MASSIVE ACTIVATIONS IN HYBRID LINEAR ATTENTION LARGE LANGUAGE MODELS: PRE-ATTENTION SPIKES AND INTER-SPIKE PLATEAUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid linear attention large language models (HLA LLMs) interleave linear and full attention to balance efficiency and capacity, yet how hybridization reshapes internal activation dynamics remains poorly understood. We present the first systematic study of massive activations (MAs) in layer-interleaved HLA LLMs, examining their *architectural organization*, *training-time emergence*, *underlying mechanisms*, and *functional significance*. Across five linear attention architectures, six hybridization configurations, and five input domains, we identify two architecture-aligned morphologies: *pre-attention spikes* (PAS) immediately before full attention and *inter-spike plateaus* (ISP) persisting through intervening linear attention layers. Denser full attention increasingly connects PAS through ISP, approaching the persistent MAs of conventional Transformers. This organization also recurs across 12 public checkpoints spanning 1.2B–397B parameters, covering linear attention and state-space hybrids. Controlled pretraining of Gated DeltaNet (GDN) hybrids up to 1.3B reveals early emergence and consolidation of both morphologies, alongside asymmetric gating effects. Specifically, full attention output gates strongly attenuate MA magnitudes without eliminating their organization, whereas removing GDN output gates yields modest amplification. Mechanistically, we develop a shared systematic-outlier account: PAS follows a localized write–sink–cancel process, while ISP is consistent with delayed cancellation. Functionally, our interventions show that deleting only the four largest-magnitude PAS coordinates at each full attention input reduces mean downstream accuracy by 21.9%–63.6% relative to normal inference. Moreover, reference-conditioned spike-to-plateau connection consistently improves mean real-world retrieval accuracy, yielding relative gains of 1.1%–12.6% without retraining. Anonymized code is provided in the supplementary material.

1 INTRODUCTION

Softmax attention enables Transformers to model expressive, content-dependent interactions across the entire context (Vaswani et al., 2017; Team et al., 2025c; 2026c;b; 2025b). However, its computational cost grows quadratically with context length, making long-context modeling expensive (Tay et al., 2022; Su et al., 2026a; 2025a; Su & Yuan, 2025). Linear attention reduces this cost through recurrent computation, but its fixed-size state restricts modeling capacity, particularly for recall-intensive tasks (Yang et al., 2023; 2025b). Hybrid linear attention (HLA) LLMs address this trade-off by interleaving linear and full attention layers. This design is increasingly common in modern LLMs, including Qwen3-Next, Qwen3.5, Kimi Linear, Kimi K3, Zamba, and Nemotron-H (Yang et al., 2025a; Team et al., 2025a; 2026a; Blakeman et al., 2025; Glorioso et al., 2024).

Despite their growing adoption, how HLA architectures reshape internal activation dynamics remains poorly understood. Of particular interest are *massive activations* (MAs), a sparse set of hidden-state entries that exceed typical values by several orders of magnitude and concentrate at specific token positions (Sun et al., 2024). Prior work attributes both their emergence and their close coupling with attention sinks to structural properties of softmax-based Transformers (Su et al., 2026b; Xiao et al., 2024; Sun et al., 2026; An et al., 2025; Su & Yuan, 2025). These findings establish MAs as a distinctive signature of Transformer dynamics and an informative probe of layerwise

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.

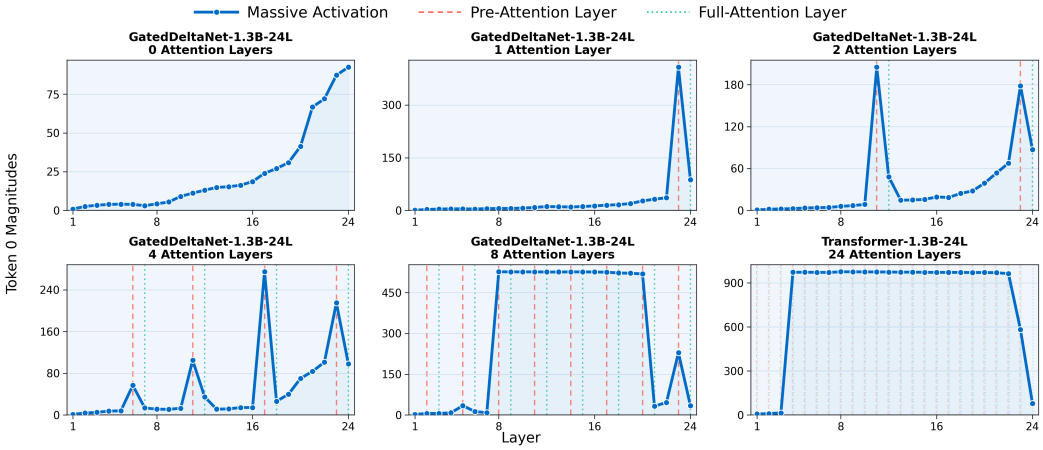


Figure 1: **MA morphology across GDN hybrid configurations.** MAs form *pre-attention spikes* (PAS) immediately before full attention layers. Denser full attention promotes persistence between successive PAS, forming *inter-spike plateaus* (ISP). Models are from the M-A-P Hybrid Linear Attention Research suite (Wang et al., 2025).

computation. Yet their behavior beyond conventional full attention LLMs remains largely unexplored, raising a central question: ***How does hybridization shape MA dynamics in HLA LLMs, and what do these dynamics reveal about their internal computation?***

In this work, we systematically investigate MAs in layer-interleaved HLA LLMs. Through extensive experiments across five linear attention architectures, six hybridization configurations, and five input domains (§3.2–§3.3), we uncover a hybridization-dependent organization distinct from that of full attention LLMs. As illustrated in Figure 1, MAs intensify immediately before full attention layers, forming *pre-attention spikes* (PAS). Denser full attention promotes persistence between PAS, producing *inter-spike plateaus* (ISP). At the full attention limit, these forms merge into the characteristic Transformer morphology, with relatively stable MAs across most intermediate layers. This organization recurs across 12 public checkpoints from Kimi Linear, Qwen3.5, Nemotron-H, and Zamba2, spanning 1.2B–397B parameters and covering both linear attention and state-space hybrids, including base and instruction-tuned variants (§3.4).

Beyond inference-time characterization, we trace the emergence of PAS and ISP through controlled pretraining of Gated DeltaNet (GDN) hybrids and examine the effects of output gating (§3.5). We further apply systematic-outlier analysis to their cross-layer evolution, linking PAS and ISP to the stable MA morphology of full attention models (§4.1–§4.2). Finally, targeted interventions demonstrate the functional importance of PAS and the potential of spike-to-plateau connection to improve real-world retrieval without retraining (§4.3). Our main contributions are:

- **Architecture-aligned MA morphology.** We present the first systematic study of MAs in layer-interleaved HLA LLMs, identifying PAS and ISP as two architecture-aligned morphologies that recur across architectures, hybridization configurations, input domains, and pretrained hybrids spanning 1.2B–397B parameters.
- **Controlled emergence and gating.** Pretraining GDN hybrids up to 1.3B shows that PAS and ISP emerge early and consolidate during optimization. Full attention output gating strongly attenuates their magnitudes but preserves their layerwise organization, whereas removing GDN output gates yields modest amplification.
- **A shared lifecycle explanation.** We develop a shared systematic-outlier account organized by MA cancellation timing. PAS follows a localized *write–sink–cancel* process, while the extended persistence of ISP is consistent with delayed cancellation; this account connects both morphologies to the stable MAs observed at the full attention limit.
- **Functional importance and real-world retrieval improvement.** Through interventions across six frozen GDN hybrids and eleven benchmarks, we establish the functional importance of PAS and show that reference-conditioned spike-to-plateau connection consistently improves mean real-world retrieval performance without retraining.

2 PRELIMINARIES ON HYBRID LINEAR ATTENTION MODELS

Hybrid Residual Blocks. For a length-T input with hidden width d, an L-block pre-normalized HLA model updates the residual stream $\mathbf{X}^{(0)} \in \mathbb{R}^{T\times d}$ as

$$\mathbf{H}^{(\ell)} = \mathbf{X}^{(\ell-1)} + \mathrm{Mixer}^{(\ell)}\left(\mathrm{Norm}^{(\ell)}_{\mathrm{mix}}\left(\mathbf{X}^{(\ell-1)}\right)\right), \tag{1}$$

$$\mathbf{X}^{(\ell)} = \mathbf{H}^{(\ell)} + \mathrm{FFN}^{(\ell)}\left(\mathrm{Norm}^{(\ell)}_{\mathrm{ffn}}\left(\mathbf{H}^{(\ell)}\right)\right), \tag{2}$$

where $\mathbf{X}^{(\ell)}$ is the residual stream after block ℓ. The mixer is full attention for $\ell \in \mathcal{I}_{\mathrm{FA}} \subseteq \{1,\dots,L\}$ and linear attention otherwise. Letting $L_{\mathrm{FA}} = |\mathcal{I}_{\mathrm{FA}}|$, we define the *hybridization ratio* as $\rho = L/L_{\mathrm{FA}}$: a ρ:1 configuration contains one full attention layer per ρ sequence-mixing layers. Larger ρ corresponds to sparser full attention, whereas $\rho = 1$ recovers a full attention model.

