AN ANALYSIS OF POSITION ENCODING IN HYBRID LLMS WITH LOCAL MIXING AND GLOBAL NOPE ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The attention operation is naively position invariant. However, positional information is fundamental to natural language, and therefore a variety of explicit position encodings have been developed in transformer-based models, such as rotary position encoding (RoPE). Although explicit position embeddings have long been assumed to be required, recent methods that interleave local mixing layers, such as sliding window attention (SWA) and gated linear attention, while *not* encoding position (NoPE) in global attention layers has recently been shown to be successful at scale. How and why this approach works is not well-understood. In this paper, we develop an explanation of how hybrid models of this sort can implicitly encode position at global NoPE layers. Supported by both theoretical and empirical evidence, our central argument is that SWA and gated linear attention induce a recency bias in the residual stream that propagates to, and is selected by, the global attention logits. Moreover, in contrast to the implicit position encodings found in models with only global NoPE attention, in which positional information arises solely from the causal mask, the recency bias in hybrid models can be maintained across long sequences. In addition to deepening our understanding of how hybrid models encode position, these findings may provide insights for how to encode position in a way that can extrapolate to longer sequence lengths indefinitely.

1 INTRODUCTION

Distinguishing tokens by their position is a vital component of language modeling (Dufter et al., 2022). Since self-attention by itself is permutation invariant, attention layers in large language models (LLMs) typically utilize position encodings (PEs), which explicitly represent either the absolute or relative position of queries and keys. However, commonly used PEs such as rotary positional encoding (RoPE; Su et al., 2024) incur computational overhead and make length extension and extrapolation difficult. Removing explicit PE (a method called NoPE) has met with mixed success: NoPE does not match the in-domain performance of the same architecture with explicit PE, although it can provide some improvements in length extrapolation while the performance of RoPE and other explicit position embeddings tend to collapse quickly beyond their training sequence length (Haviv et al., 2022; Kazemnejad et al., 2023).

Separately, to mitigate the computational and memory costs of attention across long sequences, so-called *hybrid* architectures that interleave either sliding window attention (SWA) or gated linear layers with global attention layers are increasingly used at scale. Interestingly, as hybrid models have become increasingly widespread, there has been a shift in the global attention layers from using RoPE (Gemma Team, 2024) to p-RoPE (Qwen Team, 2025) to, most recently, NoPE. Notably, Kimi K3 interleaved global NoPE with a gated linear variant called KDA (Kimi Team et al., 2026).

However, there are remaining questions concerning how such models utilize and represent position in their NoPE attention layers, which is critical to their long context performance. To our knowledge, only one prior paper has studied how hybrid models represent position in NoPE layers, and their main finding is that they do not represent *absolute* position (Puvvada et al., 2025) at all. However, this result leaves open the question of whether such models may somehow learn to encode *relative position* in their NoPE layers.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this paper, we present novel theoretical and empirical evidence that hybrid architectures can and do develop relative PEs at their NoPE layers. Unlike other *explicit* relative PEs such as RoPE, hybrid architectures *implicitly* encode relative position via the interactions between local or linear layers and subsequent NoPE layers. Specifically, the SWA or gated linear layers introduce a recency bias in the residual stream that propagates to, and is selected by, the global NoPE attention logits and then accentuated through depth. We demonstrate this both empirically and theoretically:

(i) theoretically, by mathematically demonstrating how SWA approximates a moving-average convolution in expectation, inducing a recency bias in its output that transfers into the residual stream, through intervening layers, and is selected by the global NoPE attention logits; and

(ii) empirically, by demonstrating the existence of the predicted recency bias in the residual stream and NoPE attention logits across depth, present at initialization and strengthening throughout training; we validate this in hybrid models interleaving global NoPE attention with SWA (with RoPE and NoPE) and KDA (Appendix E), trained on natural language and then inputted both natural language and random tokens.

An overview of our hypothesized mechanism is provided in Figure 1. In the following sections, we first note that existing relative PEs generally induce a recency bias (Section 2), then we introduce our core argument in two steps (Section 3). First, SWA produces a residual stream recency bias (Section 4); and second, this recency bias propagates to and is selected by the global NoPE attention logits (Section 5).

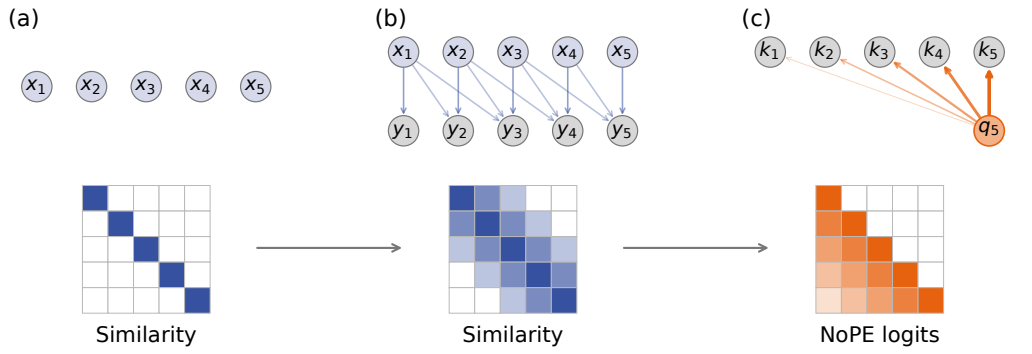


Figure 1: *Proposed mechanism for implicit relative position encoding.* (a) Tokens begin as initially uncorrelated. (b) Local mixing correlates nearby residual states. (c) Learned query and key projections can read out this lag-dependent structure as recency-biased global NoPE logits. The shading strength denotes correlation or logit strength while line width denotes mixing or attention weight.

2 Background: Recency Bias as Relative Position Encoding

Relative PEs, as opposed to absolute PEs, are the industry standard in LLM architectures (Shaw et al., 2018; Dufter et al., 2022; Su et al., 2024). These PEs alter attention logits, directly or indirectly, in a way that depends on the relative distance between query and key. There are a variety of methods for encoding relative positions in attention, but the majority are tied by the following fact:

> *Common and SOTA relative position encoding methods encode relative position by creating a recency bias in attention logits.* That is, given a query q_i at position i, and key k_j at j, relative position encodings bias the attention logits to be smaller the greater the relative distance $i - j$.

Relative PEs encoding a recency bias is not strictly necessary (Chen et al., 2025). In principle, a relative PE applied to attention can induce any kind of bias in attention logits, as long as this bias depends directly on relative position. However, using a recency bias to encode relative position appears to work well in practice, because: (i) it provides a particularly simple positional signal for

the model to learn and use (i.e., in which distance is represented directly by a smooth, monotonic change in the attention strength); and (ii) it fits the statistics of natural sequential data (i.e., the tokens most important for predicting next tokens tend to be the most recent ones; Tan et al., 2025; Wu et al., 2025; Kim et al., 2026).

Survey of relative PEs. There are generally two ways to create a relative PE in attention logits: additively or multiplicatively (Zhang et al., 2026). Additive encodings directly add a bias, b_{i-j}, to attention logits; i.e., $s = q_i^\top k_j + b_{i-j}$, where i and j are the position of query and key and $i - j$ the relative position. Examples include Alibi and Fire (Press et al., 2022; Li et al., 2024), which add a fixed, data-independent bias that decays linearly with relative distance, as well as the more recent Fox PE (Lin et al., 2025) that uses a data-dependent bias. All of these explicitly induce a recency bias in attention logits, such that keys farther from the query have a larger negative bias added to their attention values. Multiplicative PEs include a matrix, R_{i-j}, in the query-key multiplication; i.e., $s = q_i^\top R_{i-j} k_j$, where R_{i-j} is unique to relative position $i - j$. RoPE, arguably the best-known relative PE, rotates pairs of channels using rotation matrices in both the queries and keys. It has been observed that learned attention heads with RoPE either focus on very low-frequency channels with little resulting influence on logit decay, or they focus on high-frequency channels in such a way that they create a strong recency bias (Barbero et al., 2025). Other multiplicative PEs, like the recent Wall attention (Yang et al., 2025), create a similar kind of recency bias in a way that is data-dependent (Yang et al., 2025; Zhang et al., 2026).

PE in NoPE. Among previous work studying the position encoding resulting from NoPE, two papers are particularly relevant to this work. First, Puvvada et al. (2025) studied implicit position encodings in models with global NoPE attention, including hybrid architectures that interleave with SWA. In particular, they trained probes to try to predict absolute position encoding in the residual stream embeddings, finding that the probes could predict absolute position in global NoPE architectures but not in hybrid architectures (see Haviv et al., 2022; Chi et al., 2023 for more evidence of absolute position encodings in global NoPE architectures). This provides evidence that hybrid NoPE architectures, unlike their global NoPE counterparts, do not learn absolute position encodings, but leaves open the possibility that hybrid architectures could encode relative position encodings.

Second, Zuo et al. (2025) showed that architectures with global NoPE attention at every layer could learn a recency bias-based PE in their residual stream, created by the asymmetric token mixing effects of attention with a causal mask. However, their analysis and empirical results only demonstrate a recency bias effect on very short sequences (approximately 30 tokens), leaving open the question of how this bias may change with sequence length. Additionally, they do not explain the conditions under which attention can make use of this recency bias in the residual stream. In this work, we apply a similar analysis to hybrid architectures with SWA layers and compare to global NoPE models. Importantly, we show that SWA has a recency bias effect on the residual stream independent of total sequence length, whereas for initialized global NoPE residuals the recency bias loses resolution even at moderate sequence lengths. Further, unlike previous works, we explain the conditions required of attention parameters to make use of this recency bias in the residual stream, and we scale our empirical results to models with 350M+ parameters trained on natural language data.

3 Thesis: Hybrid Architectures Implicitly Encode Relative Position at Global NoPE Layers

We develop and support the following thesis about hybrid architectures with global NoPE attention:

> *Local layers (e.g., SWA) output a sequence of vectors which, in expectation, have a recency bias (i.e., smaller cosine similarity with greater relative distance). These outputs then create a recency bias in the post-attention residual stream that is persistent across sequence lengths. This residual stream bias can (and does after training) get transferred through the intervening layers and to attention logits at NoPE layers, thereby creating an* implicit *relative position encoding in NoPE layers.*

We provide two lines of evidence that present complementary views of this effect. The theoretical analysis isolates the structural effect of one type of local layer (SWA), and outlines how the resulting recency bias can propagate through intervening modules and into the global NoPE logits. However, our analysis does not fully account for the variety of computation the model is being trained for, particularly the effects of content-dependent attention and multi-layer perceptrons (MLPs). Our empirical analysis then verifies the existence of the residual stream and logit-based recency biases in multiple models, at varying stages of training, on both random and natural language data (for random-token controls, see Appendix D.2). We also show how these effects generalize to hybrid architectures that use KDA (Appendix E) as the local layer instead of SWA. Below, we provide first theoretical and then empirical preliminaries common to each step in our proposed mechanism.

3.1 Theory

First, we formally introduce several notions of similarity in the residual stream, and what we mean by a recency bias. Let $X = [x_1, \ldots, x_T] \in \mathbb{R}^{D \times T}$ be the residual stream at some layer. The most simple and intuitive measurement of similarity is cosine similarity; i.e., $c(d) = \mathbb{E}[x_i^\top x_{i-d} / \|x_i\| \|x_{i-d}\|]$. However, cosine similarity captures just one slice of the full similarity structure; in particular, it does not account for cross-channel correlations that may be picked up by the global attention projections. Therefore, we also introduce the lagged cross-moment matrix, $G_X(d) = \mathbb{E}[x_i x_{i-d}^\top]$, where the x_i are unit-normalized.

To obtain scalar measurements of similarity, one can consider various slices of this matrix; i.e., the Frobenius inner product $\langle G_X(d), M \rangle_F = \mathrm{tr}(G_X(d)^\top M)$ with some matrix M. Different slices of the lagged cross-moment correspond precisely to our similarity measures of interest: cosine similarity is the isotropic component, and the NoPE logits capture alignment with the composed query-key projection. That is,

$$\begin{aligned} \text{Cosine:} \quad & c(d) = \langle G_X(d), I \rangle_F, \\ \text{NoPE logits:} \quad & \ell(d) = \langle G_Z(d), M \rangle_F. \end{aligned}$$

Here, I is the identity matrix, Z is the actual post-normalization input to a NoPE head at its (unnormalized) model scale, and $M = W_Q^\top W_K / \sqrt{D_H}$ with head dimension D_H. Provided one of the above scalar similarity profiles $f(d)$, we compute the recency gap as $f(1) - f(L)$, where L is the far-distance reference lag (4096 in our experiments); then, $f(L)$ is the similarity floor, or the underlying correlation structure in the sequence present at long distances.

