ and add the same two-dimensional monotone clock:

$$p_t = \sqrt{1-\rho^2}\, u_t \oplus \rho\bigl(\sin(t/T), \cos(t/T)\bigr), \qquad R = (1-\rho^2)^{-1} R_h \oplus 0. \tag{38}$$

All p_t have unit norm. Their span has rank $r = 2k + 2$: complexifying the harmonic and clock coordinates produces the $2k + 2$ distinct exponential bases $e^{\pm 2\pi \mathrm{i} f_j/T}$ and $e^{\pm \mathrm{i}/T}$. The first $2k + 2$

samples form a nonsingular Vandermonde system, so the real coordinate functions have full column rank. The predecessor score is exactly one, and all competing ordinary positions have score at most $1/2$. Also,

$$\|P\|_{\mathrm{op}}^2 \le (1-\rho^2)\frac{T}{2k} + \rho^2 T = \frac{2T}{2k+1}, \qquad \|R\|_* = \frac{2k}{1-\rho^2} = 2k+1 < r.$$

Thus $B = 2r/(2k+1) < 3$, $L = 1$ and $\gamma = 1/2$ work in (8)–(9); $p_1^\top R p_1 \le 1$ follows from the harmonic cosine bound. Choose $d = r + C + D + 1$ and embed the token directions orthogonally to the positional subspace. The finite-length theorem gives

$$\mathcal{R}^*_{\mathrm{sink}} \le O\left(r + \frac{\kappa r}{\sqrt{d}}\right) = O(\log T + \kappa\sqrt{\log T}),$$

which is $O(\log T)$ at fixed κ. Since $r/d = \Theta(1)$ and B is bounded, the same theorem gives $\mathcal{R}^*_{\mathrm{diag}} = \Omega(\kappa T)$.

We verify that the diagonal class is realizable with cost $O(T)$. Let e be the first unit coordinate vector of the clock subspace, let $h_{\mathcal{D}} = \sum_{b\in\mathcal{D}} h_b$, and write

$$\xi = p_t^\top R p_t = \frac{1}{k}\sum_{j=1}^{k} \cos\left(\frac{2\pi f_j}{T}\right), \qquad \delta = 1 - \xi \in [0, 2].$$

Both quantities are independent of t. Set

$$\beta = \frac{2\kappa}{\sqrt{d}\,\gamma}, \qquad \eta = \frac{T}{\rho\cos(1)}\left(\beta\delta + \frac{2\kappa}{\sqrt{d}}\right), \qquad M_{\mathrm{diag}} = \beta R + \eta e h_{\mathcal{D}}^\top. \tag{39}$$

For copy-paste queries, the second term vanishes, so the predecessor remains preferred by margin at least κ. For a dormant query, the clock contributes $\eta\rho\sin(s/T)$ to the key at position s. Since, for $1 \le s < t \le T$, $\sin(t/T) - \sin(s/T) \ge \cos(1)/T$ and $p_s^\top R p_t \le 1$, the current-key advantage over earlier ordinary keys is at least

$$\frac{\sqrt{d}}{2}\left[-\beta\delta + \eta\rho\frac{\cos(1)}{T}\right] = \kappa.$$

Moreover, $\sin(t/T) \ge \cos(1)/T$ for every $t \ge 1$, so the current-key logit exceeds the BOS logit zero by at least

$$\frac{\sqrt{d}}{2}\left[\beta\xi + \eta\rho\frac{\cos(1)}{T}\right] = \frac{\sqrt{d}}{2}\left[\beta + \frac{2\kappa}{\sqrt{d}}\right] \ge \kappa.$$

Thus this product enforces the diagonal routing. With $W_V = W_O = \Pi$ as in the sink proof and zero MLP, every dormant token receives its own input and every copy-paste token receives its predecessor. BOS remains unchanged; all targets are exact in the hard-attention model.

Finally, $\|R\|_* = O(k)$, $\beta = O(\kappa/\sqrt{k})$, $\rho^{-1} = O(\sqrt{k})$ and $\delta \le 2$. Therefore $\eta = O(\kappa T)$ and

$$\mathcal{R}^*_{\mathrm{diag}} \le 2\|M_{\mathrm{diag}}\|_* + 2(r+C+D) \le 2\beta\|R\|_* + 2\eta\sqrt{D} + 2(r+C+D) = O(\kappa T + \log T) = O(T)$$

for fixed κ. Together with the lower bound this proves $\mathcal{R}^*_{\mathrm{diag}} = \Theta(T)$. □

Why the dimension must grow for subpolynomial cost. The fact that $r, d = \Theta(\log T)$ above is not merely an artifact of the harmonic construction. With fixed r, d and κ, every admissible representation (sink or diagonal) in this data model satisfies the packing lower bound

$$\mathcal{R}(\Theta) \ge \frac{\sqrt{2}\,\kappa}{\sqrt{d}}\big((T-1)^{1/r} - 1\big). \tag{40}$$

To see this, fix any $1 \le i < j \le T-1$. The support contains an input sequence with the same dormant token at positions i, j and a copy-paste token at $j+1$. Because the latter must select its immediate predecessor, the attention margin gives

$$\kappa \le \frac{1}{\sqrt{d}}(x_{j,b} - x_{i,b})^\top M x_{j+1,a} \le \sqrt{\frac{d}{2}}\,\|M\|_{\mathrm{op}}\,\|p_j - p_i\|_2.$$

Using $\mathcal{R}(\Theta) \geq 2\|M\|_*$ and $\|M\|_{\mathrm{op}} \leq \|M\|_*$, we obtain $\|p_j - p_i\|_2 \geq 2\sqrt{2}\kappa/(\sqrt{d}\,\mathcal{R}(\Theta))$. Packing $T-1$ such unit vectors into an r-dimensional unit ball, with pairwise separation Δ, gives $T - 1 \leq (1 + 2/\Delta)^r$, which rearranges to (40). In particular, a single fixed-dimensional family cannot have $T^{o(1)}$ cost at fixed κ under these routing and support assumptions. This does not contradict the fixed-k result: for every prescribed $\varepsilon > 0$ one can first choose a sufficiently large fixed $k \geq \max\{3, 2/\varepsilon\}$ and corresponding fixed d, and then let $T = q^k \to \infty$, obtaining a sink cost $O(T^\varepsilon)$ along those lengths.

D.2 Sink-diagonal separation for the generic clustered dormant group distribution

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:*

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

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 $|G_{ij}| \leq \phi$ for $i \neq j$, Gershgorin's circle theorem gives

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

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, $\|\Pi\|_2^2 \leq \gamma$, and $\|U\|_2 \leq \sqrt{1 + 2C\phi}$. Recall $r_\eta = \dim \mathrm{span}\{\eta_{ci} : \text{all } c, i\}$. 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

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

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

We assume

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

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

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

Now, Theorem 7.2 is implied by the following result.