Full Attention Layers. Omitting multi-head notation, full attention computes

$$[\mathbf{Q},\mathbf{K},\mathbf{V}] = \mathbf{X}[\mathbf{W}_Q,\mathbf{W}_K,\mathbf{W}_V], \qquad \mathrm{FA}(\mathbf{X}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}} + \mathbf{M}\right)\mathbf{V}\mathbf{W}_O, \tag{3}$$

where d_h is the head dimension and $\mathbf{M}$ the causal mask. Full attention enables direct, content-dependent interactions with preceding tokens but incurs quadratic computation in T.

Linear Attention Layers. Linear attention replaces pairwise token interactions with a fixed-size recurrent state. A canonical recurrent formulation is

$$[\mathbf{q}_t,\mathbf{k}_t,\mathbf{v}_t] = \mathbf{x}_t[\mathbf{W}_Q,\mathbf{W}_K,\mathbf{W}_V], \qquad \mathbf{S}_t = \mathbf{F}_t\mathbf{S}_{t-1} + \mathbf{k}_t^\top\mathbf{v}_t, \qquad \mathbf{y}_t = \mathbf{q}_t\mathbf{S}_t\mathbf{W}_O, \tag{4}$$

where $\mathbf{S}_t$ is the recurrent state and $\mathbf{F}_t$ governs state retention. Variants differ in their state-transition and update rules, but all maintain states whose size is independent of sequence length, enabling linear-time sequence processing without a growing KV cache.

By interleaving full and linear attention across depth, HLA LLMs combine complementary sequence-mixing mechanisms whose interaction may reorganize internal computation across layers. How this interaction shapes activation dynamics remains poorly understood, motivating our study of the layerwise organization of MAs. A broader discussion of linear and hybrid linear attention, together with related work on MAs in LLMs, is provided in Appendix A.

3 MASSIVE ACTIVATION DYNAMICS UNDER HYBRIDIZATION

3.1 EXPERIMENTAL SETUP AND SINK-GUIDED TRACKING

Analysis Roadmap. We first characterize PAS across linear attention architectures (§3.2) and ISP across hybridization ratios (§3.3). Next, we examine whether this organization recurs in large-scale linear attention and state-space hybrids (§3.4). Finally, controlled pretraining traces the emergence of both morphologies and isolates their response to output gating (§3.5).

Models and Inputs. We use the M-A-P suite (Wang et al., 2025): RetNet, HGRN, Gated Linear Attention (GLA), DeltaNet, and Gated DeltaNet (GDN), at 340M and 1.3B under a unified pretraining recipe.[1] These backbones span fixed-decay, input-dependent gating, and delta-rule updates (Sun et al., 2023; Qin et al., 2023; Yang et al., 2023; 2024; 2025b). We use *"Summer is warm. Winter is cold."* as a running example. For cross-domain comparisons, we evaluate 100 inputs each from general prose in WikiText-103, scientific writing in Scientific Papers, mathematical reasoning in GSM8K, Python code in CodeSearchNet, and multilingual text in FLORES-200 (Merity et al., 2016; Cohan et al., 2018; Cobbe et al., 2021; Husain et al., 2019; Costa-Jussà et al., 2022). Checkpoint and pretraining details appear in Appendix B.1.

[1] M-A-P Hybrid Linear Attention Research model collection on Hugging Face.

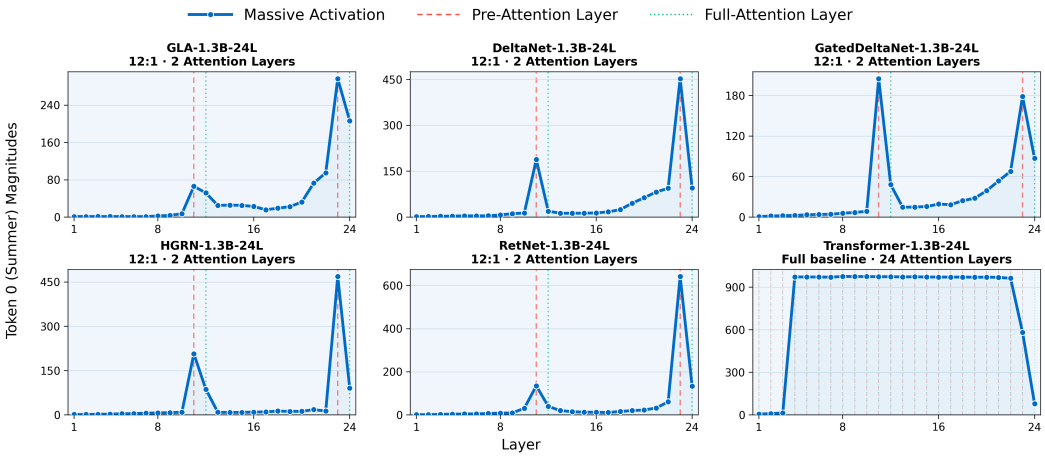


Figure 2: **PAS across linear attention architectures.** First-token MA trajectories for 1.3B M-A-P hybrids at 12:1, with a full attention Transformer for reference. All five hybrid backbones exhibit activation maxima immediately before full attention.

Sink-Guided Tracking of MAs. Prior work traces MAs through layerwise magnitude ranking, with highly activated tokens closely corresponding to attention sinks in full attention models (Sun et al., 2024; An et al., 2025). Under hybridization, however, sink tokens exhibit lower and less stable magnitude rankings, while the maximally activated token switches more frequently across layers. Magnitude ranking alone can therefore conflate trajectories of different tokens. We complement it with an attention-derived anchor that preserves token identity across depth.

For each input x, let $A^{(\ell,h)}_{x,q,t}$ denote the attention probability assigned by query token q to source token t in head h of full attention layer ℓ. We select the dominant *consensus sink* as

$$t^\star_x = \arg\max_{1\le t<T} \frac{1}{|\mathcal{I}_{\mathrm{FA}}|H|\mathcal{Q}_t|} \sum_{\ell\in\mathcal{I}_{\mathrm{FA}}} \sum_{h=1}^{H} \sum_{q\in\mathcal{Q}_t} A^{(\ell,h)}_{x,q,t}, \qquad \mathcal{Q}_t = \{t+1,\dots,T\}, \tag{5}$$

where H is the number of attention heads. Averaging across layers and heads reduces local fluctuations, while normalization by $|\mathcal{Q}_t|$ accounts for differences in the number of eligible queries. Holding this token fixed, we trace its maximum absolute activation as $m^{(\ell)}_{x,t^\star_x} = \|\mathbf{X}^{(\ell)}_{x,t^\star_x,:}\|_\infty = \max_j |X^{(\ell)}_{x,t^\star_x,j}|$. This preserves token identity while allowing the maximally activated feature to change across depth. The attention distribution determines only which token is tracked; whether and how strongly that token exhibits an MA remains determined by its activation magnitude. Further details and diagnostic evidence are provided in Appendix B.2.

3.2 Pre-Attention Spikes across Linear Attention Architectures

Recurring Pre-Attention Spikes. To isolate the effect of the linear attention mechanism, we fix model scale and hybridization ratio at 1.3B and 12:1. For qualitative comparison, we visualize first-token trajectories for the running example, using the most prevalent sink position as a representative anchor. Despite differing recurrent updates, all five backbones exhibit pronounced activation maxima immediately before full attention (Figure 2). We term these architecture-aligned local maxima *pre-attention spikes* (PAS). Beyond this example, Appendix B.3 provides token-wise analyses, while Appendix B.4 documents recurrence across five domains and both model scales.