3.2 Empirics

For the empirical analysis, we study hybrid architectures at both the 120M and 350M scales, varying the SWA window size $w \in \{64, 128, 256, 512, 1024, 2048, 4096\}$ and comparing to a global NoPE baseline (i.e., no SWA interleave) at each scale. Unless otherwise noted, the SWA layers use RoPE and are interleaved with global NoPE at a 1:1 ratio. The main text shows results for the 350M models, and the $w = 128$ variant for single-model analyses. We train the models on the Prolong dataset (Gao et al., 2025). For evaluations periodically throughout training, we use TextbookChapters (Chevalier et al., 2024).

Using the measures defined in Section 3.1, we track residual stream cosine similarity and global NoPE attention logits across training, depth, and SWA window size. For more details on the experimental setup, including training setup, statistics sampling, and head aggregation, see Appendix A.

4 Step 1: SWA Produces a Recency Bias in the Residual Stream

4.1 Theory

Unlike full attention, SWA restricts its attention operation to a local window of keys and values. Here, we show that this local operation gives the attention output a distance-dependent cosine similarity profile. Furthermore, because its scale is fixed by the window size rather than the sequence length (in contrast to full attention), this effect does not dilute as the total sequence length grows.

For a single attention head, define $q_i = W_Q x_i$, $k_i = W_K x_i$, $v_i = W_V x_i$. For a given q_i and k_{i-d}, the attention weight at distance d is:

$$\alpha_i(d) = \frac{\exp\left(q_i^\top k_{i-d}/\sqrt{D_H}\right)}{\sum_{r=0}^{w-1} \exp\left(q_i^\top k_{i-r}/\sqrt{D_H}\right)}, \qquad 0 \le d < w,$$

where w is the sliding window size and D_H the head dimension. Writing $o_i = W_O W_V x_i$, the SWA output is

$$y_i = W_O \sum_{d=0}^{w-1} \alpha_i(d) v_{i-d} = \sum_{d=0}^{w-1} \alpha_i(d) o_{i-d}.$$

Hence, SWA acts akin to a data-dependent convolution.

Intuition. To illustrate why such a local filter may produce a recency bias in its output, consider a simplified scenario that closely reflects SWA in an LLM at initialization, in which the projected values o_i are zero-mean, of similar magnitude, and uncorrelated across positions, while the attention weights are roughly uniform. Then we may approximate the SWA output as a moving average,

$$y_i \approx \frac{1}{w} \sum_{r=0}^{w-1} o_{i-r}.$$

In this simplified scenario, the cosine similarity of two outputs, say $y_i = \frac{1}{w}(o_{i-w+1} + \cdots + o_i)$ and $y_{i-d} = \frac{1}{w}(o_{i-d-w+1} + \cdots + o_{i-d})$, depends on the fraction of projected values shared by their sums. The closer they are in position, the larger this shared fraction; e.g., for a window size 64, y_i shares 63 terms with y_{i-1}, 62 with y_{i-2}, and so on. Thus, the expected output cosine decreases with relative distance and approaches zero once the windows no longer overlap.

Length scaling and comparison with global NoPE. As noted in Section 2, Zuo et al. (2025) did a similar analysis, but for global NoPE architectures; i.e., they showed that a global NoPE attention layer could induce an adjacency pattern in its outputs. However, they considered only tiny sequences (~ 30 tokens) and did not explore whether this effect scaled with sequence length. Here, the uniform-attention toy case helps explain why global NoPE cannot sustain a well-resolved recency bias even out to 1000-token sequences. Specifically, the output of a global NoPE layer will be

$$y_i \approx \frac{1}{i+1} \sum_{t=0}^{i} o_t.$$

Unlike SWA, the model averages over the entire preceding sequence. Therefore, if $i = 1000$, then y_i and y_{i-10} share 99% of their terms, so the uniform-attention cosine changes only slightly over fixed lags. Together, these calculations predict a persistent window-scale recency profile for SWA, but not a comparably resolved fixed-lag profile for global NoPE.

To make the contrast with SWA more concrete, under the same initialization-like assumptions as above, the output's cosine similarity with distance is:

$$c_{\text{SWA}}(d) \approx \left[1 - \frac{d}{w}\right]_+, \qquad c_{\text{global}}(d) \approx \sqrt{(i-d+1)/(i+1)} \propto 1 - O(d/i).$$

Hence, at initialization, the strength of SWA's recency bias decays with the window size w, with a smaller window resulting in a stronger, steeper recency bias; conversely, global attention's positional signal decays with the sequence length i, and therefore dilutes for longer sequences. That is, SWA's locality scale is length-independent by construction, whereas global attention's is not.

General mean-attention case. We now describe how SWA affects the similarity structure in its output after training, where the attention profile is no longer uniform and the incoming residual stream no longer zero-mean Gaussian. Again, we approximate attention by its mean profile $a_d = \mathbb{E}[\alpha_i(d)]$ (estimated across some number of relevant sequences). Suppose that X has some incoming lagged cross-moment structure $G_X(d)$; let $A = W_O W_V$ be the composed projected-value matrix

and $g_a(k) = \sum_s a_s a_{s+k}$ the mean attention autocorrelation. In this setting, the SWA output is $y_i \approx A \sum_{d=0}^{w-1} a_d x_{i-d}$, and its lagged cross-moment is

$$G_Y(d) \approx A \sum_k g_a(k) G_X(d-k) A^\top = A(g_a * G_X)(d) A^\top,$$

if output norms are concentrated (see Appendix B.1 for the full derivation). Thus, the similarity structure in the output is the incoming structure mixed across distances by the mean attention autocorrelation, and then mixed across channels by the composed attention matrices. The structural effect of SWA is still present, as $g_a(k) > 0$ if $|k| < w$, and 0 otherwise. So, the window bounds the scale of the recency bias, but its effect can be amplified or dampened by the attention profile within the window. Finally, the extent to which this mixing promotes or dilutes a recency bias in the output with respect to each similarity metric depends on its alignment to the relevant projections.

4.2 Empirics

Initialization and length scaling. Empirically, we first validate the argument that SWA's effect is structurally different than that of global NoPE, and that only the former produces a well-resolved recency bias at long sequence lengths. Figure 2 compares the isolated first attention-branch outputs before residual addition at initialization. With increasing sequence lengths (probed via query position along an 8k-token sequence), global NoPE activations become so highly correlated that positions cannot be resolved for most distances; in contrast, SWA retains a strong distance-dependent profile across the window while flattening beyond, regardless of sequence length.

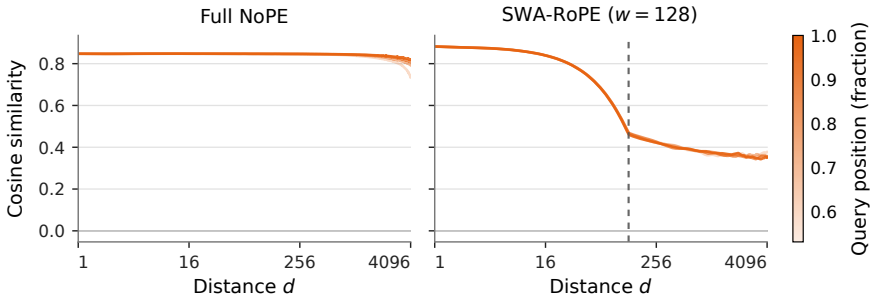


Figure 2: *Attention outputs at initialization.* Raw cosine of the isolated first attention-branch outputs for global NoPE and SWA ($w = 128$; marked by dashed line), across four fixed query-position bins. The recency bias in the SWA output is stronger and remains constant with increased sequence length, whereas full attention's weakens significantly at longer sequences.

Emergence with training and depth. Though the above corroborates that SWA produces a recency bias at initialization, it is possible that this effect dissipates during training as the model focuses on learning content-dependent interactions. However, this is not the case – in fact, the recency bias strengthens during training. Figure 3 visualizes the recency profile during training at the first, middle, and last SWA layers. Across depth, the trained profiles show strengthening and subsequent attenuation relative to their peak, while remaining stronger than at initialization. Note that this effect stabilized early in training (the colorbar is logarithmic in the fraction of that scale's total training tokens). Furthermore, the recency bias strengthens significantly with increased depth. Further analysis as well as layer-by-layer profiles are shown in Appendix B.2.

Window-size dependence. The theoretical analysis in Section 4.1 suggests that smaller SWA windows will produce a narrower and steeper recency bias in the outgoing residual stream. The distance profiles (top) and layer-wise recency gaps (bottom) in Figure 4a verify this trend; smaller windows lead to stronger recency biases across depth (for full profile comparisons, see Figure 8 in Appendix B). Further, to mitigate the confound of the strong logit-based recency bias induced by RoPE within the window (which as we discussed in Section 4.1 has an important contribution to the recency bias), we also verify this for models with NoPE-in-SWA. Figure 4b shows the distance

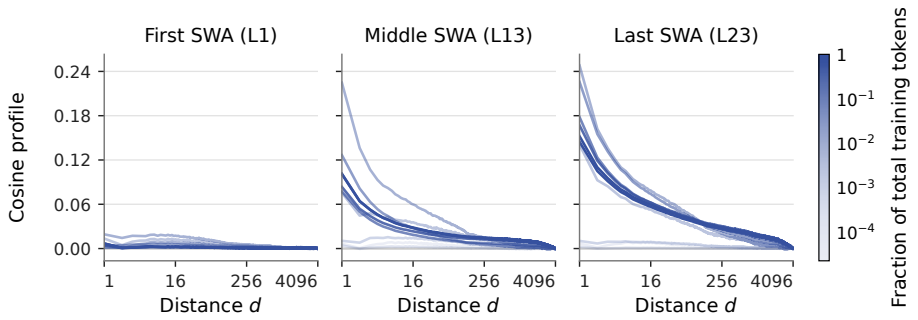


Figure 3: *Residual cosine during training.* Post-attention cosine profiles at the first, middle, and last SWA layers ($w = 128$), with color denoting the fraction of training completed on a logarithmic scale. This demonstrates that residual stream recency bias increases with training and depth.

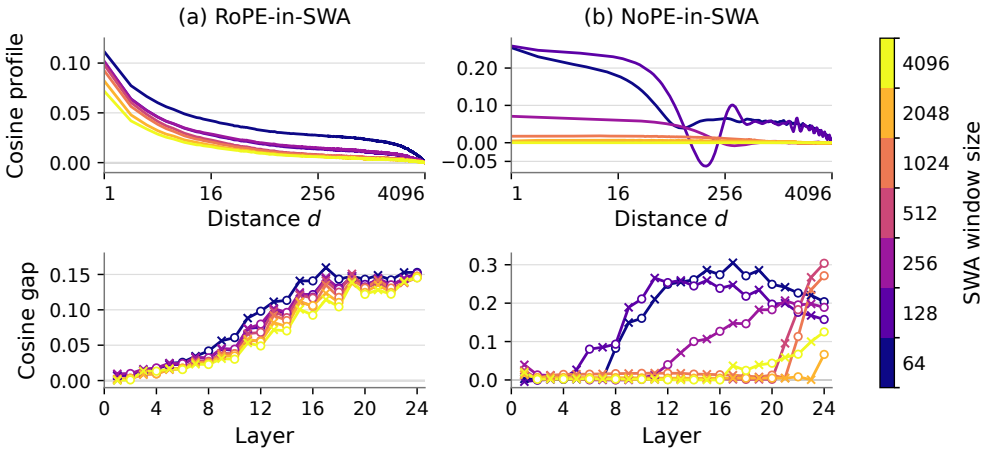


Figure 4: *Window dependence of residual cosine.* RoPE-in-SWA (left) and NoPE-in-SWA (right), with top showing the cosine gap $c(d) - c(4096)$ at the middle SWA layer (L13), and bottom the cosine gap $c(1) - c(4096)$ across depth (crosses mark SWA layers and circles global NoPE layers). With or without RoPE, smaller windows result in stronger recency biases across depth.

profiles and layer-wise recency gap across window sizes for NoPE-in-SWA; in fact, the effect of window size is more pronounced, with smaller windows producing significantly stronger recency biases (for further NoPE-in-SWA results, please see Figure 15 in Appendix D).

Given the pronounced effect of SWA window size on recency bias strength, a further question one might ask is how window size affects model performance. We find that smaller windows achieve lower loss than their larger-windowed counterparts (see Figures 9 and 15 in the appendix, which report the late-training validation-loss trajectories; note that larger NoPE-in-SWA windows cause loss collapse). Though preliminary, this is promising evidence that the recency bias strength is important to such hybrid models' success; in fact, strengthening the recency bias effect by decreasing window size may outweigh the associated reduction in training compute. Also note that, since these experiments did not match compute across window sizes, the advantage for small windows is even stronger than shown here.