Theorem D.4. *Under the setup above, the following four bounds hold.*

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

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

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

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

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

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

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

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

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,

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

Therefore

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

The variational form of nuclear norm then gives

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

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

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

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 $T_{\text{base}}c_c = 0$, 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 $T_{\text{base}}c_c = 0$ and $T_{\text{base}}\mu = \mu$, 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

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

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

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

and

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

Adding the two inequalities gives

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

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

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

Thus

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

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

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

with

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

Note that

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

implying that

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

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:

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

The variational form of nuclear norm then gives

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

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

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

Thus, we have the desired lower bound:

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

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

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

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

where

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

with

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

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 $\Pi^\top M_{\mathrm{diag}}\Pi$ gives

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

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

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

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

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

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

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

For $c' \ne c$, 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 $d_{c'i}$,

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

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

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

Now

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

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

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

The variational form of nuclear norm then gives

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

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

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

Thus

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

For the value-output map, use

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

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

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

Set the MLP to zero. Therefore

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

4. Sink lower bound. 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:

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

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

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

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

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

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,

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

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,

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

By nuclear/operator norm duality,

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

Using the operator-norm bound above, we obtain

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

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

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

Therefore

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

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

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

Therefore

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

which completes the proof. □

E Proof of recovery of Theorem 1 from [31] from Theorem C.1

Here we give a full proof of how Theorem 1 from [31] can be recovered from Theorem C.1.

Setting. Fix integers $n \geq 5$ and $L \geq 4$. Let P be a probability measure on $\mathbb{R}$ with a bounded density, and let $\mathcal{C} := \operatorname{span}\{e_4, \ldots, e_n\}$ and $\mathcal{A} := \operatorname{span}\{e_2, e_3\} \oplus \mathcal{C}$, where e_k denotes the k-th standard basis vector of $\mathbb{R}^n$. Consider the following distribution of sequences $x = (x^{(1)}, \ldots, x^{(L)})$: choose a trigger position $j \in \{2, \ldots, L\}$ uniformly, set $x^{(1)} = e_1$, and independently sample $u^{(i)} \in \mathcal{C}$, $i \geq 2$, with each of its content coordinates drawn from P. Set

$$x^{(i)} = \begin{cases} e_2 + u^{(i)}, & i = j, \\ e_3 + u^{(i)}, & i \in \{2, \ldots, L\} \setminus \{j\}. \end{cases} \tag{41}$$

The target is

$$y^{(i)} = \begin{cases} \dfrac{1}{j-1} \displaystyle\sum_{k=2}^{j} x^{(k)}, & i = j, \\ 0, & i \neq j. \end{cases} \tag{42}$$

Let f be a single-layer causal attention model with combined value–output map $V = W_O W_V \in \mathbb{R}^{n \times n}$, namely,

$$f(x)^{(i)} = V \sum_{k=1}^{i} \alpha_{i,k}(x) x^{(k)}, \qquad \alpha_{i,k}(x) \geq 0, \qquad \sum_{k=1}^{i} \alpha_{i,k}(x) = 1. \tag{43}$$

In particular, the usual causal softmax attention satisfies these conditions. Define the worst-case Euclidean error

$$\mathcal{L}(f) := \sup_{(x,y) \in \operatorname{supp}(\mathcal{D})} \max_{1 \leq i \leq L} \| f(x)^{(i)} - y^{(i)} \|_2, \tag{44}$$

where $\mathcal{D}$ is the joint input–target distribution above.

Theorem E.1 (Single-layer sink necessity for the trigger-conditional task)**.** *There are constants $\eta_0 > 0$ and $C_P > 0$, depending only on n and P, such that every model satisfying $\mathcal{L}(f) \leq \eta \leq \eta_0$ obeys*

$$\alpha_{i,1}(x) \geq 1 - C_P \eta \tag{45}$$

for every $(x, y) \in \operatorname{supp}(\mathcal{D})$ *and* every *non-trigger position $i \neq j$. Consequently, for any $\epsilon, \delta > 0$ there exists $\eta_\epsilon > 0$ such that $\mathcal{L}(f) \leq \eta_\epsilon$ implies*

$$\mathbb{P}_{(x,y) \sim \mathcal{D}}(\alpha_{i,1}(x) \geq 1 - \epsilon \text{ for all } i \neq j) \geq 1 - \delta. \tag{46}$$

In the exact case $\mathcal{L}(f) = 0$, necessarily $\alpha_{i,1}(x) = 1$ at every non-trigger position. [2]

Proof. We construct a finite set of trigger examples witnessing the same input–output rank condition as in the hard-switch-to-sink direction of Theorem C.1, and then derive its quantitative counterpart.

Step 1: A full-rank set of target outputs. Since P admits a probability density, its support contains at least two distinct real numbers $a \neq b$. Define

$$u^0 := a \sum_{m=4}^{n} e_m, \qquad u^m := u^0 + (b - a) e_m \quad (m = 4, \ldots, n). \tag{47}$$

Independence of the content coordinates implies that all these vectors belong to the support of the content law. For each $u \in \{u^0, u^4, \ldots, u^n\}$, choose an admissible sequence whose trigger is at $j = 2$ and whose second token is $e_2 + u$. Its target at position 2 is $e_2 + u$. Choose one more admissible sequence with trigger at $j = 3$, second token $e_3 + u^0$, and third token $e_2 + u^0$. Its target at position 3 is $\frac{1}{2}(e_2 + e_3) + u^0$. (All the displayed prefixes can be extended to length L using ordinary tokens with content u^0.)

Put $r := n - 1 = \dim \mathcal{A}$ and form the matrix of these r target columns,

$$Y := \left[e_2 + u^0,\ e_2 + u^4, \ldots, e_2 + u^n,\ \tfrac{1}{2}(e_2 + e_3) + u^0 \right] \in \mathbb{R}^{n \times r}. \tag{48}$$

[2] No restriction of the form $L < n$ and no global-optimality assumption are required.

For $m = 4, \ldots, n$, subtracting the first column from the column indexed by m gives $(b-a)e_m$; hence the columns of Y span $\mathcal{C}$. Subtracting $u^0 \in \mathcal{C}$ from the first column yields e_2, and the last column then provides e_3. Thus $\mathrm{col}(Y) = \mathcal{A}$ and

$$s := \sigma_r(Y) > 0, \qquad R := \max\left\{ \|e_2 + u^0\|_2,\ \|e_3 + u^0\|_2,\ \max_{4 \le m \le n} \|e_2 + u^m\|_2 \right\} < \infty. \tag{49}$$

In particular, all non-BOS inputs span $\mathcal{A}$, and the trigger targets span $\mathcal{A}$ as well. For an exact realizer, this verifies the input–output rank equality in Theorem C.1.