Pre-Attention Localization. We quantify alignment using the input-specific consensus sink $t^\star_x$ from §3.1. For each full attention layer f, let $\mathcal{B}_f$ denote the contiguous preceding linear attention block, including $f-1$. For an evaluation set $\mathcal{D}$, the *sink–spike alignment rate* is

$$\mathrm{Align}(\mathcal{D}) = \frac{1}{|\mathcal{D}|\,|\mathcal{I}_{\mathrm{FA}}|} \sum_{x\in\mathcal{D}} \sum_{f\in\mathcal{I}_{\mathrm{FA}}} \mathbf{1}\left[f-1 \in \arg\max_{\ell\in\mathcal{B}_f} m^{(\ell)}_{x,t^\star_x}\right], \tag{6}$$

Table 1: **Sink–spike alignment (%).** 12:1 hybrids; macro-average over five domains with 100 inputs each.

Backbone	1.3B	340M
RetNet	99.9	100.0
HGRN	100.0	100.0
GLA	100.0	99.6
DeltaNet	100.0	99.4
GDN	100.0	100.0

Table 2: **Inter-spike retention (%).** Each entry reports 1.3B/340M, averaged over the same five domains.

Backbone	**12:1**	**6:1**	**3:1**
RetNet	18.8 / 36.8	60.0 / 69.8	85.4 / 87.2
HGRN	7.2 / 5.0	40.2 / 49.6	92.5 / 81.2
GLA	27.3 / 31.9	51.7 / 61.6	88.0 / 88.1
DeltaNet	23.1 / 46.5	44.6 / 82.8	84.4 / 99.8
GDN	18.4 / 39.7	26.6 / 45.2	77.8 / 86.7

where $\mathbf{1}[\cdot]$ is the indicator function. This measures the fraction of input–attention-layer pairs with blockwise maxima immediately before full attention. Across 500 inputs (100 per domain), domain-macro-averaged alignment reaches 99.4–100% for all ten backbone–scale pairs at 340M and 1.3B (Table 1). Table 7 in Appendix B.2 provides domain-specific results. We assess peak strength through prominence, comparing consensus sinks with first-token and non-sink anchors; a random-layer baseline tests localization. Paired bootstrap analyses support stronger localization and prominence for consensus sinks than non-sink controls (Appendix B.6).

3.3 From Spikes to Plateaus across Hybridization Ratios

To isolate hybridization effects, we fix the backbone and model scale while varying $\rho \in \{24, 12, 6, 3\}$. Pure linear attention and full attention provide limiting reference configurations.

From Localized Spikes to Sustained Plateaus. Figure 1 illustrates this progression in GDN. PAS persists across hybrid configurations, but inter-spike activity changes with attention density. Sparse full attention produces sharply localized peaks separated by activation troughs. Denser full attention elevates inter-spike activations, connecting successive PAS into sustained *inter-spike plateaus* (ISP). The onset is backbone-dependent: prominent ISP appears at 6:1 in DeltaNet but primarily at 3:1 in GDN (Appendix B.5). Thus, attention density organizes the progression, with backbone-dependent transition points. At the full attention limit, spikes and plateaus merge into stable MAs across most intermediate layers, linking hybrid morphologies to those of conventional Transformers.

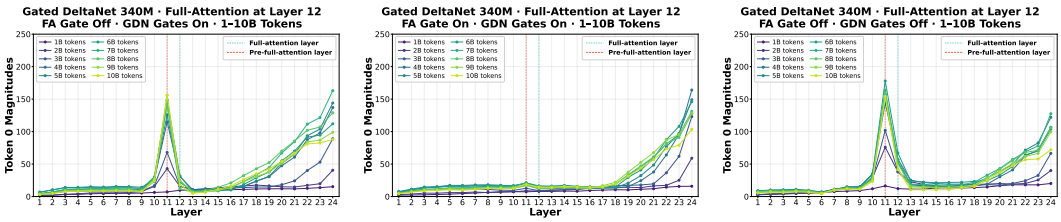
Inter-Spike Persistence. To quantify persistence, let $\mathcal{I}_{\mathrm{FA}} = \{f_1 < \cdots < f_K\}$ denote the full attention layers, and let $p_i = f_i - 1$ denote their input positions. We define $\mathcal{J}_i = \{\ell : p_i < \ell < p_{i+1},\ \ell \notin \mathcal{I}_{\mathrm{FA}}\}$ as the intervening linear-attention layers, excluding the output of the full-attention block at f_i. Using the fixed consensus sink from §3.1, we define the *inter-spike retention score* as

$$\mathrm{ISR}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \frac{1}{K-1} \sum_{i=1}^{K-1} \frac{1}{|\mathcal{J}_i|} \sum_{\ell \in \mathcal{J}_i} \min\left(1,\ \frac{m^{(\ell)}_{x,t^\star_x}}{\min\left(m^{(p_i)}_{x,t^\star_x}, m^{(p_{i+1})}_{x,t^\star_x}\right)}\right). \tag{7}$$

ISR normalizes activation at the intervening linear-attention layers by the weaker of the two PAS endpoints, p_i and p_{i+1}; values near zero indicate decay and near one sustained persistence. Clipping at one prevents above-reference activations from dominating the score. The metric requires at least two PAS and a nonempty $\mathcal{J}_i$ for every adjacent pair; it is inapplicable to single-spike configurations. On the same 500 inputs, ISR increases with full attention density across all five backbones and both scales (Table 2). Paired 95% bootstrap confidence intervals for all 20 adjacent-ratio differences lie above zero. To distinguish persistence from endpoint normalization effects, Appendix B.6 reports absolute inter-spike activation.

3.4 Recurrence in Large-Scale Pretrained Hybrid Models

We evaluate 12 public checkpoints from Kimi Linear, Qwen3.5, Nemotron-H, and Zamba2, spanning 1.2B–397B parameters (Team et al., 2025a; Yang et al., 2025a; Blakeman et al., 2025; Glorioso et al., 2024). These families cover linear attention and state-space mixers, periodic and nonuniform attention schedules, and base and instruction-tuned variants. Kimi Linear combines Kimi Delta Attention with periodic global MLA layers; Qwen3.5 uses three GDN layers followed by one output-gated full attention layer ($\rho = 4$ under our definition). Nemotron-H combines Mamba-2 with sparse, partly nonuniform self-attention, whereas Zamba2 periodically inserts shared Transformer



(a) Standard GDN. (b) Full attention output gating. (c) GDN output gates removed.

Figure 3: **PAS emergence and gating response.** First-token trajectories at successive 1B-token checkpoints in a 340M GDN model with full attention at layer 12. PAS emerges early and strengthens; full attention output gating attenuates but does not eliminate it, whereas removing GDN output gates moderately amplifies it.

blocks into a Mamba-2 backbone. These designs test whether PAS–ISP organization extends beyond a particular recurrent mechanism, ungated attention, or periodic layout.

Consistent Organization, Variable Magnitude. First, matched Kimi Linear Base and Instruct checkpoints retain closely aligned PAS locations and ISP boundaries despite magnitude differences, showing persistence across post-training. Second, across Qwen3.5 scales, PAS and ISP track full attention placement even as magnitude and prominence vary non-monotonically. Layerwise organization is thus more stable than amplitude across these checkpoints. Third, recurrence in Nemotron-H and Zamba2 extends the pattern to state-space hybrids with sparse and partly nonuniform attention schedules. Fourth, across representative inputs from five domains, PAS locations and ISP spans remain stable while magnitudes vary. Together, these representative traces extend the architectural pattern beyond the M-A-P suite. Appendices C.1 and C.2 provide configurations and all checkpoint and domain traces.

3.5 Controlled Pretraining Studies of PAS and ISP

Controlled Pretraining Models. Inference-time recurrence leaves open when these morphologies arise and how architectural choices shape their development. Using Flash Linear Attention (Yang & Zhang, 2024), we pretrain 24-layer GDN hybrids from scratch at 340M and 1.3B. Within each comparison, only the targeted layer placement or gating condition changes; all other architectural and training settings remain fixed (Appendix D.1).

Training-Time Emergence of PAS and ISP. To isolate PAS, we place a single full attention layer at an early, middle, or late depth in 340M models and repeat the middle placement at 1.3B. For ISP, we train both scales at 3:1. We trace first-token MA trajectories every 1B tokens at 340M and every 5B at 1.3B. Across these configurations, PAS consistently forms immediately before full attention and shifts with its placement. Under matched training conditions, this relocation links PAS to full attention placement rather than a preferred absolute depth. PAS is visible by 1B tokens in the 340M layer-12 configuration (Figure 3a) and becomes more pronounced and stable. ISP likewise emerges early and consolidates during optimization (Figure 12a). Both morphologies therefore develop during pretraining, rather than being artifacts of checkpoint selection (Appendices D.3 and D.4).