5 Step 2: Global NoPE Inherits the Residual Stream Recency Bias

5.1 Theory

Preservation by intervening transformations. Now, we analyze the conditions for propagation of the residual stream recency bias into the global NoPE logits. First, the recency bias must be preserved

by the intervening modules between SWA and global NoPE: residual addition, normalization, and MLPs. Broadly speaking, residual addition moderately dilutes the recency bias in the SWA output, normalization modules increase the recency bias by reducing the similarity floor, and MLPs can suppress, amplify, or reorient the recency bias depending on dynamics learned in training. Though we omit further details here for brevity, please see Appendix C for a module-level theoretical and empirical analysis where we find these largely preserve the incoming cosine gap.

Selection by NoPE logits. For now, assume the recency bias has been preserved until the input of global NoPE attention. Let x_i denote the post-normalization input to global NoPE's query and key projections. For a fixed attention head, let $M = W_Q^\top W_K / \sqrt{D_H}$ and recall $G_X(d) = \mathbb{E}[x_i x_{i-d}^\top]$. Recall that the expected logit profile is

$$\ell(d) = \mathbb{E}[x_i^\top M x_{i-d}] = \langle G_X(d), M \rangle_F.$$

Thus, the logit recency gap $\ell(1) - \ell(L)$ is positive precisely when M aligns positively with $G_X(1) - G_X(L)$; its magnitude depends on the extent of that alignment. Notably, in contrast to cosine, this similarity profile can depend on the skew component of the full cross-moment structure. Depending on the slice orientation, the projections can preserve, distort, or reverse the residual-similarity ordering (see Appendix C for further discussion).

Note also that, for independent zero-mean query and key initializations, $\mathbb{E}[M] = 0$; therefore, the recency bias does not immediately propagate to the full attention logits in expectation over initialization. However, training can align W_Q and W_K, allowing their logits to inherit the residual stream recency bias, which predicts that the effect should become visible in NoPE logits as training induces favorable alignment.

5.2 Empirics

Emergence with training. To empirically verify this, Figure 5 tracks the global NoPE attention logits during training. Indeed, at initialization the distance dependence is weak; but, as training progresses, a recency-biased profile emerges across depth, consistent with learning the projection alignment described above. Similarly to the development of the residual stream recency bias following SWA, the recency bias in the full attention logits strengthens and stabilizes very early in training (the colorbar is logarithmic in the fraction of training completed). But, whereas the residual stream recency bias increases consistently with depth, the logit recency bias quickly strengthens then stabilizes (layer-by-layer profiles are shown in Appendix C.2).

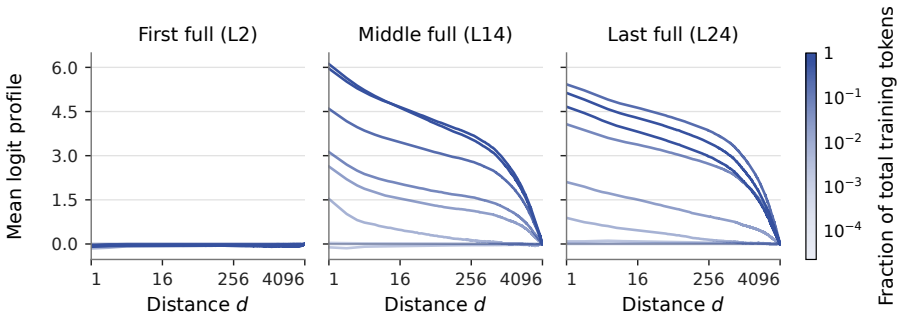


Figure 5: *Global NoPE logits during training.* Head-mean profiles at the first, middle, and last global layers ($w = 128$), with color denoting the fraction of training completed on a logarithmic scale. This shows that, through training, the global NoPE heads learn to select the residual stream recency bias and thereby develop a corresponding recency bias in their logits.

Window-size dependence. We found in Section 4 that smaller windows induce a stronger residual stream recency bias. To investigate the extent to which this trend propagates into the global NoPE logits, Figure 6 summarizes the layer and head distribution of the recency gap. In comparison to the residual stream, a few distinct patterns emerge. For both RoPE-in-SWA and NoPE-in-SWA, the logit gap remains fairly stable across depth after initially strengthening (as seen above). For RoPE-in-SWA

(Figure 6a), there is little variation by window size too; however, for NoPE-in-SWA (Figure 6b), the logit gaps for larger windows swing drastically positive and negative (note that the recency gap axis is symmetric-log scaled to account for the wider variation). Further analysis as well as layer-by-layer profiles are shown in Appendix C.2.

These patterns are suggestive that the model seeks to learn a consistent logit recency bias across depth. However, if window size is too large (and therefore, the residual stream recency bias too weak), then the model is unable to develop a logit recency bias, resulting in performance collapse. This aligns with the spirit of our proposed mechanism: the role of local layers is to produce a residual stream recency bias available to be picked up by the global NoPE logits; within the global NoPE logits, the model learns to attenuate a consistent recency bias across heads to serve as its position encoding in attention.

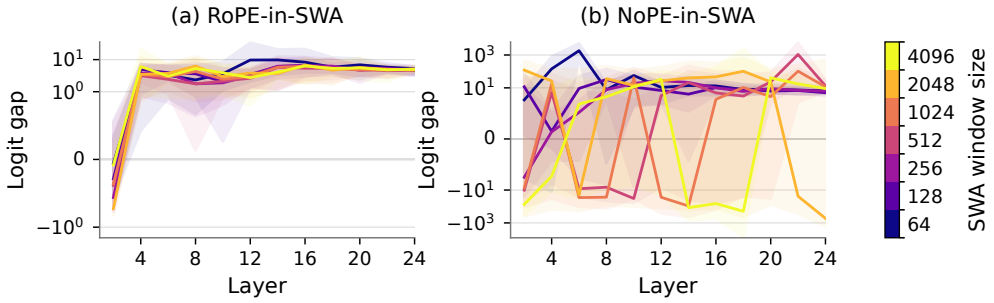


Figure 6: *Global NoPE logit gaps.* RoPE-in-SWA (left) and NoPE-in-SWA (right), showing the head-mean gap $\bar{\ell}(1) - \bar{\ell}(4096)$ (shading spans the full headwise range and vertical axes are symmetric-log). Though the window dependence is less clear in RoPE-in-SWA, it appears the logit recency bias collapses for large windows with NoPE-in-SWA.

6 Discussion

Despite their empirical success, how hybrid models interleaving SWA and global NoPE attention encode position in their global NoPE layers has remained a mystery. In this paper we develop a novel mechanistic explanation for how such models accomplish this encoding. Firstly, SWA creates a recency bias in its output by correlating neighboring embeddings, which emerges as a fundamental and in-built inductive bias present from initialization. In the second stage, this recency bias is propagated to the global attention logits, first preserved by the intervening transformations (including residual stream addition, normalization, and MLPs), and then selected by the global attention projections (to varying degrees across heads). Furthermore, this proposed mechanism may be extended to hybrid models that interleave linear layers like KDA in place of SWA (Appendix E). This recency bias acts as a kind of *implicit* relative position encoding, creating effects similar to explicit relative position encodings.

Our thesis of how such models encode position is particularly interesting given the special capacity of hybrid models with global NoPE attention for length extension and generalization. We hypothesize this is for a few reasons. For one, the window provides a finite-length local scale that does not dilute as sequence length increases, and which is invariant to the absolute sequence length. Also, the position encoding learned does not radically change at distances beyond training, as is the case with RoPE, implying that extrapolation to longer sequence lengths does not introduce any serious discontinuities. Finally, this position encoding may be uniquely flexible due to residing within the residual stream geometry, which may enable expressive and easily-learned interactions with content that may be adaptively picked up by the global attention heads.

Nevertheless, there are still many unknowns about how hybrid models with SWA and global NoPE attention encode position. For example, we still have not developed a complete quantitative account of how the recency bias propagates through all modules of the model, or how this affects the attenuation of the recency bias across depth and varying sequence lengths. Nor have we established a quantitative relationship between these characteristics and the extent of length generalizability. These topics are left to future work.

AI USE STATEMENT

In this work, we used generative AI tools to help with literature review, creation of code and execution of experiments/analysis, formalizing math in parts of the theoretical analysis, generating figures, and polishing some writing. We have not used generative AI tools to generate core research ideas or directions, frame the experimental methodology, or write sections of the paper. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work (including text, claims or artifacts produced with the aid of generative AI).

REFERENCES

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.06205.

Yuhan Chen, Ang Lv, Jian Luan, Bin Wang, and Wei Liu. HoPE: A novel positional encoding without long-term decay for enhanced context awareness and extrapolation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23044–23056, 2025. doi: 10.18653/v1/2025.acl-long.1123. URL https://aclanthology.org/2025.acl-long.1123/.

Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, Toni Annala, Max Aragon, Arturo Rodriguez Fanlo, Simon Frieder, Simon Machado, Akshara Prabhakar, Ellie Thieu, Jiachen T. Wang, Zirui Wang, Xindi Wu, Mengzhou Xia, Wenhan Xia, Jiatong Yu, Junjie Zhu, Zhiyong Ren, Sanjeev Arora, and Danqi Chen. Language models as science tutors. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8310–8335, 2024. URL https://proceedings.mlr.press/v235/chevalier24a.html.

Ta-Chung Chi, Ting-Han Fan, Li-Wei Chen, Alexander Rudnicky, and Peter Ramadge. Latent positional information is in the self-attention variance of transformer language models without positional embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1183–1193, 2023. doi: 10.18653/v1/2023.acl-short.102. URL https://aclanthology.org/2023.acl-short.102/.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/abea47ba24142ed16b7d8fbf2c740e0d-Paper.pdf.

Philipp Dufter, Martin Schmitt, and Hinrich Sch
"utze. Position information in transformers: An overview. *Computational Linguistics*, 48(3): 733–763, 2022. doi: 10.1162/coli_a_00445.

Tianyu Gao, Alexander Wettig, Howard Yen, and Danqi Chen. How to train long-context language models (effectively). *arXiv preprint arXiv:2410.02660*, 2025. URL https://arxiv.org/abs/2410.02660.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118.

Akshat Gupta, Atahan Ozdemir, Caoqinwei Gong, and Gopala Anumanchipalli. Geometric interpretation of layer normalization and a comparative analysis with rmsnorm. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 375–407, 2026.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1382–1390, 2022. doi: 10.18653/v1/2022.findings-emnlp.99. URL https://aclanthology.org/2022.findings-emnlp.99/.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24892–24928, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4e85362c02172c0c6567ce593122d31c-Abstract-Conference.html.

Junu Kim, Xiao Liu, Zhenghao Lin, Lei Ji, Yeyun Gong, and Edward Choi. LayerNorm induces recency bias in transformer decoders. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. URL https://aclanthology.org/2026.findings-acl.1430/.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, et al. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026. URL https://arxiv.org/abs/2607.24653.

Shanda Li, Chong You, Guru Guruganesh, Joshua Ainslie, Santiago Ontanon, Manzil Zaheer, Sumit Sanghai, Yiming Yang, Sanjiv Kumar, and Srinadh Bhojanapalli. Functional interpolation for relative positions improves long context transformers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.04418.

Zhixuan Lin, Evgenii Nikishin, Xu Owen He, and Aaron Courville. Forgetting transformer: Softmax attention with a forget gate. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.02130.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2108.12409.

Krishna C. Puvvada, Faisal Ladhak, Santiago Akle Serrano, Cheng-Ping Hsieh, Shantanu Acharya, Somshubra Majumdar, Fei Jia, Samuel Kriman, Simeng Sun, Dima Rekesh, and Boris Ginsburg. SWAN-GPT: An efficient and scalable approach for long-context language modeling. *arXiv preprint arXiv:2504.08719*, 2025. URL https://arxiv.org/abs/2504.08719.

Qwen Team. Qwen3-Next-80B-A3B-Instruct model card. Hugging Face, 2025. URL https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 464–468, 2018. doi: 10.18653/v1/N18-2074. URL https://aclanthology.org/N18-2074/.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://arxiv.org/abs/2104.09864.

Shawn Tan, Songlin Yang, Aaron Courville, Rameswar Panda, and Yikang Shen. Scaling stick-breaking attention: An efficient implementation and in-depth study. *arXiv preprint arXiv:2410.17980*, 2025. URL https://arxiv.org/abs/2410.17980.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.01951.