Step 2: Quantitative injectivity of V on $\mathcal{A}$. Assume $\mathcal{L}(f) \le \eta$. At the BOS position, causality gives $\alpha_{1,1} = 1$ and the target is zero; hence

$$\|Ve_1\|_2 = \|f(x)^{(1)}\|_2 \le \eta. \tag{50}$$

Index the witness sequences from Step 1 by $\ell = 1, \ldots, r$; let $j_\ell \in \{2, 3\}$ be their respective trigger positions and let y_ℓ be the corresponding columns of Y. For each witness, set

$$z_\ell := \sum_{k=2}^{j_\ell} \alpha_{j_\ell,k}(x_\ell) x_\ell^{(k)} \in \mathcal{A}, \qquad Z := [z_1, \ldots, z_r]. \tag{51}$$

Because the weights in this sum are nonnegative and add up to at most one, the definition of R gives

$$\|Z\|_{\mathrm{op}} \le \|Z\|_F \le R\sqrt{r}. \tag{52}$$

The model output at the trigger position of witness ℓ is $\alpha_{j_\ell,1}(x_\ell)Ve_1 + Vz_\ell$. By the loss bound and (50),

$$\|Vz_\ell - y_\ell\|_2 \le \eta + \alpha_{j_\ell,1}(x_\ell)\|Ve_1\|_2 \le 2\eta, \qquad \|VZ - Y\|_{\mathrm{op}} \le 2\sqrt{r}\,\eta. \tag{53}$$

If $\eta \le \eta_0 := s/(4\sqrt{r})$, the singular-value perturbation inequality yields

$$\sigma_r(VZ) \ge \sigma_r(Y) - \|VZ - Y\|_{\mathrm{op}} \ge s/2 > 0. \tag{54}$$

It follows that Z has rank r. Since its columns lie in the r-dimensional subspace $\mathcal{A}$, Z is a linear bijection from $\mathbb{R}^r$ onto $\mathcal{A}$. For any $u \in \mathcal{A}$, write $u = Zw$ and use (52)–(54) to obtain

$$\|Vu\|_2 = \|VZw\|_2 \ge \sigma_r(VZ)\|w\|_2 \ge \frac{s}{2R\sqrt{r}}\|u\|_2. \tag{55}$$

Thus V is quantitatively injective on $\mathcal{A}$ with lower bound $\mu := s/(2R\sqrt{r}) > 0$.

Step 3: The strict-half-space margin forces the sink. Set $h := e_2 + e_3$. By (41), every non-BOS token, whether ordinary or a trigger, satisfies

$$\langle h, x^{(k)} \rangle = 1 \qquad (k \ge 2), \qquad \|h\|_2 = \sqrt{2}. \tag{56}$$

Fix any admissible sequence and any non-trigger position $i \ge 2$. Define the non-BOS attention mixture

$$z_i := \sum_{k=2}^{i} \alpha_{i,k}(x) x^{(k)} \in \mathcal{A}. \tag{57}$$

Using (56) and normalization of attention weights gives

$$\langle h, z_i \rangle = \sum_{k=2}^{i} \alpha_{i,k}(x) = 1 - \alpha_{i,1}(x), \qquad \|z_i\|_2 \ge \frac{1 - \alpha_{i,1}(x)}{\sqrt{2}}. \tag{58}$$

Since the target at i is zero, the worst-case loss bound and (50) imply

$$\|Vz_i\|_2 \le \|f(x)^{(i)}\|_2 + \alpha_{i,1}(x)\|Ve_1\|_2 \le 2\eta. \tag{59}$$

Combining (55), (58), and (59), we obtain

$$\frac{s}{2R\sqrt{2r}}\big(1 - \alpha_{i,1}(x)\big) \le 2\eta, \qquad \text{so} \qquad 1 - \alpha_{i,1}(x) \le \underbrace{\frac{4R\sqrt{2r}}{s}}_{=:C_P}\,\eta. \tag{60}$$

For $i = 1$ the claim is automatic, since $\alpha_{1,1} = 1$. All the estimates hold for every sequence in the support, independently of its trigger position, establishing (45).

Finally, for any $\epsilon, \delta > 0$, take

$$\eta_\epsilon := \min\left\{ \frac{s}{4\sqrt{r}}, \frac{\epsilon s}{4R\sqrt{2r}} \right\} > 0. \tag{61}$$

Then $\mathcal{L}(f) \le \eta_\epsilon$ implies $\alpha_{i,1}(x) \ge 1 - \epsilon$ simultaneously for all non-trigger positions and all admissible sequences, so the event in (46) has probability one (hence at least $1 - \delta$). Setting $\eta = 0$ in (60) gives the exact implication. □

F Extension of Section 7 to Relative Positional Embeddings

Although our theory is stated for absolute positional embeddings, a similar geometric interpretation can be developed for RoPE-style relative positional encodings. Standard RoPE applies position-dependent orthogonal rotations to queries and keys, namely $q_i = \Phi_i W_Q x_i$ and $k_j = \Phi_j W_K x_j$, so that

$$S_{ij} = \frac{1}{\sqrt{d}} x_i^\top W_Q^\top \Phi_i^\top \Phi_j W_K x_j.$$

Thus, even for a fixed context length, the rotations Φ_i cannot generally be replaced by a single common rotation.

The appropriate interpretation is instead to impose the geometric assumptions on position-indexed embeddings. More precisely, for a context embedding x_a occurring at position i, define

$$z_{a,i} = \Phi_i x_a, \qquad z_{a,i}^\top z_{b,j} = x_a^\top \Phi_i^\top \Phi_j x_b.$$

Hence, the similarity between two occurrences depends on both their context embeddings and their positions. For a fixed context length T, the collection of rotations $\{\Phi_0, \dots, \Phi_T\}$ is fixed and finite, allowing us to consider a fixed position-indexed geometry.

The absorption can be performed separately at each position. In particular, defining $\widetilde{W}_{Q,i} = \Phi_i W_Q \Phi_i^\top$ and $\widetilde{W}_{K,i} = \Phi_i W_K \Phi_i^\top$, we obtain $q_i = \widetilde{W}_{Q,i} z_{a,i}$ and $k_i = \widetilde{W}_{K,i} z_{a,i}$. Since Φ_i is orthogonal, $\|\widetilde{W}_{Q,i}\|_F = \|W_Q\|_F$ and $\|\widetilde{W}_{K,i}\|_F = \|W_K\|_F$. Importantly, however, these transformed matrices depend on position and cannot be selected independently, as they must all originate from the same shared matrices W_Q and W_K.

Under this interpretation, dormant groups should be defined in the position-indexed query-key geometry. In particular, the same residual-stream embedding may belong to different dormant groups when occurring at sufficiently different positions, while two different residual-stream embeddings may belong to the same dormant group when their position-indexed representatives are sufficiently close. This is analogous to the absolute-positional setting, where the input to the attention layer already depends on both token identity and position.

Importantly, the dormant-group geometry used by the query-key mechanism need not coincide with the geometry of the residual stream. The two parts of the proof of Theorem 7.2 use these geometric assumptions for different purposes.

Sink upper bound. The query-key part determines the attention pattern. For the sink upper bound, the problem becomes a finite collection of margin constraints over token-position pairs: dormant queries must select BOS, while copy-paste queries must select the preceding dormant embeddings belonging to the correct position-indexed group. Analogues of the separation and incoherence assumptions in Theorem 7.2 can then be formulated for the position-indexed geometry.