Asymmetric Effects of Output Gating. Recent work reports that element-wise output gating can eliminate MAs and their associated sinks in full attention Transformers (Su et al., 2026b; Qiu et al., 2026). We test this in hybrids by adding output gates to full attention layers and removing native GDN output gates. Full attention output gating markedly attenuates PAS, although a small pre-attention peak persists throughout training (Figure 3b). Removing GDN output gates produces only moderate amplification, despite modifying more layers (Figure 3c). ISP exhibits the same asymmetry: full attention gating substantially reduces the plateau, whereas removing GDN gates has a comparatively modest effect (Figure 12; Appendix D.5). Crucially, full attention output gating attenuates MA magnitudes without eliminating their architecture-aligned organization: weak PAS and ISP persist and become more visible during training. Consistently, Qwen3.5 models with native full attention output gating retain both morphologies (Figures 19c–19f; Appendix C.2).

4 A SYSTEMATIC-OUTLIER ACCOUNT OF PAS AND ISP

The preceding analyses establish MA locations and training-time emergence; we now examine their cross-layer mechanisms and functional significance. Building on systematic-outlier analysis (An et al., 2025; Su & Yuan, 2025; Su et al., 2025b), we trace signed residual updates and verify a representative PAS at a fixed token–feature coordinate. Together with recurring cross-model patterns, this supports a shared lifecycle distinguished by cancellation timing. We examine localized PAS and persistent ISP, then test how spike suppression and extended persistence affect performance.

4.1 PRE-ATTENTION SPIKES: LOCALIZED WRITE–SINK–CANCEL

Localized Write–Sink–Cancel. Across backbones and hybridization configurations, PAS appears as a coordinated cross-layer event rather than an isolated numerical fluctuation. In a representative 1.3B GDN hybrid at 12:1, large outliers recur immediately before full attention and dissipate afterward (Figure 4). Signed decomposition reveals recurring write–cancel pairs; fixed-coordinate tracing then tests whether the same feature is written and cancelled.

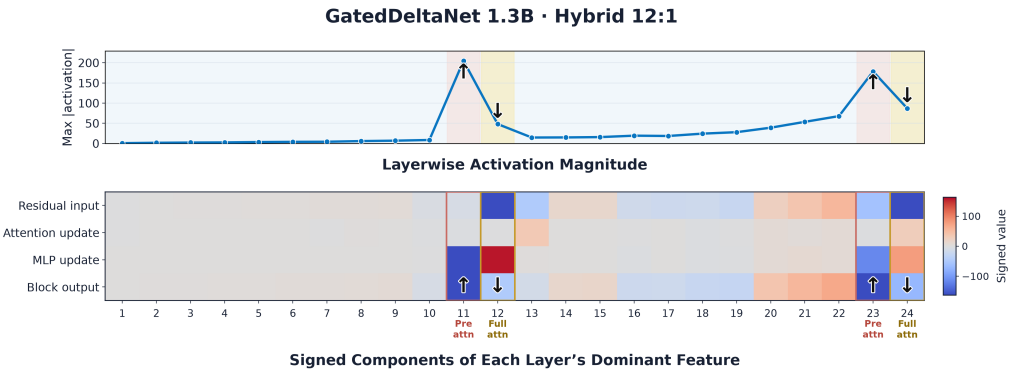


Figure 4: **Localized write–sink–cancel lifecycle.** A 1.3B GDN hybrid at 12:1 writes an outlier before full attention and cancels it with an opposite-signed update.

Fixed-Coordinate Verification. We therefore examine PAS preceding $f = 12$ in the same model. Fixing the consensus sink $t^\star$ and feature $j^\star = \arg\max_j |X^{(f-1)}_{t^\star,j}|$, we trace signed activations and module-level updates at this coordinate. Unlike the morphological envelope, this trace follows one feature throughout. The residual decomposition separates the incoming state, mixer update, FFN update, and block output. The signs of these components distinguish reinforcement from cancellation. This trace identifies three stages: layer 11 writes the outlier; during layer-12 full attention, its token attracts a disproportionate share of attention from later queries and acts as a sink; an opposite-signed update in that layer substantially cancels the same outlier. Signed residual addition verifies cancellation at this fixed coordinate. Together, these stages verify a localized *write–sink–cancel* event linking residual dynamics to attention-sink behavior. Appendix E.1 visualizes each stage; Appendix E.2 extends the layerwise analysis to M-A-P and large pretrained linear attention and state-space hybrids.

4.2 INTER-SPIKE PLATEAUS: EXTENDED OUTLIER PERSISTENCE

Extended Persistence through Delayed Cancellation. Signed ISP analysis reveals the same write–cancel pattern over greater depth. Outlier magnitudes rise immediately before full attention and remain elevated across intervening linear attention layers, with opposite-signed updates as the plateau dissipates (Figure 5). Persistence followed by an opposite-signed update supports two regimes of a shared lifecycle: prompt cancellation localizes PAS, whereas delayed cancellation is consistent with sustained ISP (cross-model traces in Appendix E.2).

A Continuum toward Full Attention. The hybridization trend in §3.3 admits the same interpretation. As full attention becomes denser, elevated inter-spike activations bridge adjacent PAS,

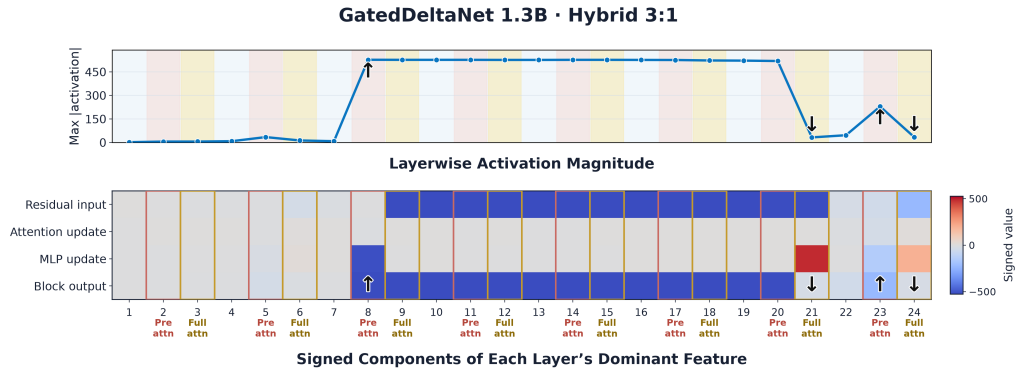


Figure 5: **ISP dynamics consistent with delayed cancellation.** Elevated activations persist between full attention layers, with opposite-signed updates as the plateau dissipates. Signed components are evaluated at each layer's dominant feature.

consistent with extended persistence. At the full attention limit, the distinction between spikes and intervening plateaus disappears, recovering stable MAs across much of the network depth. PAS, ISP, and persistent full attention MAs thus form a morphological continuum that the cancellation-timing account helps interpret. Here, persistence refers to the residual activation trajectory across *depth*. We next test how altering attention-input outliers and their persistence affects model behavior.

4.3 Functional Interventions: Removing Spikes and Extending Persistence

We investigate the functional significance of PAS–ISP organization through two complementary interventions: suppressing PAS at full attention inputs and connecting successive spikes into sustained plateaus. We evaluate six released M-A-P GDN checkpoints spanning two scales (340M and 1.3B) and three hybridization ratios ($\rho \in \{12, 6, 3\}$), with model weights and attention-layer placement fixed. Both interventions use the same examples and scoring rules as normal inference across eleven benchmarks: ten accuracy tasks and one perplexity benchmark. Real-world retrieval, including SQuAD, is evaluated by cloze-completion accuracy. Table 3 summarizes results; Appendix F provides complete protocols and model-specific scores.

Table 3: **PAS suppression and spike-to-plateau connection across six frozen GDN hybrids.** $k = 4$ coordinates are selected per full attention input. Scores are averaged equally across models; parentheses report relative changes from normal inference, $100(\bar{S}_{\text{int}}/\bar{S}_{\text{normal}} - 1)$. Accuracy is in percent (↑); perplexity is reported separately (↓). The ten-task mean weights accuracy tasks equally, excluding perplexity; the real-world retrieval mean averages FDA, SWDE, SQuAD, and NQ. Complete model-specific results for all k appear in Appendix F.2.