Songlin Yang, Yikang Shen, Kaiyue Wen, Shawn Tan, Mayank Mishra, Liliang Ren, Rameswar Panda, and Yoon Kim. PaTH attention: Position encoding via accumulating householder transformations. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2080. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/59c27bf8d56d3d50c7aeaf7535dee975-Abstract-Conference.html.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Yifan Zhang, Zixiang Chen, Yifeng Liu, Zhen Qin, Huizhuo Yuan, Kangping Xu, Yang Yuan, Quanquan Gu, and Andrew Chi-Chih Yao. Group representational position encoding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2512.07805.

Chunsheng Zuo, Pavel Guerzhoy, and Michael Guerzhoy. Position information emerges in causal transformers without positional encodings via similarity of nearby embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9418–9430, 2025. URL https://aclanthology.org/2025.coling-main.632/.

Appendix

A Measurement and Experimental Details **13**

A.1 Similarity measures and estimation 13

A.2 Models, training, and evaluation inputs 13

A.3 Aggregation and uncertainty conventions 14

B Supplement to Step 1: SWA Produces a Recency Bias in the Residual Stream **14**

B.1 Theory 14

B.2 Empirics 14

C Supplement to Step 2: Global NoPE Inherits the Residual Stream Recency Bias **15**

C.1 Theory 16

C.2 Empirics 17

D Controls **19**

D.1 120M-scale robustness 19

D.2 Random-token controls 21

E KDA extension **26**

E.1 Theory 26

E.2 Empirics 26

A Measurement and Experimental Details

A.1 Similarity measures and estimation

Cosine is measured after attention is added to the residual stream, except for isolated attention outputs at initialization and explicitly named module states. For matrix plots, we average outer products of unit-length states before taking norms; this captures shared structure, though sampling noise can contribute. Cosine and logit profiles use every integer lag from 1 to 4096.

A.2 Models, training, and evaluation inputs

SWA configurations. We study the 120M and 350M hybrid models described in Section 3.2, alternating SWA with global NoPE attention. SWA windows are $64, 128, 256, 512, 1024, 2048, 4096$; single-model comparisons use $w = 128$. The SWA layers use RoPE unless stated otherwise. The two model scales were trained for 12B and 24B tokens, respectively, with sequence length 8192. Training and natural-language evaluation text use the GPT-2 tokenizer's 50,257 IDs; the model embedding vocabulary is padded to 50,304 IDs. Main-text figures use 350M models and TextbookChapters unless their captions state otherwise.

KDA configurations. The KDA controls interleave KDA and global NoPE layers at both model scales. The supplementary KDA figures use final-checkpoint measurements on both TextbookChapters and random-token inputs; their aggregation conventions are stated below.

Evaluation support. Recency measurements use 48 sequences of length 8192. The natural-language inputs come from TextbookChapters; random-token controls sample IDs uniformly from the padded model vocabulary with seed 0. Queries are fixed at zero-based positions 4352–8191 for

every lag $1 \le d \le 4096$, giving 184320 query–key pairs per lag. Initialization plots partition this support into four bins of 960 query positions. Validation-loss plots instead use DCLM sequences of length 4096 and show raw trajectories over the final 30% of training.

A.3 Aggregation and uncertainty conventions

Global NoPE logits. In the SWA and KDA profile and gap figures, we average actual pre-softmax logits across heads at each lag. The logit gap is $\bar{\ell}(1) - \bar{\ell}(4096)$ from this head-mean profile. Where shown, shading spans the range of headwise gaps, not uncertainty.

Module changes. The module figures use the final 350M RoPE-in-SWA, $w = 128$ checkpoint. Each point is an after-minus-before change, averaged equally over the twelve SWA or twelve full-attention blocks. Residual addition steps compare with the isolated branch; block summaries compare the updated residual with its incoming residual. Matrix norms are taken after averaging matrices across queries and sequences, separately at each state and layer. Query–key pairs are not treated as independent samples.

B Supplement to Step 1: SWA Produces a Recency Bias in the Residual Stream

This appendix provides additional theoretical and empirical results that corroborate those in Section 4, using the same setting of 350M scale models inputted natural language (TextbookChapters).

B.1 Theory

Mean-attention cross-moment derivation. As in Section 4.1, we assume that the norms of vectors entering and leaving the attention branch are concentrated, and therefore we work directly on unit-normalized moments G_X and G_Y. Assume these moments are approximately stationary away from the causal boundary. For a fixed mean attention profile a_d, zero outside $0 \le d < w$, let $y_i = A \sum_d a_d x_{i-d}$. Expanding both outputs gives

$$\begin{aligned} G_Y(d) &= A \sum_{r,s=0}^{w-1} a_r a_s \, \mathbb{E}[x_{i-r} x_{i-d-s}^\top] A^\top \\ &= A \sum_{r,s=0}^{w-1} a_r a_s \, G_X(d+s-r) A^\top \\ &= A \sum_{k=-(w-1)}^{w-1} g_a(k) G_X(d-k) A^\top, \qquad g_a(k) = \sum_s a_s a_{s+k}. \end{aligned}$$

Here d is the output lag and $k = r - s$ is the difference between two filter offsets. For learned attention, taking $a_r = \mathbb{E}[\alpha_i(r)]$ is a fixed-profile approximation that omits weight fluctuations and their dependence on the inputs.

For uncorrelated white inputs, $G_X(d) = \Sigma \mathbb{1}\{d = 0\}$, and therefore $G_Y(d) = g_a(d) A \Sigma A^\top$. If we further assume attention is uniform, then the triangular profile introduced in the text is a special case of the same convolution.

B.2 Empirics

Depth-wise accumulation. As noted in Section 4.2, the residual stream recency bias accumulates across depth. To investigate this in finer resolution, Figure 7 shows the residual stream recency profile following every SWA layer at the last checkpoint of training, which demonstrates that the trend of increasing recency bias through depth is notably smooth across layers. Comparing Figure 7 to its 120M-scale companion Figure 14c in Appendix D.1, it is particularly notable how consistent this layer-wise recency bias trend is across model scales, not only in ordering, but magnitude. This supports that the residual stream recency bias provides a functional role regardless of scale.

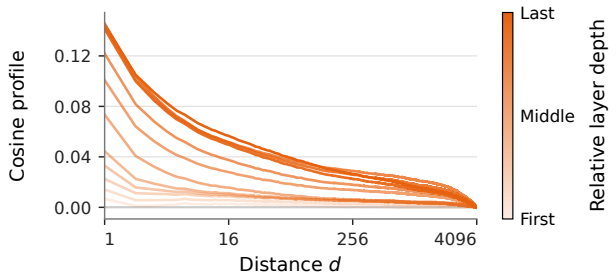


Figure 7: *Residual cosine across depth.* 350M post-attention cosine gap profiles $c(d) - c(4096)$ at the final $w = 128$ checkpoint. Darker orange denotes deeper SWA layers.

Window-size dependence and validation loss. To supplement the comparison of window sizes in Figure 4, here we present two figures: Figure 8 displays the full cosine profiles at the first, middle, and last SWA layers of different window sizes; and Figure 9 shows the validation loss curves for the same range of window sizes. These display that varying window size has a clear and predictable effect on both cosine profiles and language modeling (though we cannot fully isolate its causal role).

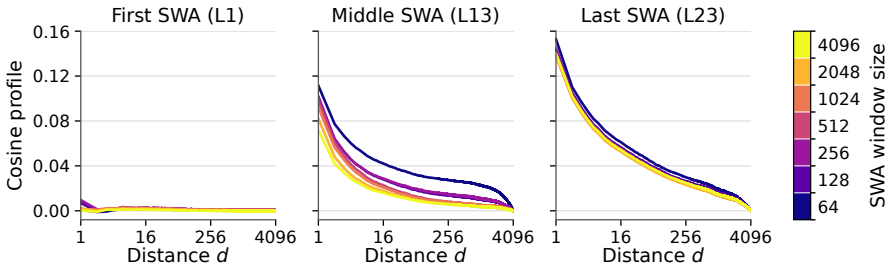


Figure 8: *Window dependence of residual structure.* Final first, middle, and last SWA layers of the 350M models; colors denote windows 64–4096. Curves show post-attention residual cosine profiles $c(d) - c(4096)$.

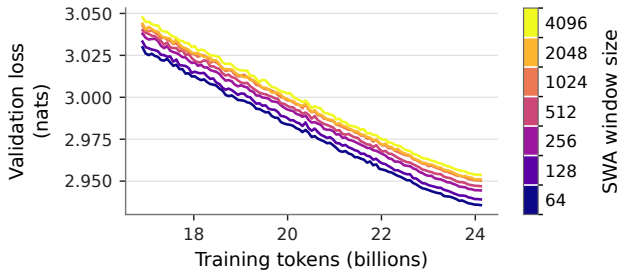


Figure 9: *Validation loss across window sizes.* Mean 350M DCLM token cross-entropy (nats) over the final 30% of training, evaluated on 4096-token sequences.

C Supplement to Step 2: Global NoPE Inherits the Residual Stream Recency Bias

This appendix examines how residual addition, normalization, and MLPs transform the recency structure induced by SWA before it is read out by global NoPE. It provides the theoretical analysis and module-level measurements supporting the transition from Section 4 to Section 5. We also

explain how the cross-moment magnitude and alignment helps understand the recency bias in both the residual stream and logits.

C.1 THEORY

Addition of the residual stream. Suppose there is some similarity structure $G_Y(d)$ in the SWA output Y, and $G_X(d)$ in its input X. Denote the following residual stream by $Z = X + Y$; we wish to understand its similarity structure $G_Z(d)$ as a function of its components. Expanding the outer-product in $G_Z(d) = \mathbb{E}[z_i z_{i-d}^\top]$, we obtain

$$G_Z(d) = G_X(d) + \mathbb{E}[x_i y_{i-d}^\top] + \mathbb{E}[y_i x_{i-d}^\top] + G_Y(d)$$

For uncorrelated white inputs, the first and second terms are 0, and the third term becomes $a_d \Sigma A^\top$, and so we obtain $G_Z(d) = a_d A\Sigma + g_a(d) A\Sigma A^\top$. In either case, the similarity structure $G_Y(d)$ is carried into the residual stream, and the resulting similarity structure depends on its interactions with the existing similarity structure $G_X(d)$.

Effect of normalization. The actual normalization operation, whether this is the (numerically stable, scaled by $\sqrt{D}$) L2 normalization for RMSNorm or first subtracting the channel mean for LayerNorm, has little effect on similarity structure. RMSNorm in fact exactly preserves the unit-normalized cross-moment structure; and, though LayerNorm additionally projects prior to L2 normalization, prior work has found that they behave very similarly in practice (Zhang & Sennrich, 2019; Gupta et al., 2026).[1]

In contrast, the gain (and offset for LayerNorm; though we omit that here for simplicity) of the normalization module can alter the similarity structure, including its floor and gap. Here, let X denote the output of the normalization stage. Decompose X into its variation around a position-independent mean $\bar{x} = \mathbb{E}[x_i]$, so that each $x_i = \bar{x} + \epsilon_i$ where $\mathbb{E}[\epsilon_i] = 0$. For per-channel gains $D_\gamma = \mathrm{diag}(\gamma)$ and output $Y = D_\gamma X$, we have

$$\mathbb{E}[y_i y_{i-d}^\top] = \underbrace{D_\gamma \bar{x}\bar{x}^\top D_\gamma}_{\text{shared-mean contribution}} + \underbrace{D_\gamma \mathbb{E}[\epsilon_i \epsilon_{i-d}^\top] D_\gamma}_{\text{fluctuation cross-moment}} .$$

Write $M = D_\gamma \bar{x}\bar{x}^\top D_\gamma$ for the shared-mean contribution and $e_M = \mathrm{tr}(M) = \|D_\gamma \bar{x}\|^2$ for its energy, and $C(d) = D_\gamma \mathbb{E}[\epsilon_i \epsilon_{i-d}^\top] D_\gamma$ for the fluctuation cross-moment and $e_C = \mathrm{tr}(C(0)) = \mathbb{E}[\|D_\gamma \epsilon_i\|^2]$ for its energy. Assuming approximately stationary second moments and concentrated norms, the unit-normalized similarity gap and far-distance floor are

$$G_Y(d) - G_Y(L) \approx \frac{C(d) - C(L)}{e_M + e_C}, \qquad G_Y(L) \approx \frac{M + C(L)}{e_M + e_C}.$$

How the normalization module's gain affects the similarity gap and floor (and in turn, the cosine similarity, or another projection or magnitude of the cross-moment) depends on its effect on the shared-mean contribution and fluctuation cross-moments and their energies. The gain's effect on each of these terms depends in turn on how the channel-pair entries are reweighted by $\gamma_c \gamma_k$, with the energies depending on the diagonal weights γ_c^2. In practice, it appears that the broad pattern is that the normalization's gain reduces the shared-mean energy relative to the fluctuation energy; when the fluctuation profile $C(d)/e_C$ is approximately preserved, this can reduce the cosine similarity floor while amplifying the similarity gap.