However, obtaining the corresponding low-cost sink construction additionally requires that all these constraints can be realized by a single shared pair W_Q, W_K. Indeed, the transformed matrices must satisfy

$$\Phi_i^\top \widetilde{W}_{Q,i} \Phi_i = W_Q, \qquad \Phi_i^\top \widetilde{W}_{K,i} \Phi_i = W_K$$

for every position i. Consequently, a collection of individually low-cost query-key maps in the position-indexed geometry does not necessarily correspond to a valid RoPE attention head. The cost

of satisfying this shared-parameter restriction may depend on the conditioning of the position-indexed geometry and the context length, and is not automatically controlled by the original geometric assumptions.

The value-output and MLP parts, on the other hand, do not need to distinguish the individual dormant groups. They only require a low-cost linear map T_{base} satisfying $T_{\mathrm{base}}d = d$ for every dormant residual embedding d, together with $T_{\mathrm{base}}b_{\mathrm{BOS}} = 0$ and $T_{\mathrm{base}}c = 0$ for every copy-paste residual embedding c.

More generally, the residual-side complexity can be expressed as

$$\tau_{\mathrm{res}} = \inf_{T} \big\{ \|T\|_* : Td = d \text{ for every dormant } d, \\ Tb_{\mathrm{BOS}} = 0,\ Tc = 0 \text{ for every copy-paste } c \big\}.$$

Using the same value-output and MLP construction as in the original proof gives a residual-side cost of at most $(3 + 2\epsilon)\tau_{\mathrm{res}}$. Under the original residual-stream assumptions of Theorem 7.2, we have $\tau_{\mathrm{res}} \le r_\eta + C\beta$, recovering the existing cost bound $5(r_\eta + C)$. Thus, changing the dormant groups in the position-indexed query-key geometry does not require the residual embeddings to form the same groups.

Diagonal lower bound. The diagonal lower bound depends even less on the residual geometry, since it uses only attention-score constraints. For example, suppose two dormant embeddings x_a, x_b occur at positions $p < q$ in both relative orders. Diagonal attention requires

$$x_b^\top W_Q^\top W_K x_b - x_b^\top W_Q^\top \Phi_q^\top \Phi_p W_K x_a \ge \kappa\sqrt{d}, \\ x_a^\top W_Q^\top W_K x_a - x_a^\top W_Q^\top \Phi_q^\top \Phi_p W_K x_b \ge \kappa\sqrt{d}.$$

These are the RoPE analogues of the inequalities used in the original proof. They can be aggregated over admissible position-indexed pairs to obtain a generalized witness replacing Σ_D. This witness measures the complexity of the query-key separation required to make every dormant occurrence attend to itself.

Unlike the original construction, the generalized witness does not necessarily reduce to a covariance matrix of embedding differences. In particular, reversing the order of two embeddings also changes their position-indexed representatives. Nevertheless, when the position dependence within a dormant group is sufficiently small, the resulting witness approaches the original construction up to perturbation terms, suggesting a corresponding effective-rank lower-bound mechanism.

Discussion and limitations. Overall, the comparison retains the same conceptual structure: the sink upper bound consists of a query-key routing cost and a residual-side cost, while the diagonal lower bound depends entirely on query-key score constraints. Schematically, we may write

$$U_{\mathrm{sink}} \le U_{\mathrm{QK,sink}} + (3 + 2\epsilon)\tau_{\mathrm{res}}, \\ L_{\mathrm{diag}} \ge L_{\mathrm{QK,diag}},$$

where the query-key terms must be defined using the RoPE-compatible position-indexed score constraints.

However, this formulation does not establish an exact equivalence with Theorem 7.2 for arbitrary RoPE. A fully rigorous extension requires controlling the cost and feasibility of the sink construction under the shared W_Q, W_K restriction, as well as deriving the precise position-indexed analogue of the effective-rank lower bound. In particular, fixed context length makes the problem finite but does not guarantee that the additional cost is bounded by a universal constant. Establishing these geometry-dependent bounds and precisely characterizing the resulting sink-versus-diagonal cost comparison is an interesting direction for future work.

G Further experimental evidence

G.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

$$X + s_l \,\mathrm{ATTN}_l^h(Z^{(l)}) \text{ , where } s_l = \frac{\frac{1}{T}\sum_{i=1}^{T}\left\|\mathrm{ATTN}_l(Z_i^{(l)})\right\|}{\frac{1}{n_l}\sum_{h=1}^{n_l}\frac{1}{T}\sum_{i=1}^{T}\left\|\mathrm{ATTN}_l^h(Z_i^{(l)})\right\|}. \tag{62}$$

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

$$u(A) = \frac{1}{T}(1 + \sum_{i=2}^{T}(-\sum_{j=1}^{i} A_{ij}\log A_{ij})/\log i) \in [0,1], \tag{63}$$

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 [25]) and code (CodeParrot[3]). 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.

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

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

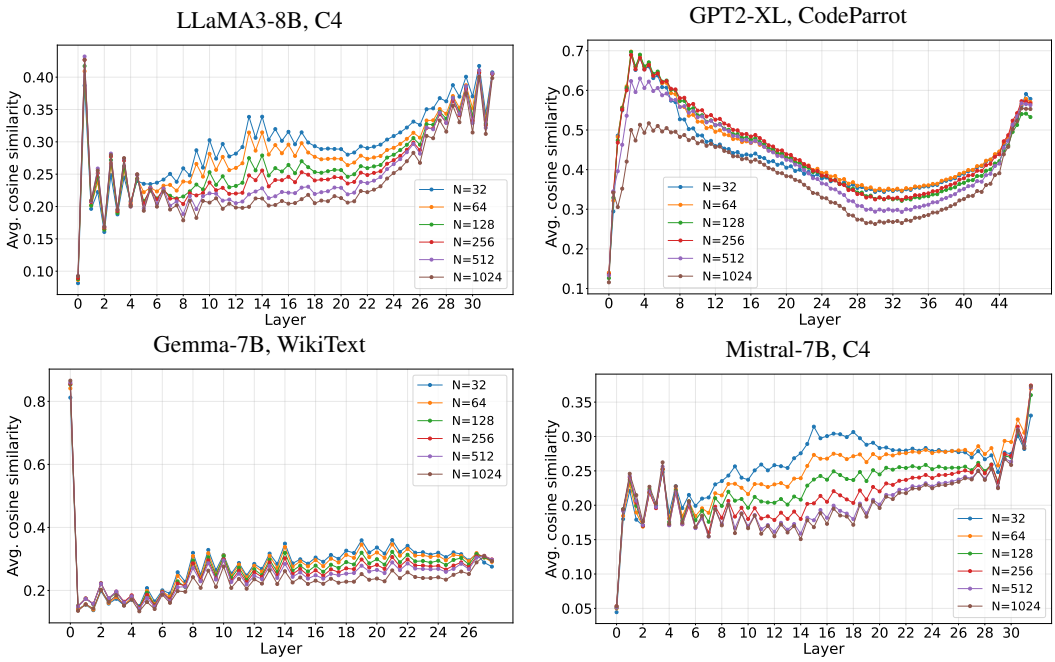

Figure 4: Evolution of average cosine similarity across layers for different models, datasets, and context lengths. For each layer, the first value corresponds to the token matrix before the layer, and the second to the token matrix after adding the attention update to the residual stream. In all cases,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