Benchmark	Normal	PAS suppression	Spike connection
PIQA	70.13	66.78 (−4.77%)	70.10 (−0.04%)
ARC-Easy	65.55	56.40 (−13.95%)	65.38 (−0.27%)
HellaSwag	51.11	45.71 (−10.57%)	51.07 (−0.08%)
FDA	45.37	15.32 (−66.23%)	46.05 (+1.50%)
SWDE	35.57	16.06 (−54.85%)	37.72 (+6.06%)
SQuAD	37.78	32.02 (−15.26%)	42.19 (+11.66%)
NQ	22.42	14.14 (−36.91%)	22.37 (−0.19%)
NIAH-1	71.00	19.33 (−72.77%)	71.13 (+0.19%)
NIAH-2	80.43	24.27 (−69.83%)	81.13 (+0.87%)
NIAH-3	36.93	4.27 (−88.45%)	37.13 (+0.54%)
WikiText-103 PPL ↓	11.6290	20.1000 (+72.84%)	11.6238 (−0.04%)
Ten-task mean	51.63	29.43 (−43.00%)	52.43 (+1.55%)
Real-world retrieval mean	35.28	19.38 (−45.06%)	37.08 (+5.10%)

Suppressing PAS. Following our first-token analyses, both interventions target the first prompt token. For each example, let $p_i = f_i - 1$ index the residual state entering full attention layer f_i, and let $z^{(\ell)} = X^{(\ell)}_{1,:}$ denote its normal reference trajectory before normalization. At every full attention input, we select $C_i = \mathrm{TopK}(|z^{(p_i)}|, k)$, with $k \in \{1, 2, 4\}$, and set $\widetilde{X}^{(p_i)}_{1,j} = 0$ for $j \in C_i$. Coordinate selections are fixed from the reference pass; all other entries remain unedited. Deleting only four coordinates per full attention input reduces ten-task mean accuracy in all six models, with relative losses of 21.91%–63.60%. The cross-model mean falls from 51.63% to 29.43%, accompanied by increased perplexity. Retrieval is particularly sensitive: cross-model relative losses reach 66.23% on FDA and 88.45% on NIAH-3, compared with 4.77% on PIQA. These results demonstrate that model performance depends strongly on the selected first-token PAS coordinates.

Connecting Spikes without Retraining. We next extend persistence from the first to last full attention input on the union $C = \bigcup_i C_i$. Here, k specifies the per-layer selection budget, so C may contain more than k distinct coordinates. Between adjacent inputs, we set

$$\widetilde{X}^{(\ell)}_{1,j} = s_{i,j} \max\left\{ |z^{(p_i)}_j|,\ s_{i,j} z^{(\ell)}_j \right\}, \qquad p_i < \ell < p_{i+1}, \quad j \in C, \tag{8}$$

where $s_{i,j} = \mathrm{sign}(z^{(p_i)}_j)$, with $+1$ at zero. The left reference endpoint sets a signed magnitude floor, while larger same-direction reference values are retained. At each full attention input, we restore the selected coordinates to their reference values, $\widetilde{X}^{(p_i)}_{1,j} = z^{(p_i)}_j$ for $j \in C$; unselected coordinates evolve under the intervention. This design extends inter-spike activity while preserving the selected PAS endpoint values. A normal forward pass supplies the reference trajectory, and the same rule is applied across examples without answer-based selection or parameter updates. Across all eighteen model–budget settings, the equal-weight mean accuracy over FDA, SWDE, SQuAD, and NQ improves by 1.06%–12.60% relative to normal inference, with exploratory paired 95% confidence intervals for the gain above zero (Appendix F.2, Table 13). At $k = 4$, the cross-model retrieval mean rises from 35.28% to 37.08%, a 5.10% relative gain. SQuAD and SWDE improve in eighteen and sixteen settings, respectively. At $k = 4$, both improve in all six models, with cross-model relative gains of 11.66% and 6.06%. Task-level regressions occur on FDA, NQ, and NIAH-3, while changes on commonsense benchmarks are comparatively small. Appendix F.3 reports supplementary Qwen3.5-27B results on reasoning, mathematics, and code generation.

Connecting Morphology to Function. Together, these interventions demonstrate the functional relevance of MA organization: suppressing selected PAS coordinates can impair performance, whereas reference-conditioned spike-to-plateau connection improves mean real-world retrieval without retraining. One possible explanation is that extending MA persistence brings hybrid activation dynamics closer to the persistent-MA morphology of full attention Transformers. These gains motivate refining connection strength and persistence depth and developing reference-free variants, with the goal of improving retrieval while reducing inference overhead.

5 Conclusion

We identify PAS and ISP as recurring, architecture-aligned MA morphologies in hybrid LLMs. Controlled pretraining reveals early emergence and asymmetric gating effects that alter magnitude while preserving layerwise organization. A shared systematic-outlier account links localized write–sink–cancel dynamics in PAS with ISP persistence consistent with delayed cancellation. Crucially, suppressing PAS coordinates can impair performance, whereas reference-conditioned spike-to-plateau connection consistently improves mean real-world retrieval in the evaluated GDN hybrids without retraining. Together, these results connect MA organization to model behavior and motivate further study of cancellation depth, effective connection rules, and reference-free retrieval improvement.

ETHICS STATEMENT

This work analyzes activation dynamics in language models using existing checkpoints and benchmark datasets. The additional interventions do not involve human participants or collect personal information. Benchmark scores characterize the evaluated tasks, not model safety or suitability for deployment. AI assistance is disclosed separately below.

REPRODUCIBILITY STATEMENT

The appendix documents the evaluated checkpoints, datasets, analysis metrics, controlled-pretraining configurations, hyperparameters, and benchmark protocols used in this work. Appendix F specifies the functional interventions, evaluation populations, and aggregation rules. The anonymized supplementary material includes the analysis and intervention code, evaluation configurations, frozen manifests, and numerical evidence needed to regenerate the new tables.

AI USE STATEMENT

The authors led the development of the core research ideas, scientific observations, and experiments. OpenAI Codex served as an auxiliary tool, used for text polishing, code debugging, and support with evaluation workflows and consistency checks. The research direction, interpretation, and scientific claims remained with the authors. All reported measurements derive from executed model evaluations. The authors reviewed the AI-assisted work and take full responsibility for the final manuscript, code, results, and scientific claims.

REFERENCES

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Systematic outliers in large language models. *arXiv preprint arXiv:2502.06415*, 2025.

Simran Arora, Aman Timalsina, Aaryan Singhal, Benjamin Spector, Sabri Eyuboglu, Xinyi Zhao, Ashish Rao, Atri Rudra, and Christopher Ré. Just read twice: closing the recall gap for recurrent language models. *arXiv preprint arXiv:2407.05483*, 2024.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 7432–7439, 2020.

Aaron Blakeman, Aarti Basant, Abhinav Khattar, Adithya Renduchintala, Akhiad Bercovich, Aleksander Ficek, Alexis Bjorlin, Ali Taghibakhshi, Amala Sanjay Deshmukh, Ameya Sunil Mahabaleshwarkar, et al. Nemotron-h: A family of accurate and efficient hybrid mamba-transformer models. *arXiv preprint arXiv:2504.03624*, 2025.

Ruisheng Cao, Mouxiang Chen, Jiawei Chen, Zeyu Cui, Yunlong Feng, Binyuan Hui, Yuheng Jing, Kaixin Li, Mingze Li, Junyang Lin, et al. Qwen3-coder-next technical report. *arXiv preprint arXiv:2603.00729*, 2026.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 615–621, 2018.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Jorge Gallego-Feliciano, S Aaron McClendon, Juan Morinelli, Stavros Zervoudakis, and Antonios Saravanos. Hidden dynamics of massive activations in transformer training. *arXiv preprint arXiv:2508.03616*, 2025.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Zenodo*, 2021.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, Anna Golubeva, Vasudev Shyam, James Whittington, Jonathan Pilault, and Beren Millidge. The zamba2 suite: Technical report. *arXiv preprint arXiv:2411.15242*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, et al. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*, 2025.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

Colin Lockard, Prashant Shiralkar, and Xin Luna Dong. Openceres: When open information extraction meets the semi-structured web. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3047–3056, 2019.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Zhen Qin, Songlin Yang, and Yiran Zhong. Hierarchically gated recurrent neural network for sequence modeling. *Advances in Neural Information Processing Systems*, 36:33202–33221, 2023.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *Advances in Neural Information Processing Systems*, 38:100092–100118, 2026.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Zeru Shi, Zhenting Wang, Fan Yang, Qifan Wang, and Ruixiang Tang. A single layer to explain them all: Understanding massive activations in large language models. *arXiv preprint arXiv:2605.08504*, 2026.

Zunhai Su and Kehong Yuan. Kvsink: Understanding and enhancing the preservation of attention sinks in kv cache quantization for llms. In *Second Conference on Language Modeling (COLM 2025)*, 2025.