Effect of the MLP. For normalized inputs, a wide MLP at random initialization admits a simple kernel description of its effect on cosine similarity (Daniely et al., 2016); i.e., at initialization an MLP preserves the ordering of the recency-bias profile. However, this description does not address how the MLP evolves during training nor its effect on the profile's floor, magnitude, or shape.

Suppose the MLP branch is $y = W_2 f(W_1 x)$, with wide random Gaussian W_1, W_2, and consider two inputs x, x' with cosine similarity p. For each hidden unit, the corresponding preactivations

[1] In particular, Gupta et al. (2026) found that the residual streams of transformers with LayerNorm are approximately orthogonal to the one-vector in the models studied, which they show to be the normal vector of the projection prior to normalization.

$U = w^\top x$ and $V = w^\top x'$ are jointly Gaussian with correlation p. We therefore reduce the problem to asking how the nonlinearity f transforms this correlation, defining the resulting kernel $k_f(p)$; the subsequent projection W_2 approximately preserves this kernel at large width. Expanding f in the Gaussian Hermite basis gives

$$k_f(p) = \frac{\sum_n a_n^2 p^n}{\sum_n a_n^2}.$$

The $n = 0$ term contributes a constant similarity floor, the $n = 1$ term transmits similarity linearly, and higher-order terms nonlinearly reshape the profile. Since the coefficients enter as a_n^2, k_f is nondecreasing for $p \geq 0$. Thus, at random initialization, a wide MLP preserves the ordering of a nonnegative recency-bias profile, although it may change its floor, magnitude, and shape.

This initialization argument does not guarantee preservation by trained MLPs. Depending on the projections and nonlinearities learned in training, the MLP can suppress, amplify, or reorient the incoming similarity structure.

Decomposing the general similarity structure. For each of the above modules, we study their transformation of the full lagged cross-moment. Central to our understanding of this general structure is that different slices of it reduce to the cosine similarity and per-head logit curves, via the Frobenius inner product. Just as a vector inner product depends on both magnitude and alignment, it is instructive to decompose the Frobenius inner product into the contributions of cross-moment magnitude and its vectorized alignment with either the isotropic component I or the composed query-key projection $M = W_Q^\top W_K / \sqrt{D_H}$.

For nonzero matrices G and X of the same shape, the Frobenius inner product decomposes as

$$\langle G, X \rangle_F = \|G\|_F \|X\|_F \cos(G, X), \qquad \cos(G, X) = \frac{\operatorname{tr}(G^\top X)}{\|G\|_F \|X\|_F}.$$

Hence, the matrix cosine is just the ordinary cosine between the vectorized matrices: the two norms set the scale, while their alignment determines the sign.

To see how this pertains to a recency bias, let $B_X = G_X(1) - G_X(L)$, and $B_Z = G_Z(1) - G_Z(L)$ for the actual post-normalization input Z to global attention. Taking $X = I$ or $X = M$ gives

$$c(1) - c(L) = \sqrt{D}\, \|B_X\|_F \cos(B_X, I),$$
$$\ell(1) - \ell(L) = \|B_Z\|_F \|M\|_F \cos(B_Z, M).$$

Here $\|I\|_F = \sqrt{D}$, and $\|M\|_F$ is the head's readout gain. Therefore, the recency gaps for both cosine similarity and the global NoPE logits depend on the underlying matrix lag contrast of the cross-moment, which is unsigned (i.e., a large magnitude alone does not imply a positive cosine or logit gap). Meanwhile, its vectorized cosine alignment determines the sign.

We therefore have an understanding of how each module transforms the cross-moment structure, and how that corresponds to cosine similarity or logits. Next, we empirically assess these effects.

C.2 Empirics

Measuring the per-module effect on similarity gap and floor. We observed in Section C.1 that residual stream addition, normalization, and MLPs each have a complex effect on the full cross-moment similarity structure, including both its gap and floor. To empirically characterize these effects, Figure 10 measures the difference in the cosine gap and floor, before and after each module (averaged across each of its appearances in depth; the grey block summaries show the cumulative effect across each block).

Proceeding through each module appearance from left to right, we generally observe the following trends: (i) normalization, including its learned affine transformation, reduces the cosine floor and increases the gap; (ii) SWA provides the strongest boost to the gap (and interestingly, reduces the floor); (iii) the residual addition reduces the gap and increases the floor; (iv) the MLP reduces the gap and increases the floor; and (v) global NoPE attention also increases the gap, though much less so than SWA. The cumulative effect across each block is much smaller, with full-attention and MLP blocks largely preserving the incoming gap.

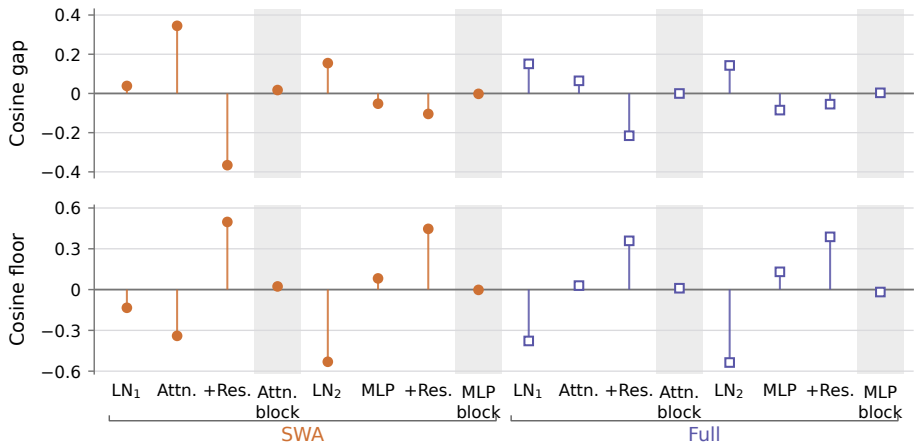


Figure 10: *Module changes in cosine gap and floor.* After-minus-before changes in cosine gap $c(1) - c(4096)$ (top) and cosine floor $c(4096)$ (bottom), averaged across SWA and full-attention blocks at $w = 128$. Residual-addition steps compare with the isolated branch; shaded block summaries compare the complete update with its incoming residual. Note that the effect of each module on the similarity structure is largely as predicted by our theory.

Cross-moment magnitude and readout alignment. For a broader view on the cross-moment magnitude and its alignment to both the isotropic component and per-head projections, Figure 11 measures the matrix lag contrast B of the cross-moment and its cosine similarity with I and M across depth. Indeed, as is expected from the separate analyses of residual stream cosine and the logits, the total similarity magnitude increases through depth along with its alignment to I (which together imply the increased recency bias through depth we see in Section 4); meanwhile, the global NoPE projections maintain a positive alignment with B that remains relatively constant, producing positive and consistent logit gaps across depth (as seen in Section 5).

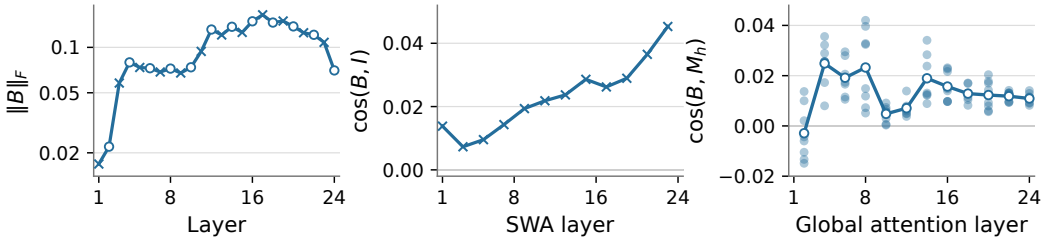


Figure 11: *Cross-moment magnitude and readout alignment.* Final 350M RoPE-in-SWA model ($w = 128$). Left: $\|B_X\|_F$ at post-block residual states; crosses mark SWA and circles global NoPE layers. Middle: $\cos(B_X, I)$ at SWA layers. Right: $\cos(B_Z, M)$ at model-scale post-LN1 inputs to global heads; small points show eight heads and connected circles their unweighted mean. Magnitude uses a logarithmic axis; alignments use linear axes.

Depth-wise and window-size dependence. For a more fine-grained view of depth-wise patterns than provided in Figure 6, Figure 12 shows the mean logit profile at every global NoPE layer. As noted in Section 5.2, the pattern is different than for the residual stream recency bias: after an initial rise, the logit recency bias is largely consistent across layers. Similarly, comparing the logit profiles for different window sizes, Figure 13 shows a weaker pattern between recency bias and window size than in the residual stream. The relative lack of variation in the logit profiles with depth and window

size suggests that, as we discussed in Section 5.2, the model seeks to develop a consistent logit-based recency bias to serve as its position encoding in attention.

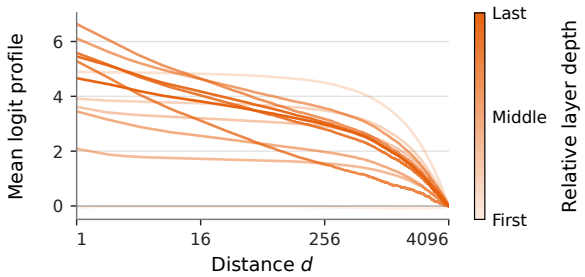


Figure 12: *Global NoPE logits across depth.* 350M head-mean logit gap profiles $\bar{\ell}(d) - \bar{\ell}(4096)$ at the final $w = 128$ checkpoint. Darker orange denotes deeper global layers.

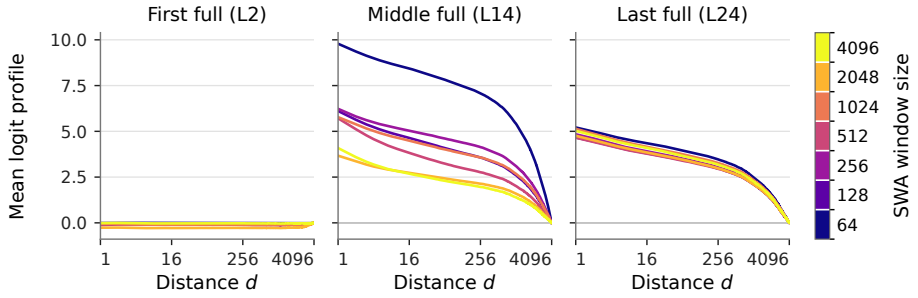


Figure 13: *Global NoPE logits across window sizes.* 350M head-mean logit gap profiles $\bar{\ell}(d) - \bar{\ell}(4096)$ at the final first, middle, and last global layers.

D CONTROLS

D.1 120M-SCALE ROBUSTNESS

To verify that the 350M-scale results of the main text are robust across model scales, we train 120M models with the same configurations of SWA window size and varying RoPE and NoPE within the window, on the same dataset and for a number of tokens comparable given the scale difference (see Appendix A). These 120M-scale results corroborate the main-text 350M-scale results, with the same qualitative conclusions from each panel provided in the composite of Figure 14. That is, SWA's recency bias is present from initialization (a), strengthens with training and depth (b–c), becomes sharper with smaller windows (d–e), and propagates to and is selected by the logits (f–i).

Furthermore, Figure 15 displays the cosine gap for NoPE-in-SWA, as well as a comparison of the validation losses for both RoPE and NoPE within the window. Of particular note is that, in both cases, loss improves with smaller windows despite fewer training flops, indicating that smaller windows may be beneficial, perhaps due to their ability to resolve a strong recency bias. In further support of this, with NoPE-in-SWA loss collapses entirely for larger windows; this is suggestive that, without a sufficiently strong architectural mechanism to resolve a recency bias and therefore encode position, such models fail to learn language effectively.