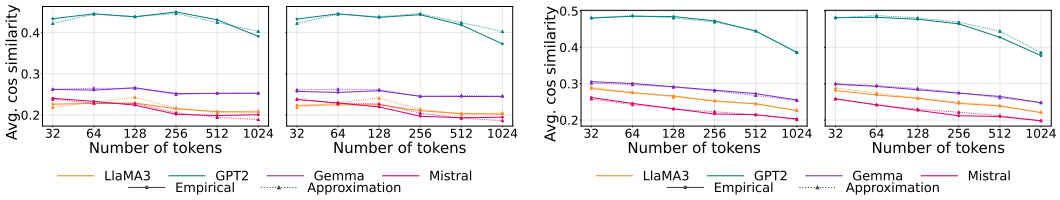

Figure 5: Empirical average cosine similarity (solid) and theoretical approximation (2) (dashed) for X (first plot) and $X + WZ$ (second plot), across models and context lengths, on the CodeParrot dataset (left) and WikiText (right).

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 $\text{avg. cossim}(X +$

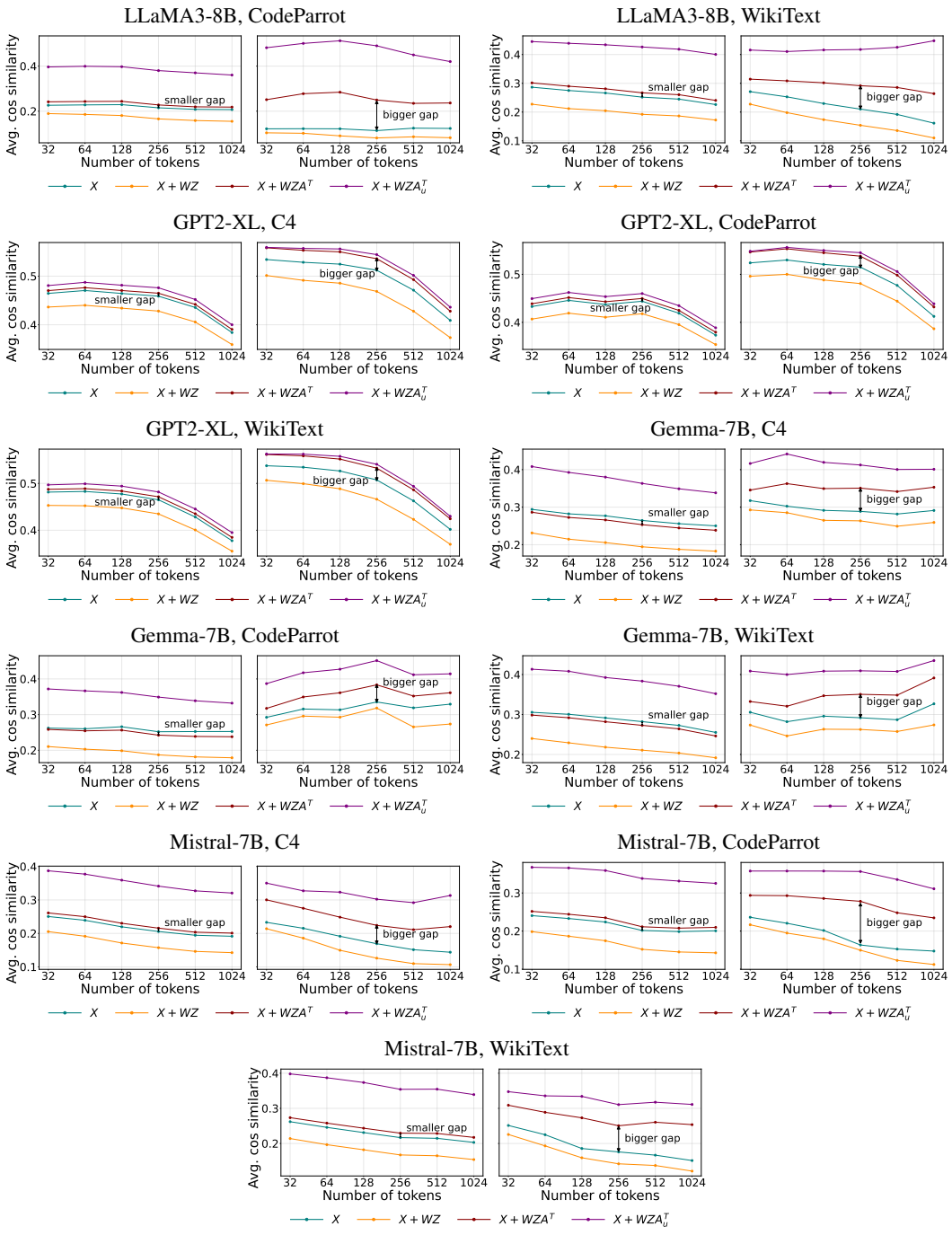

Figure 6: Empirical average cosine similarity for different models, datasets and context lengths for different components of the attention step. Results averaged over all heads (first plot) and over heads with uniformity coefficient larger than 0.6 (second plot). Cosine similarity generally increases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

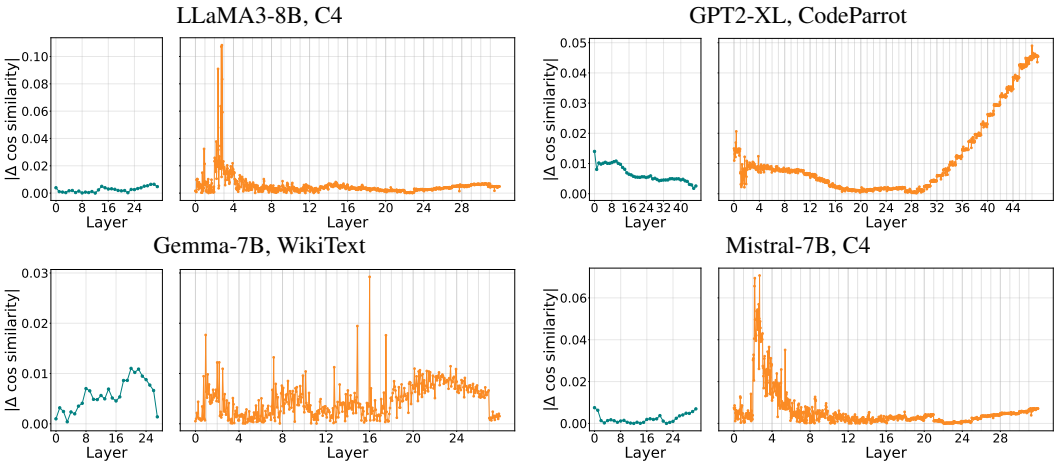

Figure 7: Absolute difference between the empirical average cosine similarity and the theoretical approximation (2) for individual heads, for X (blue plot) and $X + WZ$ (orange plot), across different models and datasets at context length 512. The x-axis indexes the heads in order, from the first to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