Zunhai Su, Zhe Chen, Wang Shen, Hanyu Wei, Linge Li, Huangqi Yu, and Kehong Yuan. Rotatekv: Accurate and robust 2-bit kv cache quantization for llms via outlier-aware adaptive rotations. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence Article No.: 690, Pages 6200 - 6208*, 2025a.

Zunhai Su, Qingyuan Li, Hao Zhang, Weihao Ye, Qibo Xue, YuLei Qian, Yuchen Xie, Ngai Wong, and Kehong Yuan. Unveiling super experts in mixture-of-experts large language models. In *International Conference on Learning Representations 2026, 118323-118354*, 2025b.

Zunhai Su, Rui Yang, Chao Zhang, Yaxiu Liu, Yifan Zhang, Wei Wu, Jing Xiong, Dayou Du, Xialie Zhuang, Yulei Qian, et al. Oscar: The occam's razor for extreme kv cache quantization in llms and beyond. *Under review*, 2026a.

Zunhai Su, Hengyuan Zhang, Wei Wu, Yifan Zhang, Yaxiu Liu, He Xiao, Qingyao Yang, Yuxuan Sun, Rui Yang, Chao Zhang, et al. Attention sink in transformers: A survey on utilization, interpretation, and mitigation. *Under review*, 2026b.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks. *arXiv preprint arXiv:2603.05498*, 2026.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

Kimi Team, Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, et al. Kimi linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025a.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026a.

Meituan LongCat Team, Anchun Gui, Bei Li, Bingyang Tao, Bole Zhou, Borun Chen, Chao Zhang, Chengcheng Han, Chenhui Yang, Chi Zhang, et al. Introducing longcat-flash-thinking: A technical report. *LongCat LLM Technical Report*, 2025b.

Meituan LongCat Team, Bei Li, Bingye Lei, Bo Wang, Bolin Rong, Chao Wang, Chao Zhang, Chen Gao, Chen Zhang, Cheng Sun, et al. Longcat-flash technical report. *LongCat LLM Technical Report*, 2025c.

Meituan LongCat Team, Anchun Gui, Bei Li, Bingyang Tao, Bole Zhou, Borun Chen, Chao Zhang, Chen Gao, Chen Zhang, Chengcheng Han, et al. Longcat-flash-thinking-2601 technical report. *LongCat LLM Technical Report*, 2026b.

Meituan LongCat Team, Bin Xiao, Chao Wang, Chengjiang Li, Chi Zhang, Chong Peng, Hang Yu, Hao Yang, Haonan Yan, Haoze Sun, et al. Longcat-next: Lexicalizing modalities as discrete tokens. *LongCat LLM Technical Report*, 2026c.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Dustin Wang, Rui-Jie Zhu, Steven Abreu, Yong Shan, Taylor Kergan, Yuqi Pan, Yuhong Chou, Zheng Li, Jibin Wu, Ge Zhang, et al. A systematic analysis of hybrid linear attention. *arXiv preprint arXiv:2507.06457*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

He Xiao, Qingyao Yang, Dirui Xie, Wendong Xu, Zunhai Su, Runming Yang, Haobo Liu, Wenyong Zhou, Zhengwu Liu, and Ngai Wong. Exploring layer-wise information effectiveness for post-training quantization in small language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15753–15764, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Songlin Yang and Yu Zhang. FLA: A triton-based library for hardware-efficient implementations of linear attention mechanisms. `https://github.com/fla-org/flash-linear-attention`, 2024. GitHub repository.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in neural information processing systems*, 37:115491–115522, 2024.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. In *International Conference on Learning Representations*, volume 2025, pp. 29687–29707, 2025b.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Hengyuan Zhang, Xinrong Chen, Zunhai Su, Xiao Liang, Jing Xiong, Wendong Xu, He Xiao, Chaofan Tao, Wei Zhang, Ruobing Xie, et al. Beyond outliers: A data-free layer-wise mixed-precision quantization approach driven by numerical and structural dual-sensitivity. *Under review*, 2026a.

Hengyuan Zhang, Zhihao Zhang, Mingyang Wang, Zunhai Su, Yiwei Wang, Qianli Wang, Shuzhou Yuan, Ercong Nie, Xufeng Duan, Qibo Xue, et al. Locate, steer, and improve: A practical survey of actionable mechanistic interpretability in large language models. *Computer Science Review*, 62:101011, 2026b.

APPENDIX CONTENTS

A Related Work **15**

A.1 Linear Attention and Hybrid Linear Attention Models 15

A.2 Massive Activations in Large Language Models 15

B MA Dynamics in the M-A-P Model Suite **15**

B.1 Model Suite and Pretraining Details 15

B.2 Attention-Sink-Guided Tracking and Alignment 16

B.3 MA Dynamics across Token Positions 18

B.4 MA Dynamics across Input Domains 18

B.5 MA Dynamics across Linear Attention Backbones and Hybridization Ratios 18

B.6 Statistical Controls and Uncertainty 18

C MA Dynamics in Large-Scale Pretrained Hybrid Models **20**

C.1 Model Families and Evaluated Checkpoints 20

C.2 MA Dynamics across Models and Input Domains 21

D Controlled Pretraining Studies of PAS and ISP **21**

D.1 Experimental Setup and Analysis Overview 21

D.2 Downstream Evaluation Protocol 22

D.3 Training-Time Emergence of PAS 22

D.4 Training-Time Emergence of ISP 24

D.5 Effects of Output Gating on PAS and ISP 25

E Additional Systematic-Outlier Analyses of PAS and ISP **26**

E.1 Fixed-Coordinate Systematic-Outlier Analysis of PAS 26

E.2 Cross-Model Systematic-Outlier Analyses 27

F Functional Intervention Protocols and Results **28**

F.1 Intervention Definition and Evaluation Scope 28

F.2 Complete Model-Specific Results 29

F.3 Exploratory Extension to Qwen3.5-27B 33

A Related Work

A.1 Linear Attention and Hybrid Linear Attention Models

Linear attention replaces pairwise token interactions with recurrent state updates, enabling linear-time sequence processing with fixed-size states. Representative architectures include RetNet, which introduces decay-based retention; HGRN and GLA, which improve state control through hierarchical and input-dependent gating; and DeltaNet, which employs delta-rule updates to selectively modify associative memory (Sun et al., 2023; Qin et al., 2023; Yang et al., 2023; 2024). GDN further augments the delta rule with data-dependent forgetting, while closely related state-space models such as Mamba achieve similar efficiency through selective state transitions (Yang et al., 2025b; Gu & Dao, 2023). Because fixed-size states can limit precise recall, recent LLMs increasingly interleave recurrent mixers with full attention layers. Representative examples include Jamba and Nemotron-H, which combine Mamba with full attention; MiniMax-01, which combines Lightning Attention with softmax attention; Qwen3-Next and Qwen3.5, which interleave GDN with gated attention; and Kimi Linear, which combines Kimi Delta Attention with Multi-Head Latent Attention (Lieber et al., 2024; Blakeman et al., 2025; Li et al., 2025; Cao et al., 2026; Team et al., 2025a). These hybrid architectures achieve favorable performance–efficiency trade-offs, but their internal activation dynamics remain poorly understood.

A.2 Massive Activations in Large Language Models

MAs were first identified as extremely sparse hidden-state outliers whose magnitudes exceed typical activations by several orders of magnitude. They persist across layers, remain largely invariant to input content, and function as implicit bias terms closely associated with attention sinks (Sun et al., 2024). Subsequent work connects MAs to a broader system of activation, weight, and attention outliers; traces their sources to super experts in MoE LLMs; and links their cross-layer evolution to attention sinks (An et al., 2025; Su et al., 2025b; Su & Yuan, 2025). Recent analyses further identify pre-normalization as central to the coupling between MAs and attention sinks, trace their emergence to RMSNorm and FFN computation, and explain their propagation through residual connections (Sun et al., 2026; Shi et al., 2026). Beyond their mechanistic role, MAs create a challenging numerical regime for low-bit inference (Zhang et al., 2026b). Although MAs are distinct from conventional channel-wise outlier features, their extreme magnitudes broaden the tensor dynamic range; under activation or KV cache quantization, a small number of such values can dominate shared scales, reduce the resolution available to typical values, and amplify quantization error. This challenge motivates outlier-aware transformations and sink-preserving KV cache quantization methods (Zhang et al., 2026a; Xiao et al., 2026; Su et al., 2025a; 2026a; Su & Yuan, 2025). Training-time studies additionally show that MAs emerge along predictable, architecture-dependent trajectories (Gallego-Feliciano et al., 2025). Despite this progress, existing work has focused almost exclusively on full attention LLMs, leaving MAs in HLA LLMs largely uncharacterized.