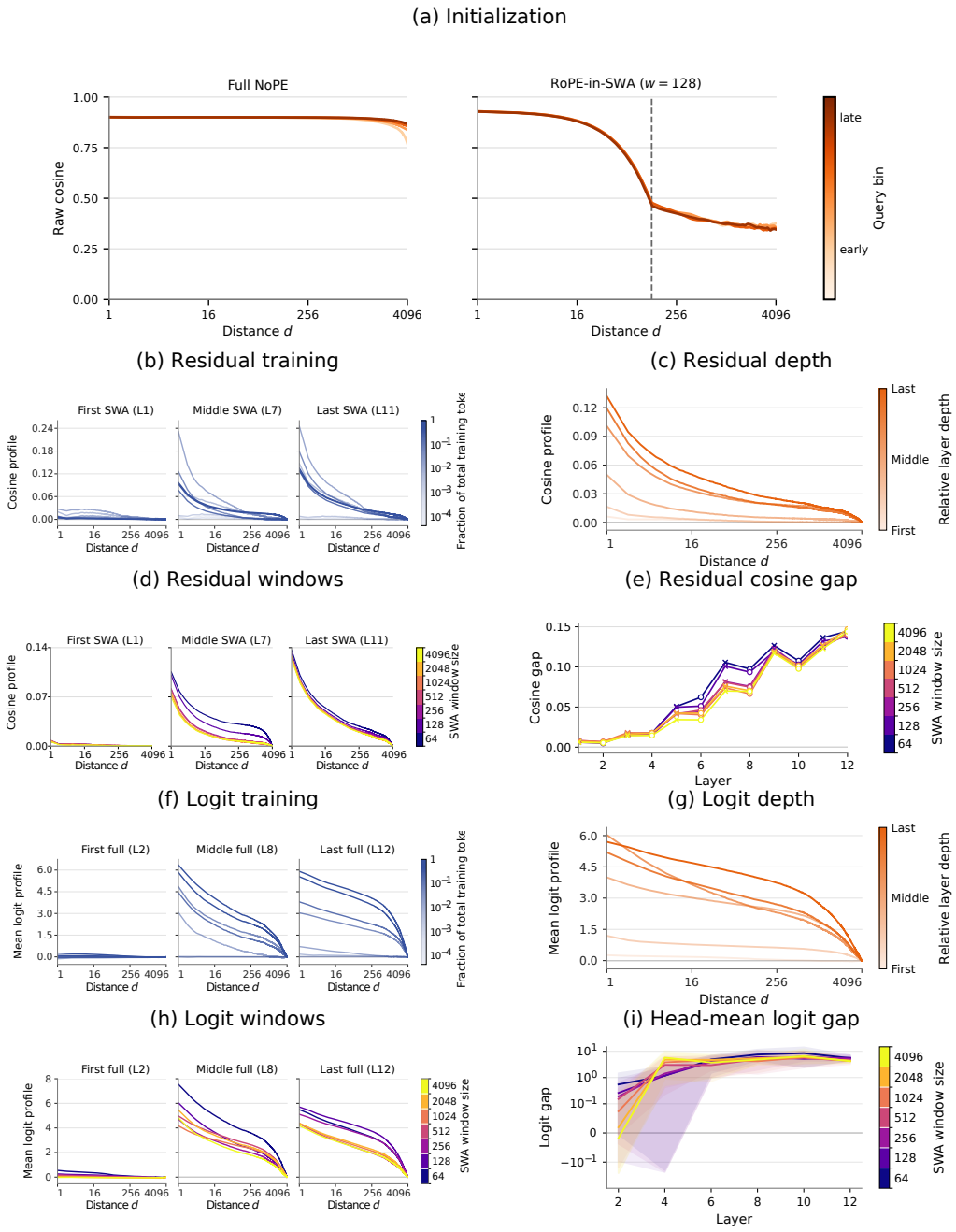


Figure 14: *120M TextbookChapters, RoPE within SWA.* (a) Raw cosine of the isolated first attention-branch outputs at initialization for global NoPE and SWA ($w = 128$), split by query-position bin. (b–e) Post-attention residual cosine during training, across SWA depth and window sizes, and the depthwise gap $c(1) - c(4096)$. (f–i) Global-NoPE logit profiles during training, across depth and windows, and the head-mean logit gap across depth. Cosine and logit profiles show gaps to $d = 4096$; window panels cover sizes 64–4096. Blue denotes training progress, orange denotes depth, and multicolor curves denote window size. In (i), shading is the full range across heads, not uncertainty.

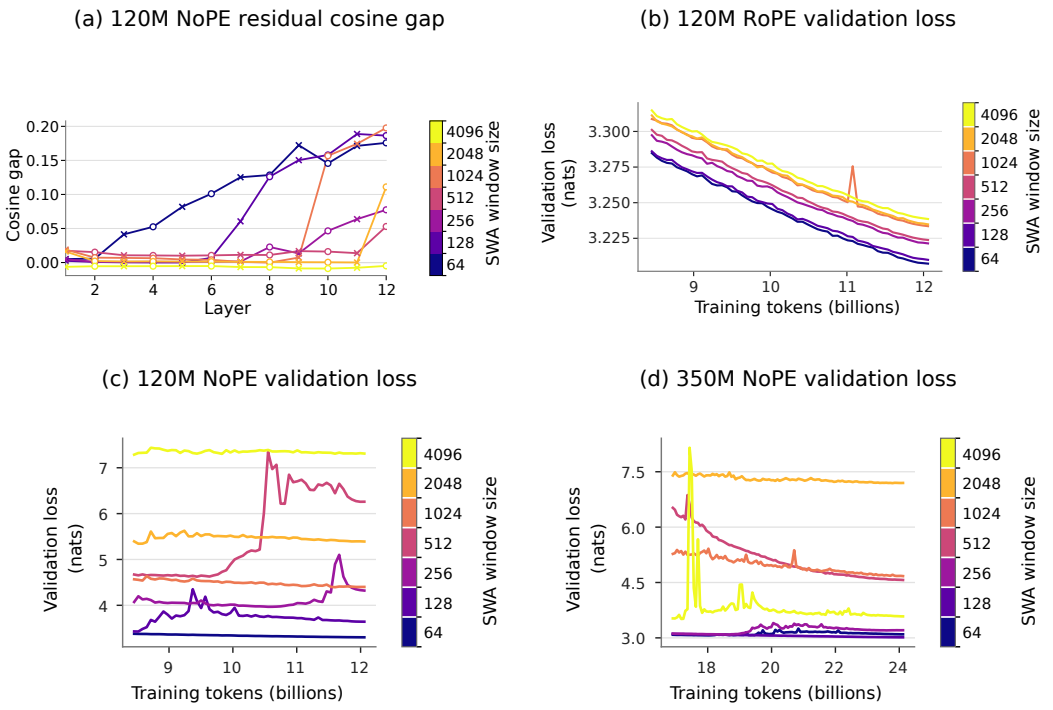


Figure 15: *Natural-language NoPE comparison and validation loss.* (a) The 120M NoPE-in-SWA residual cosine gap $c(1) - c(4096)$ across depth for seven window sizes; crosses mark SWA layers and circles mark global NoPE layers. (b–d) Mean DCLM token cross-entropy (nats) over the final 30% of training on 4096-token sequences: 120M RoPE-in-SWA (b), 120M NoPE-in-SWA (c), and 350M NoPE-in-SWA (d). Window size changes several model properties, so these loss differences do not isolate the contribution of recency.

D.2 Random-token controls

The main analyses use natural-language sequences, as these are the natural domain of the language models we train. However, language has underlying correlation structure that may confound the evidence for our proposed architectural mechanism. Hence, repeating the complete SWA analysis on uniformly sampled tokens from the model's vocabulary is an important control to isolate the structural contribution from natural-language statistics. The following controls use the same query positions and lags as the TextbookChapters analyses and report results for both RoPE and NoPE within SWA (Figures 16–19; each composite groups the observables for one SWA encoding and model scale).

These controls also corroborate the main-text findings. Across all four composites, we observe the recency bias present at initialization and strengthening with training and depth (a–c). SWA's position encoding had a larger impact on the effect of window size (d–e): for NoPE-in-SWA at both model scales, we observe the same pattern of smaller windows promoting stronger recency biases; however, for RoPE-in-SWA at both scales, the trends were inconclusive, likely due random tokens altering the learned RoPE decay profile. Finally, across all four settings, the recency bias transfers to the logits particularly for smaller windows, but this transfer is less consistent than in natural language (f–i). Overall, these results are consistent with our proposed architectural mechanism functioning but being untuned for data outside its training distribution.

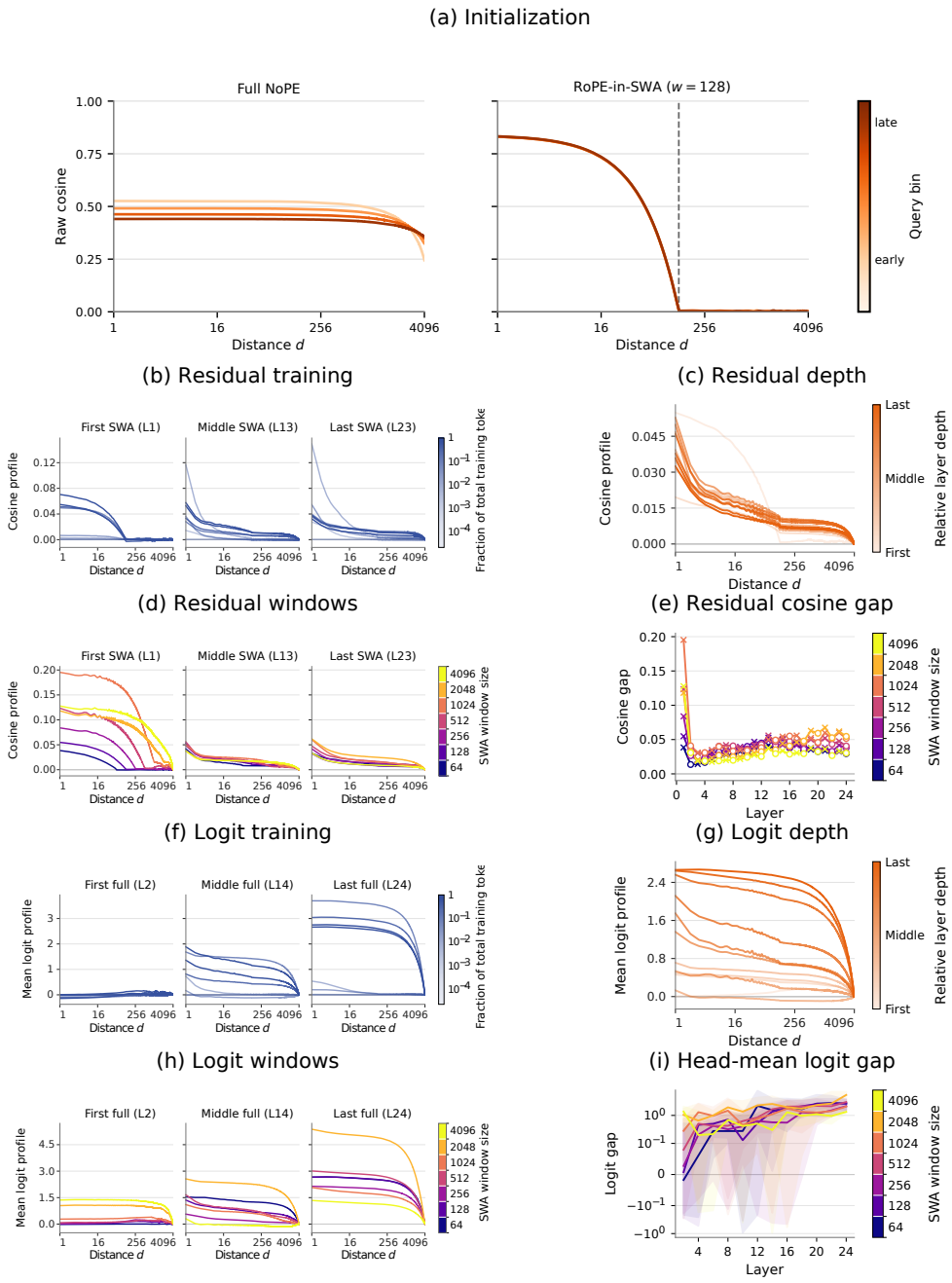


Figure 16: *Random-token control: 350M, RoPE within SWA.* (a) Raw cosine of isolated first attention-branch outputs at initialization for global NoPE and SWA ($w = 128$), split by query-position bin. (b–e) Post-attention residual cosine during training, across SWA depth and window sizes, and the depthwise gap $c(1) - c(4096)$. (f–i) Global-NoPE logit profiles during training, across depth and windows, and the head-mean logit gap across depth. Cosine and logit profiles show gaps to $d = 4096$; window panels cover sizes 64–4096. Blue denotes training progress, orange denotes depth, and multicolor curves denote window size. In (i), shading is the full range across heads, not uncertainty.