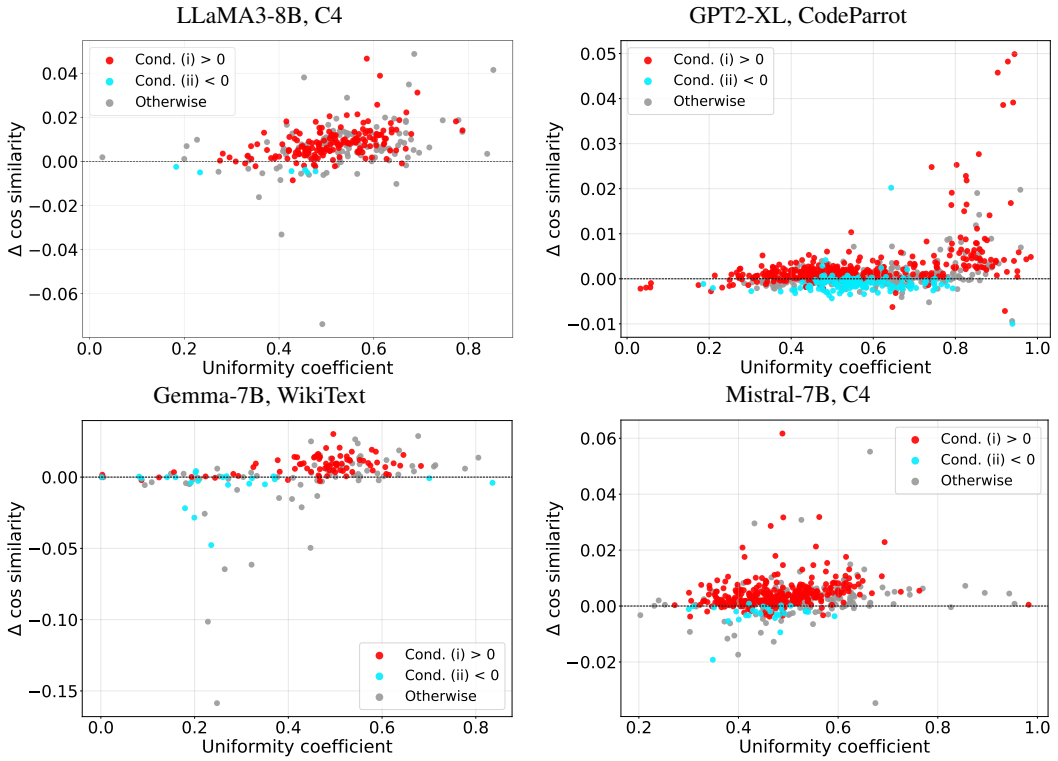

Figure 8: avg. $\cos\mathrm{sim}(X + WZA^\top)$ − avg. $\cos\mathrm{sim}(X + WZ)$ for each head of different models and datasets at context length 512. Heads satisfying $\beta\,\mathrm{tr}(BW) > 0$ (red) tend to yield positive values, whereas heads satisfying $\beta\,\mathrm{tr}(BW) + \mathrm{tr}(BW^\top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

$WZA^T) - \text{avg. cossim}(X + WZ)$ 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. 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).

G.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` [10], `tinyshakespeare` [20], `WikiText` [24], and `CodeSearchNet-Python` [16]). 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.

G.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 [17]. The sequence we tested this on is:

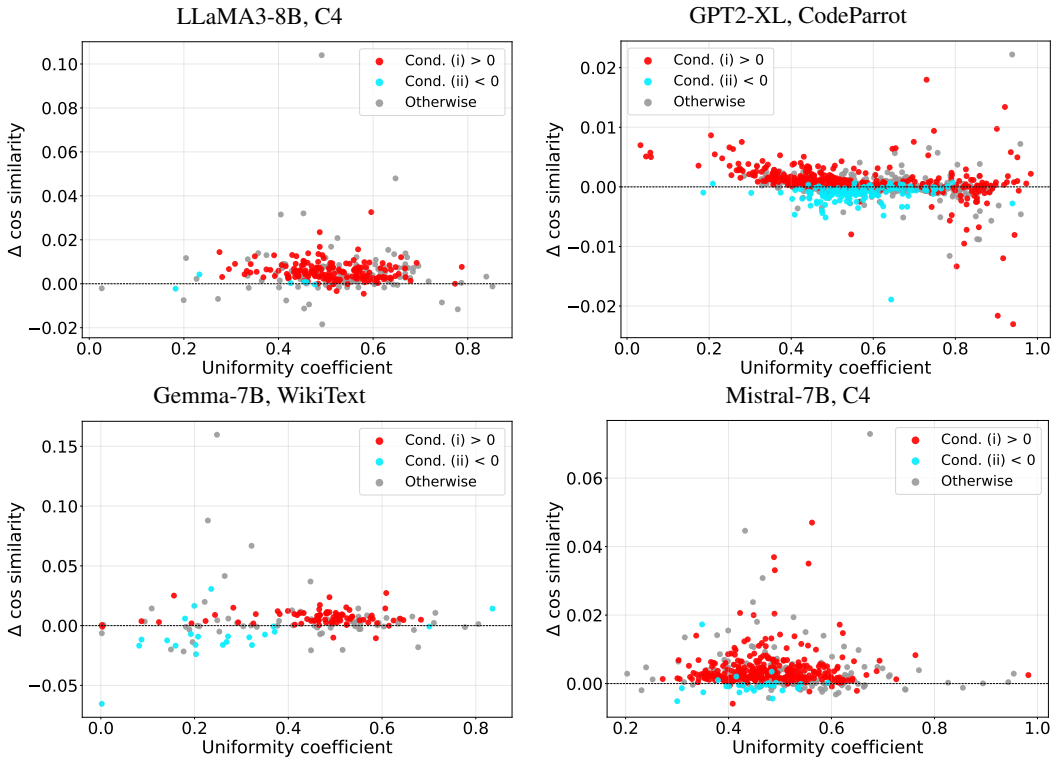

Figure 9: avg. $\cos\mathrm{sim}(X + WZA_u^\top) - \text{avg. } \cos\mathrm{sim}(X + WZA^\top)$ for each head of different models and datasets at context length 512. Heads satisfying $\beta\,\mathrm{tr}(BW) > 0$ (red) tend to yield positive values, whereas heads satisfying $\beta\,\mathrm{tr}(BW) + \mathrm{tr}(BW^\top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

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.