B MA Dynamics in the M-A-P Model Suite

B.1 Model Suite and Pretraining Details

Our inference-time analysis uses 52 pretrained checkpoints from the model suite released by Wang et al. (2025). The evaluated checkpoints span five linear attention architectures—RetNet, HGRN, GLA, DeltaNet, and GDN—at two parameter scales, 340M and 1.3B. For each architecture and scale, we consider four hybridization configurations and a pure linear attention baseline, together with a full attention Transformer reference at each scale. All evaluated checkpoints are publicly available through the M-A-P Hybrid Linear Attention Research collection.[2]

Linear Attention Families. The five architectures considered in our analysis span vector-valued recurrence, matrix-valued associative memory, input-dependent retention, and delta-rule state updates. They differ primarily in the structure of their recurrent states, the mechanisms governing

[2]M-A-P model collection on Hugging Face.

Table 4: Linear attention architectures evaluated in our inference-time analysis. The architectures differ in state structure, retention mechanism, and update rule.

Architecture	State Structure	Retention Mechanism	Update Rule
HGRN	Vector	Input-dependent forgetting	Gated recurrence
RetNet	Matrix	Fixed exponential decay	Outer-product accumulation
GLA	Matrix	Input-dependent decay	Gated outer-product accumulation
DeltaNet	Matrix	No explicit decay	Delta-rule correction
GDN	Matrix	Input-dependent forgetting	Gated delta-rule correction

Table 5: Hybridization configurations considered in our inference-time analysis. All models comprise 24 sequence-mixing layers.

Configuration	ρ	Linear Attention	Full Attention
Full attention	1	0	24
3:1	3	16	8
6:1	6	20	4
12:1	12	22	2
24:1	24	23	1
Pure linear attention	∞	24	0

information retention, and the rules used to incorporate new information. HGRN maintains a compact vector-valued state controlled by input-dependent forget gates. RetNet employs a matrix-valued state with fixed exponential decay. GLA replaces fixed decay with input-dependent multiplicative gates, allowing retention to vary across tokens and state dimensions. DeltaNet applies the delta rule to correct existing key–value associations rather than unconditionally accumulating new outer products. GDN combines delta-rule correction with input-dependent forgetting, enabling flexible control over state retention and overwriting. Table 4 summarizes these distinctions.

Hybridization Configurations. All evaluated models comprise 24 sequence-mixing layers. Within each model family, we hold the linear attention architecture fixed while varying the number of full attention layers, enabling controlled comparisons of how hybridization shapes MA dynamics.

We define the hybridization ratio as $\rho = L/L_{\mathrm{FA}}$, where $L = 24$ is the total number of sequence-mixing layers and L_{FA} is the number of full attention layers. A ρ:1 configuration therefore contains one full attention layer per ρ sequence-mixing layers. Larger values of ρ correspond to sparser full attention. The evaluated checkpoints cover four hybrid configurations, with $\rho \in \{3, 6, 12, 24\}$. We additionally include a full attention Transformer with $\rho = 1$ and, by convention, a pure linear attention model corresponding to $\rho = \infty$. Together, these configurations span the spectrum from exclusively full attention to exclusively linear attention.

Original Pretraining Recipe. All checkpoints were pretrained from scratch on FineWeb-Edu (Penedo et al., 2024) using implementations from the Flash Linear Attention library (Yang & Zhang, 2024). Within each parameter scale, all architectures and hybridization configurations share the same training corpus and optimization recipe, minimizing variation unrelated to the sequence-mixing architecture or full attention density. The 340M models were trained on 20B tokens with a global batch size of 50K tokens, whereas the 1.3B models were trained on 100B tokens with a global batch size of 1M tokens. Both scales were optimized using AdamW with a cosine learning-rate schedule. Table 6 summarizes the original pretraining configurations.

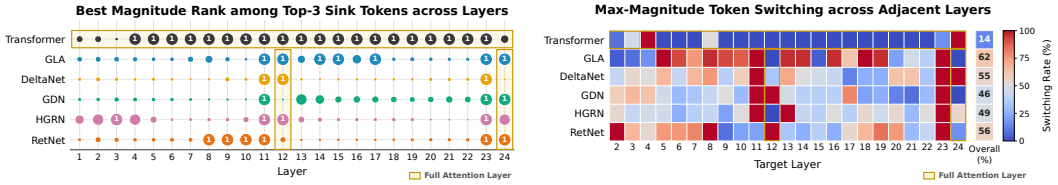
The shared architecture depth and pretraining recipe make this suite a controlled testbed for comparing MA dynamics across linear attention architectures, hybridization ratios, and model scales.

B.2 Attention-Sink-Guided Tracking and Alignment

Magnitude Ranking and Sink Alignment. Prior work detects MAs by selecting the hidden state entries with the largest absolute magnitudes at each layer (Sun et al., 2024). Subsequent studies show that these activations are not merely numerical anomalies but a key component of a systematic

Table 6: Original pretraining configurations of the checkpoints used in our inference-time analysis.

Configuration	340M	1.3B
Number of layers	24	24
Training corpus	FineWeb-Edu	FineWeb-Edu
Training tokens	20B	100B
Global batch size	50K tokens	1M tokens
Optimizer	AdamW	AdamW
Learning-rate schedule	Cosine	Cosine
Implementation	Flash Linear Attention	Flash Linear Attention



(a) Best magnitude rank among the top-3 consensus sink tokens across layers. Larger bubbles indicate better magnitude–sink alignment.

(b) Mean frequency of changes in the maximally activated token between adjacent layers.

Figure 6: **Magnitude–sink correspondence in 1.3B HLA LLMs at a** 12:1 **hybridization ratio.** Compared with the full attention Transformer, HLA models exhibit less consistent alignment between attention sinks and magnitude-ranked tokens (left), together with more frequent switching of maximally activated tokens across layers (right). Gold boxes mark full attention layers.

outlier mechanism in Transformers; tokens carrying MAs frequently attract a disproportionate share of attention mass and coincide with attention sinks (An et al., 2025; Sun et al., 2026; Su & Yuan, 2025; Su et al., 2026b). In full attention LLMs, this alignment between numerical magnitude and functional attention behavior makes layerwise magnitude ranking a useful proxy for tracing MAs across depth. Under hybridization, however, attention sink tokens exhibit lower and less stable magnitude rankings than in the full attention baseline, while the maximally activated token switches more frequently between adjacent layers (Figure 6). We therefore complement magnitude ranking with attention behavior, using consensus attention sinks as stable, attention-derived token anchors.

Consensus Attention Sinks as Token Anchors. Equation 5 in §3.1 defines the input-specific consensus sink using attention received from later queries, averaged across full attention layers and heads and normalized by the number of eligible queries. The diagnostic in Figure 6a considers the three highest-scoring tokens; the quantitative alignment and retention analyses use the single dominant sink $t_x^\star$. Its activation trajectory $m^{(\ell)}_{x,t_x^\star}$ preserves token identity, while the maximally activated feature may change with depth. Attention selects the anchor independently of activation magnitude; fixed-coordinate lifecycle analysis is conducted separately in Appendix E.1.

Sink Token Categories. Across the examined models and inputs, sink positions include the first token, punctuation and delimiters (newlines, periods, commas), and semantically light words such as *the*, *is*, *of*, and *in*, consistent with earlier MA observations (Sun et al., 2024). The first token is the most prevalent; the main qualitative analysis uses it, while quantitative alignment uses the input-specific consensus sink. Sink selection and the MA definition allow any eligible token position. Under causal masking, a fixed first token's hidden states are independent of subsequent text, although its received attention can depend on later queries. First-token plots illustrate layerwise organization; other positions appear in Appendix B.3.

Table 7 reports the domain-specific results underlying Table 1 for both parameter scales. Each domain contains 100 inputs; Overall denotes their macro-average. Token/layer controls and uncertainty analyses are provided in Appendix B.6.

Table 7: **Sink–spike alignment rates in M-A-P HLA models (%).** Results are reported for different linear attention backbones under the 12:1 hybridization ratio. Each entry denotes 1.3B/340M, and Overall is the macro-average across the five evaluation domains.