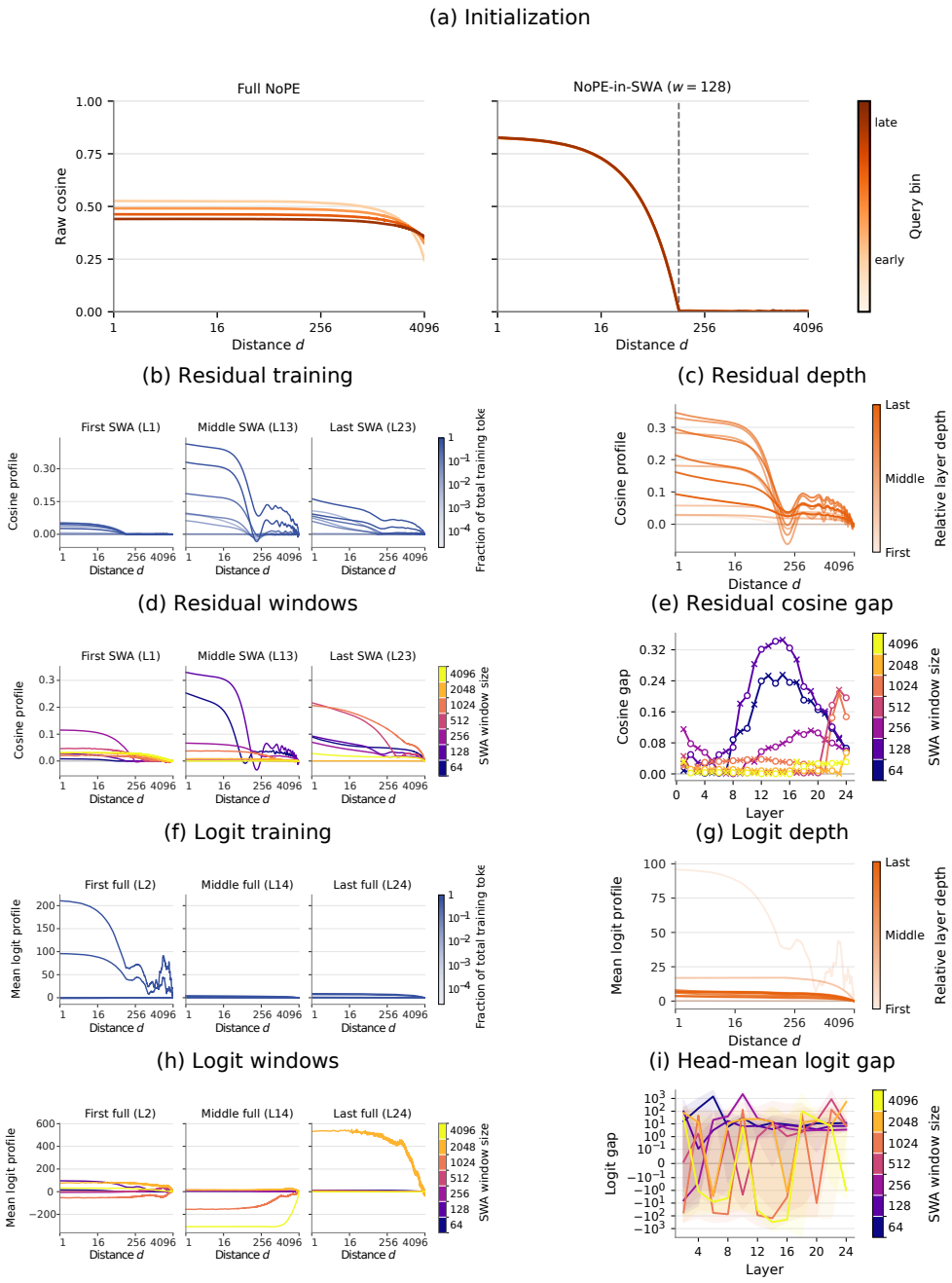


Figure 17: *Random-token control: 350M, NoPE within SWA.* (a) Raw cosine of isolated first attention-branch outputs at initialization for global NoPE and SWA ($w = 128$), split by query-position bin. (b–e) Post-attention residual cosine during training, across SWA depth and window sizes, and the depthwise gap $c(1) - c(4096)$. (f–i) Global-NoPE logit profiles during training, across depth and windows, and the head-mean logit gap across depth. Cosine and logit profiles show gaps to $d = 4096$; window panels cover sizes 64–4096. Blue denotes training progress, orange denotes depth, and multicolor curves denote window size. In (i), shading is the full range across heads, not uncertainty.

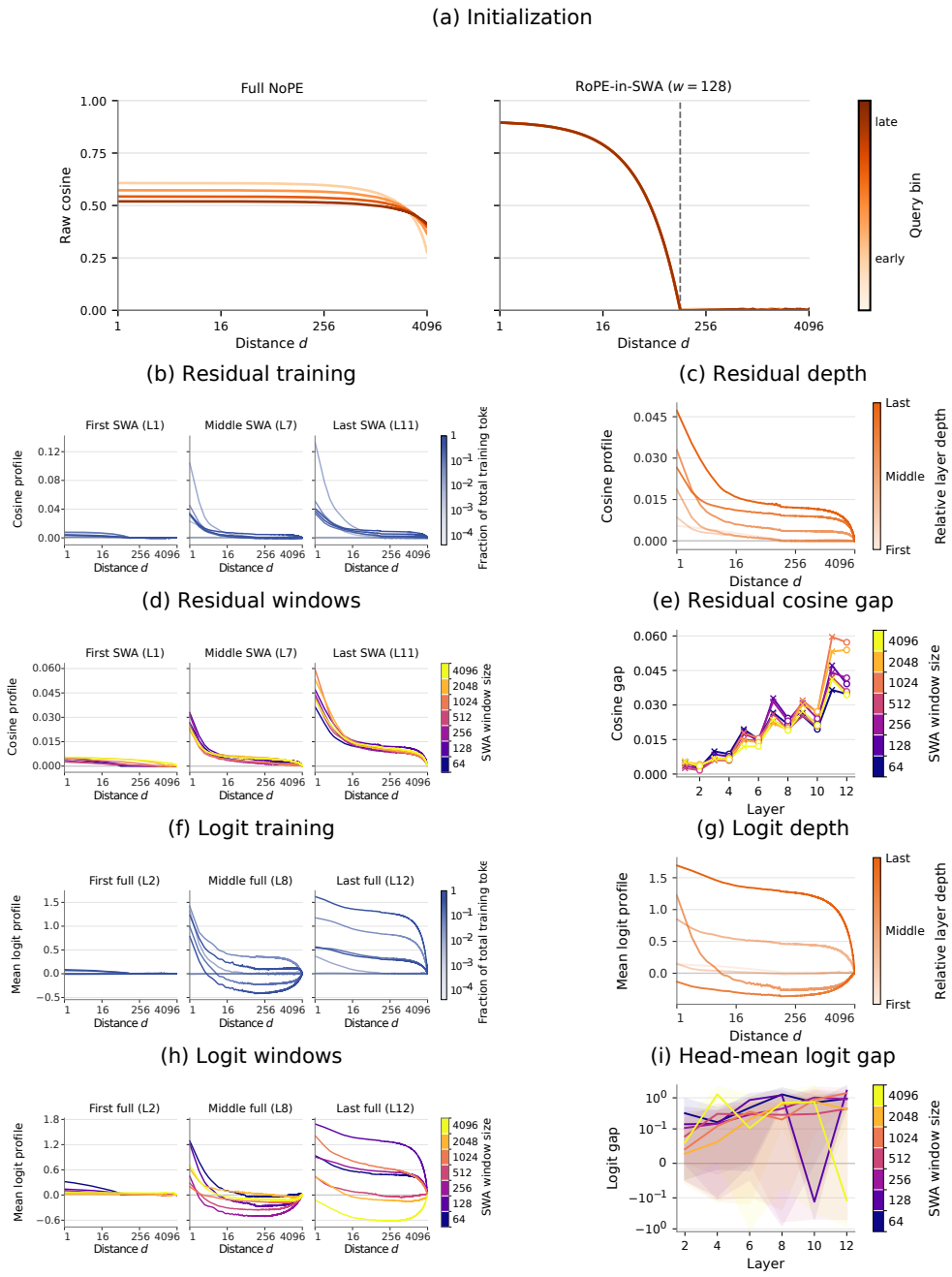


Figure 18: *Random-token control: 120M, RoPE within SWA.* (a) Raw cosine of isolated first attention-branch outputs at initialization for global NoPE and SWA ($w = 128$), split by query-position bin. (b–e) Post-attention residual cosine during training, across SWA depth and window sizes, and the depthwise gap $c(1) - c(4096)$. (f–i) Global-NoPE logit profiles during training, across depth and windows, and the head-mean logit gap across depth. Cosine and logit profiles show gaps to $d = 4096$; window panels cover sizes 64–4096. Blue denotes training progress, orange denotes depth, and multicolor curves denote window size. In (i), shading is the full range across heads, not uncertainty.

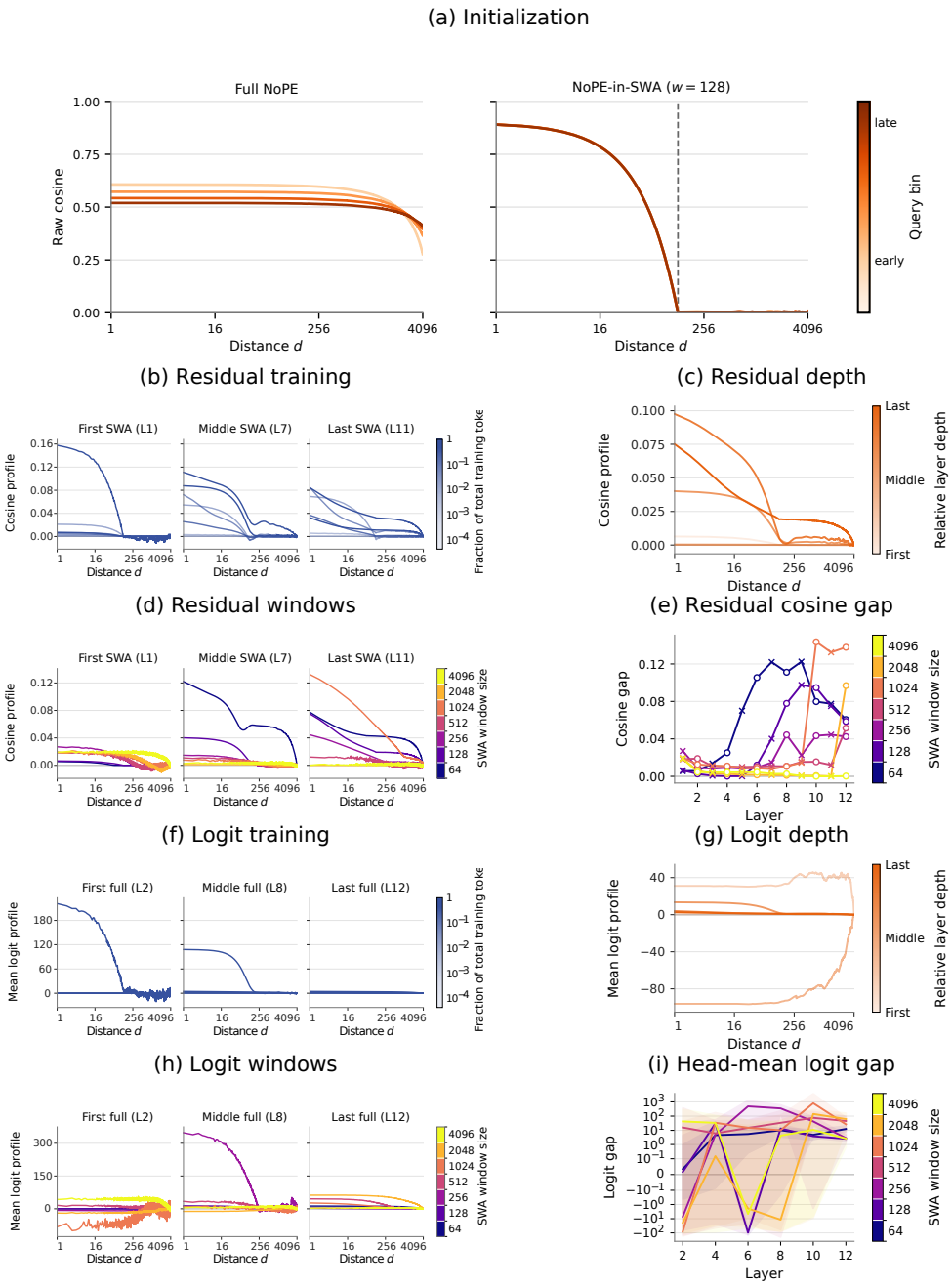


Figure 19: *Random-token control: 120M, NoPE within SWA.* (a) Raw cosine of isolated first attention-branch outputs at initialization for global NoPE and SWA ($w = 128$), split by query-position bin. (b–e) Post-attention residual cosine during training, across SWA depth and window sizes, and the depthwise gap $c(1) - c(4096)$. (f–i) Global-NoPE logit profiles during training, across depth and windows, and the head-mean logit gap across depth. Cosine and logit profiles show gaps to $d = 4096$; window panels cover sizes 64–4096. Blue denotes training progress, orange denotes depth, and multicolor curves denote window size. In (i), shading is the full range across heads, not uncertainty.