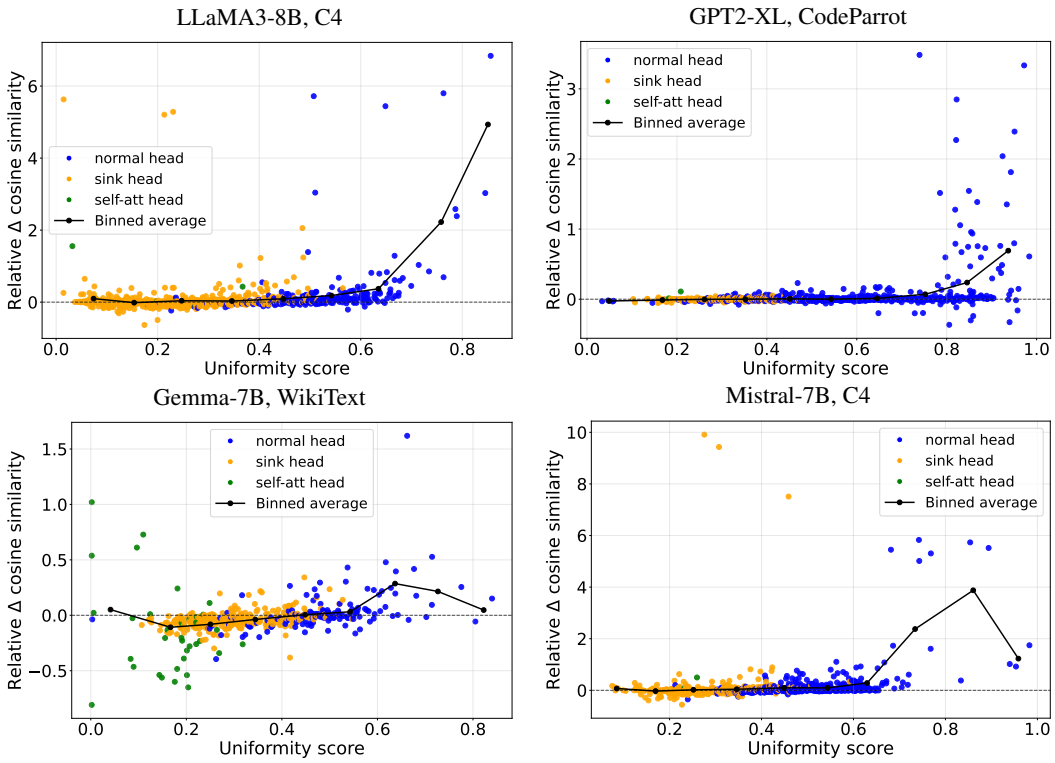

Figure 10: Relative change in average cosine similarity, $(\text{avg. } \cos\text{sim}(X + WZA^\top) - \text{avg. } \cos\text{sim}(X))/\,|\text{avg. } \cos\text{sim}(X)|$, for each head of different models and datasets at context length 512, plotted against the head's uniformity coefficient. Heads are 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.

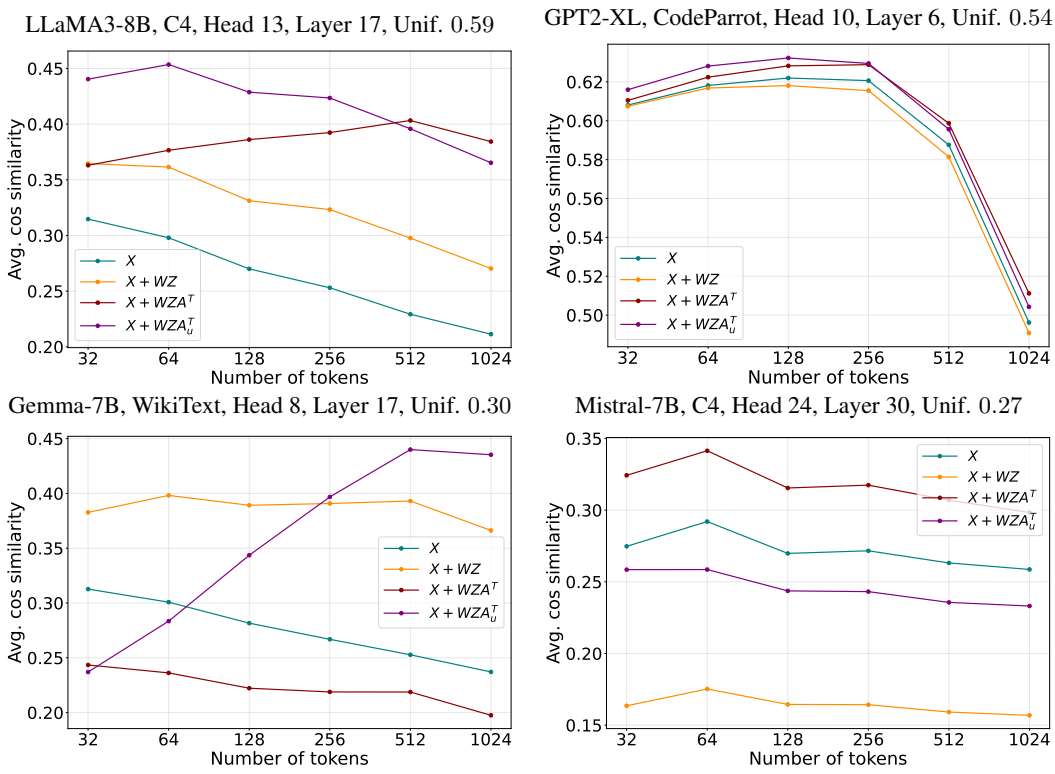

Figure 11: Empirical average cosine similarity for selected heads from different models and datasets, shown for different components of the attention step. These panels illustrate heads that deviate from the typical behavior described in the main text. The reported uniformity coefficient is averaged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

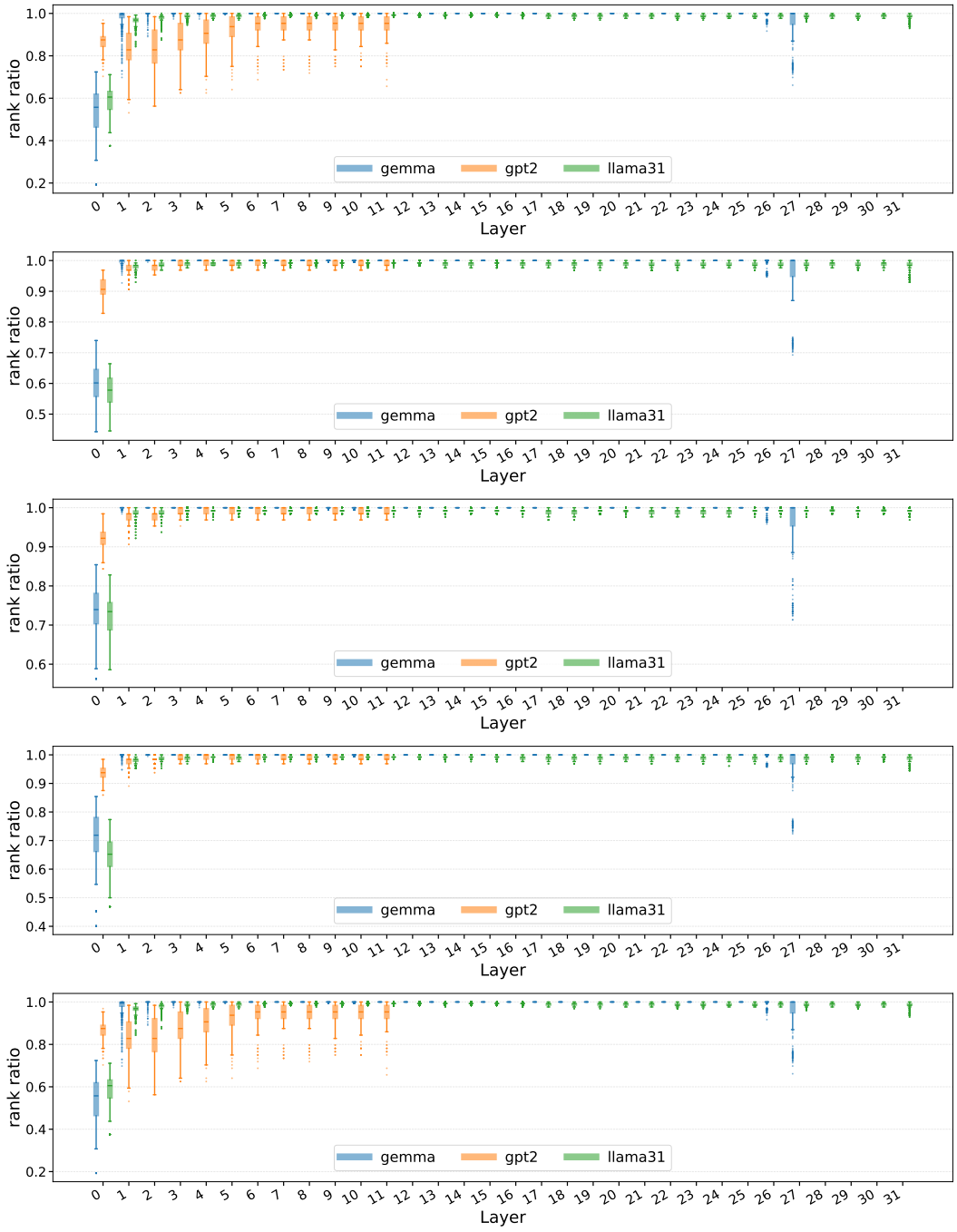

Figure 12: Rank distributions across layers, pooled over heads, 50 sequences and 4 datasets. The whiskers of the box plot point to absolute minimum and maximum. The relative rank tolerance is 0.001. **Upper:** Average for all four datasets. **Lower four:** Individual results for `TinyStories`, `tinyshakespeare`,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

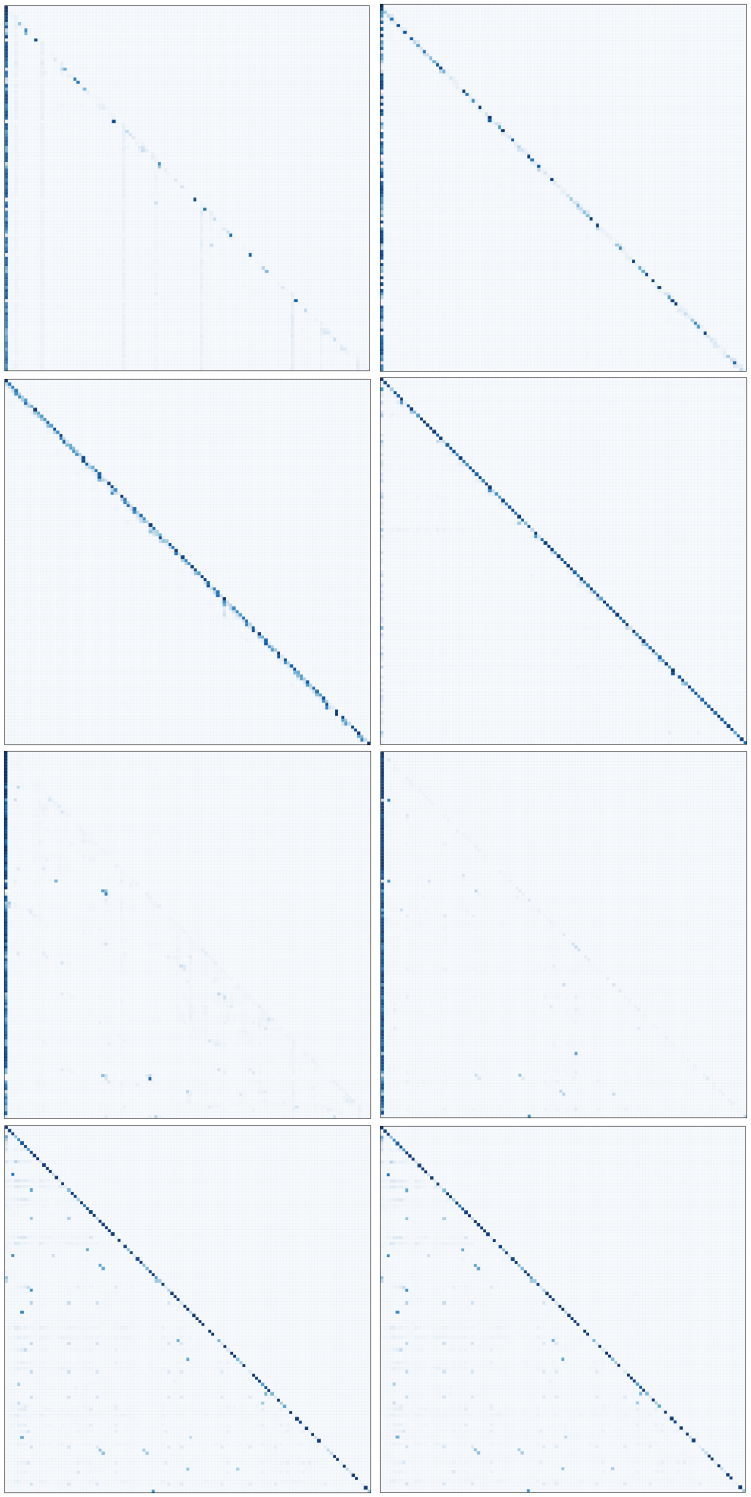

Figure 13: Attention patterns in pretrained transformers (taken randomly) from a pool of models available at `https://attention-motifs.github.io/v2/index.html`. The patterns satisfy the description of our distributions as copy-paste distribution sink (first row), copy-paste distribution diagonal (second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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims precisely correspond to the results presented in the rest of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We always discuss the scope and robustness of our claims. The theory assumptions are verified empirically or otherwise justified.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: At the best of our knowledge, our proofs are correct and the sets of assumptions full.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide enough information for our results to be reproducible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We submit the code as a supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide enough experimental details to understand the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We do not always report error bars, but we always comment about the variance of our results, how many runs we use to obtain the results, and how robust our results are with respect to randomness.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experiments are light-weight and this information is not necessary in this context.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper does conform with the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This theory-focused paper explains observed phenomena and does not have any societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This is a theory-focused paper with no place for potential misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the assets we use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not do crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The LLMs were not used for any core development. We used LLMs for code development and for some proof computations, which we carefully verified.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.