E KDA EXTENSION

E.1 THEORY

KDA updates its recurrent state as

$$S_t = (I - \beta_t k_t k_t^\top)\,\mathrm{Diag}(\alpha_t) S_{t-1} + \beta_t k_t v_t^\top.$$

Here, α_t controls how much of each state channel is retained, while the delta update revises the association with the current key. A write from position i reaches position t through the intervening transition products. We now use a simplified approximation to show how this repeated mixing can produce a recency bias, following the same intuition as for SWA.

Intuition. Consider a simplified scenario in which the projected values o_t are zero-mean, of similar magnitude, and uncorrelated across positions, and the effective contribution of a past value decreases by a fixed factor $0 < \lambda < 1$ at each step. Averaging over the content-dependent gates, writes, and query readout, we approximate the attention output before normalization and gating as an exponential moving average,

$$y_t \approx \lambda y_{t-1} + (1-\lambda) o_t = (1-\lambda) \sum_{r=0}^{\infty} \lambda^r o_{t-r},$$

where the infinite sum describes positions sufficiently far from the start of the sequence. As in SWA, nearby outputs are similar because they combine many of the same values; take for example y_t and y_{t-d}. In particular, all values contributing to y_{t-d} also contribute to y_t, but with their weights reduced by λ^d. Hence, if output norms are concentrated, their expected cosine similarity is approximately

$$c_{\mathrm{KDA}}(d) \approx \lambda^d.$$

Thus, the same shared-values argument that gives a triangular profile for uniform SWA gives an exponential profile in this simplified KDA case.

Length scaling. The retention factor plays a role analogous to the SWA window size: a smaller λ produces a narrower, steeper recency profile, while a larger λ retains correlations over longer distances. Critically, like SWA, this scale is length-independent and therefore remains equally-resolved with sequence length.

Correlated inputs. As in the general SWA case, the incoming similarity structure also affects the output. Let $G_O(d)$ and $G_Y(d)$ denote the lagged cross-moments of the unit-normalized projected values and outputs, respectively. Assume norms are concentrated, then for the same exponential weights $a_r = (1-\lambda)\lambda^r$, the filter calculation then gives the approximate relation

$$G_Y(d) \propto \sum_{k \in \mathbb{Z}} g_a(k) G_O(d-k), \qquad g_a(k) = \frac{1-\lambda}{1+\lambda} \lambda^{|k|}.$$

Thus, the incoming structure is mixed across distances by a smoothly decaying overlap profile. As in SWA, the subsequent transformations determine how this structure enters the residual stream and whether later NoPE projections read it out as larger logits for recent keys.

E.2 EMPIRICS

We empirically test whether our proposed mechanism extends to KDA-interleaved models at different scales, when inputted both natural language data and random-token controls. Figures 20 and 21 below provide a subset of panels (those relevant to KDA) from previous composites; each row pairs both model scales, and all logit panels use the head-mean convention specified in Appendix A.3. Indeed, the same empirical pattern appears: KDA layers primarily increase residual stream recency, increasingly with training and depth (a–d); and later NoPE layers learn recency-biased logits (e–h). This indicates that the general mechanism we identify is not specific to SWA, but can apply to many local sequence mixers including the variety of increasingly-common linear gated attention types.

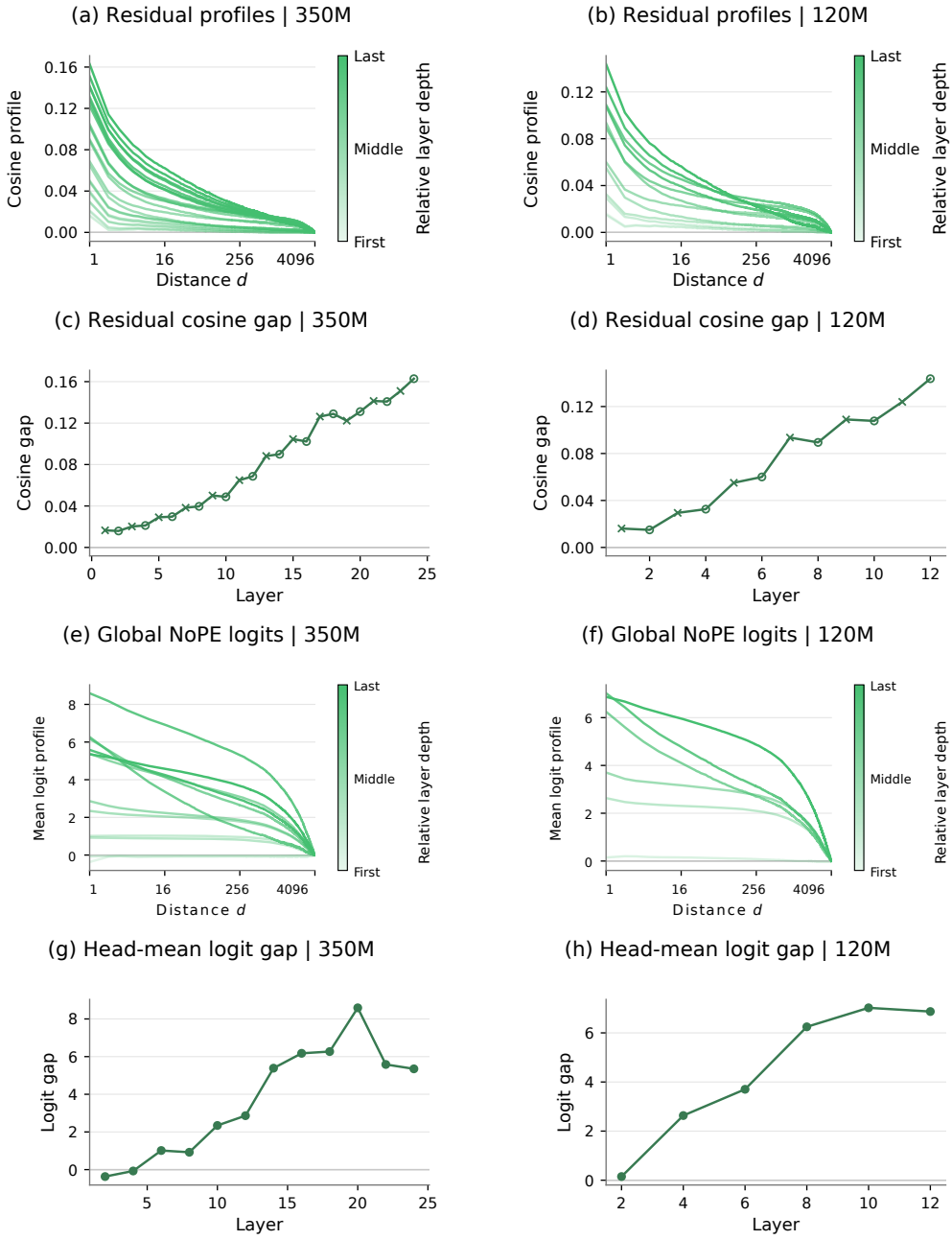


Figure 20: *KDA–NoPE natural-language extension.* Each row pairs 350M (left) and 120M (right). (a–b) Post-attention residual cosine profiles across KDA and global NoPE depth, shown as gaps $c(d) - c(4096)$; darker green denotes greater depth. (c–d) Residual cosine gap $c(1) - c(4096)$ after each attention update. (e–f) Head-mean global-NoPE logit gap profiles $\bar{\ell}(d) - \bar{\ell}(4096)$. (g–h) Gap of the head-mean logit profile $\bar{\ell}(1) - \bar{\ell}(4096)$.

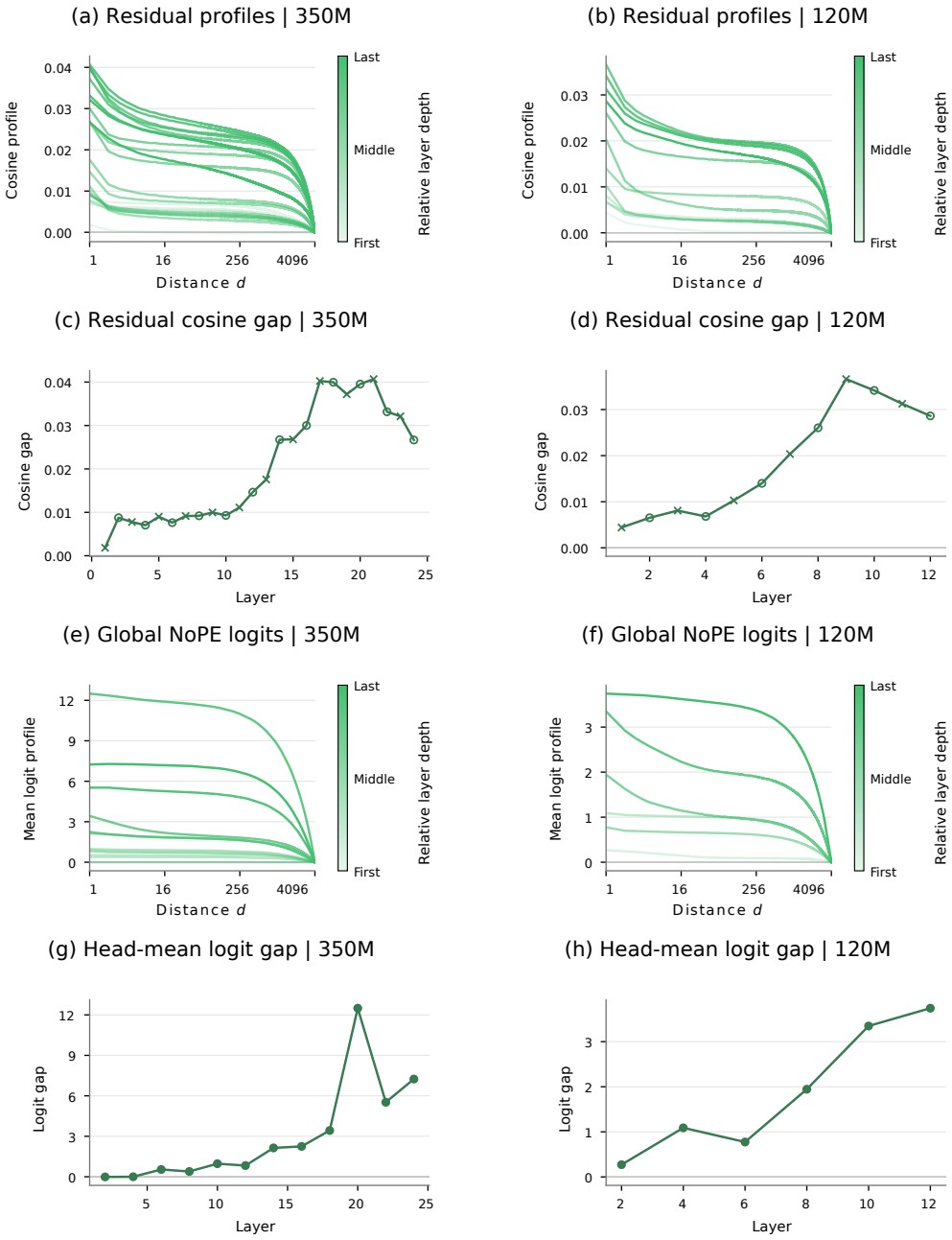


Figure 21: *KDA–NoPE random-token extension.* The same 350M (left) and 120M (right) checkpoints are evaluated on random tokens sampled from their vocabulary. (a–b) Residual cosine profiles after KDA and global NoPE updates across depth, shown as gaps $c(d) - c(4096)$; darker green denotes greater depth. (c–d) Residual cosine gap $c(1) - c(4096)$ after each attention update. (e–f) Head-mean global-NoPE logit gap profiles $\bar{\ell}(d) - \bar{\ell}(4096)$. (g–h) Gap of the head-mean logit profile $\bar{\ell}(1) - \bar{\ell}(4096)$.