Linear Attention	WikiText	Scientific Papers	GSM8K	CodeSearchNet	FLORES	Overall
RetNet	99.5 / 100.0	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	**99.9 / 100.0**
HGRN	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	**100.0 / 100.0**
GLA	100.0 / 99.5	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	100.0 / 98.5	**100.0 / 99.6**
DeltaNet	100.0 / 100.0	100.0 / 98.0	100.0 / 100.0	100.0 / 99.5	100.0 / 99.5	**100.0 / 99.4**
GDN	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	100.0 / 100.0	**100.0 / 100.0**

B.3 MA Dynamics across Token Positions

To complement the first-token analysis in the main text, we examine all eight token positions in the running example, *"Summer is warm. Winter is cold."* For each token, we trace its maximum absolute hidden-state activation across depth in GDN models under pure linear attention, the 24:1, 12:1, 6:1, and 3:1 hybrid configurations, and full attention. The resulting trajectories are presented in Figure 16. Red dashed lines mark layers immediately preceding full attention, while green dotted lines mark full attention layers. Pronounced MAs concentrate at attention-sink positions, particularly the initial token, "Summer," and the first period. These tokens exhibit the same ratio-dependent organization identified in the main analysis: isolated PAS arise under sparse full attention, become progressively connected through ISP as full attention becomes denser, and ultimately approach the stable MA morphology of the full attention model. By contrast, tokens that do not serve as attention sinks exhibit neither pronounced PAS nor ISP. For this running example, these results show that PAS and ISP emerge selectively at sink-associated positions, supporting the use of the first token as a representative anchor in the main analysis.

B.4 MA Dynamics across Input Domains

We visualize PAS in representative inputs from different domains. For the running example and representative inputs from WikiText-103, Scientific Papers, GSM8K, CodeSearchNet, and FLORES-200 (Merity et al., 2016; Cohan et al., 2018; Cobbe et al., 2021; Husain et al., 2019; Costa-Jussà et al., 2022), we trace the maximum absolute hidden-state activation of the first token across five HLA architectures at the 1.3B scale and a fixed 12:1 hybridization ratio. A full attention Transformer is included as a reference. The resulting trajectories are presented in Figure 17.

These representative first-token trajectories show qualitatively similar layerwise organization, using the anchor described in Appendix B.2. Across all five linear attention backbones, sink-associated activations attain pronounced maxima immediately before full attention layers. They complement the input-specific consensus-sink statistics in Table 1, computed over 100 inputs per domain, which quantify recurrence across domains.

B.5 MA Dynamics across Linear Attention Backbones and Hybridization Ratios

We complement the cross-backbone analysis with within-family comparisons that characterize how MA dynamics vary with the hybridization ratio. Figure 18 presents results for GDN, DeltaNet, GLA, HGRN, and RetNet at both the 1.3B and 340M parameter scales. For each backbone and scale, we trace the maximum absolute hidden-state activation of the first token, "Summer," across model depth. The figure compares configurations containing 0, 1, 2, 4, 8, or 24 full attention layers among the 24 sequence-mixing layers, corresponding to pure linear attention, the 24:1, 12:1, 6:1, and 3:1 hybrid configurations, and a full attention Transformer, respectively. Red dashed lines mark layers immediately preceding full attention, while green dotted lines mark full attention layers.

B.6 Statistical Controls and Uncertainty

We complement the main M-A-P analysis with token and layer controls for sink–spike alignment and bootstrap uncertainty for both alignment and ISR. All analyses use the same 500 inputs as the

Table 8: **Token controls for PAS localization and prominence in M-A-P HLA models.** Each entry reports 1.3B/340M under the 12:1 hybridization ratio, macro-averaged over five domains with 100 inputs each. Alignment is reported in percent; peak excess is measured in $\log_2$ units using Equation 9. Non-sink denotes the average over tokens outside the top three consensus sinks.

	Alignment (%)			Peak Excess ($\log_2$)		
Linear Attention	**Sink**	**Non-sink**	**First Token**	**Sink**	**Non-sink**	**First Token**
RetNet	99.9 / 100.0	57.1 / 71.5	99.9 / 100.0	2.70 / 2.70	0.07 / 0.18	2.69 / 2.62
HGRN	100.0 / 100.0	93.3 / 84.3	100.0 / 100.0	4.32 / 4.83	0.52 / 0.41	4.32 / 4.83
GLA	100.0 / 99.6	75.9 / 69.4	100.0 / 99.7	3.70 / 3.18	0.21 / 0.17	2.34 / 2.81
DeltaNet	100.0 / 99.4	57.6 / 40.8	99.8 / 99.9	2.86 / 0.89	0.04 / −0.08	2.85 / 1.01
GDN	100.0 / 100.0	81.2 / 93.5	100.0 / 100.0	3.04 / 2.01	0.28 / 0.54	3.03 / 2.01

main experiments, with 100 inputs from each of the five evaluation domains. Confidence intervals are obtained from 10,000 stratified bootstrap resamples, sampling inputs with replacement within each domain before recomputing the macro-average.

Controls for Sink–Spike Alignment. For each input, we aggregate received attention across all full attention layers and select the consensus sink as in Equation 5. We compare it with two token controls: the first token and the average over eligible tokens outside the top three consensus sinks. We also include a random-layer baseline, which selects uniformly among the 11 layers in each preceding linear attention block and therefore has an expected alignment rate of $1/11 = 9.1\%$.

To measure prominence beyond the location of the blockwise maximum, we define the continuous *peak excess*

$$E_{x,t,f} = \log_2 \frac{m_{x,t}^{(f-1)}}{\max_{\ell \in \mathcal{B}_f \setminus \{f-1\}} m_{x,t}^{(\ell)}}. \tag{9}$$

A positive value indicates that the pre-attention activation exceeds every earlier activation in the block; one unit corresponds to a twofold increase over the strongest earlier layer.

Table 8 shows that consensus sinks are more strongly localized and substantially more prominent than the non-sink controls. Across all ten architecture–scale pairs, the sink-minus-non-sink alignment gap ranges from 6.5 to 58.6 percentage points, while the peak-excess gap ranges from 0.97 to 4.42 $\log_2$ units; every paired 95% bootstrap confidence interval excludes zero. The first-token control is often comparable to the consensus sink, reflecting that the initial position is itself the dominant sink for most inputs. Thus, the controls establish that PAS concentrates at sink-associated positions relative to generic tokens.

Uncertainty of the ISR Trend. We assess the monotonic ISR trend in Table 2 with a paired stratified bootstrap, matching each input across adjacent hybridization ratios. All 20 architecture–scale comparisons increase with denser full attention. The 12:1 $\rightarrow$ 6:1 increases range from 5.5 to 44.5 percentage points, and the 6:1 $\rightarrow$ 3:1 increases range from 17.0 to 52.3 points; all paired 95% confidence intervals exclude zero. These paired comparisons support higher mean inter-spike retention with denser full attention across the evaluated architectures and scales.

Absolute Inter-Spike Activation. Because ISR normalizes each inter-spike activation by its adjacent PAS, we additionally report the unnormalized mean activation magnitude

$$A_{\mathrm{ISP}}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \frac{1}{K-1} \sum_{i=1}^{K-1} \frac{1}{|\mathcal{J}_i|} \sum_{\ell \in \mathcal{J}_i} m_{x,t_x^\star}^{(\ell)}. \tag{10}$$

This companion statistic measures absolute plateau amplitude, rather than retention relative to the neighboring spikes. Table 9 shows that denser full attention increases mean absolute inter-spike activation in 19 of the 20 adjacent-ratio comparisons; the only exception is the 12:1 $\rightarrow$ 6:1 transition for 340M GDN. All ten 6:1 $\rightarrow$ 3:1 comparisons increase substantially. Thus, increasing ISR generally

Table 9: **Mean absolute inter-spike activation in M-A-P HLA models.** Each entry denotes $1.3\mathrm{B}/340\mathrm{M}$ and is macro-averaged across five domains with 100 inputs per domain. Values are computed using Equation 10; comparisons are meaningful within an architecture and scale because activation scales differ across checkpoints.

Linear Attention	**12:1**	**6:1**	**3:1**
RetNet	25.2 / 15.9	102.7 / 23.2	236.8 / 89.1
HGRN	14.7 / 8.5	132.8 / 70.2	368.9 / 165.5
GLA	37.8 / 13.1	71.4 / 20.9	360.5 / 89.7
DeltaNet	42.4 / 35.4	85.2 / 57.1	405.6 / 290.4
GDN	33.4 / 19.7	34.0 / 16.3	304.4 / 122.8

Table 10: **Large-scale pretrained hybrid models used for cross-model evaluation.** Mixer counts exclude MLP-only layers where applicable. Parameter counts denote total parameters; AxB indicates approximately xB activated parameters.

Model Family	**Linear-Time Mixer**	**Hybridization Pattern**	**Evaluated Checkpoints**
Kimi Linear	Kimi Delta Attention	20 KDA and 7 global MLA layers; $\rho = 27/7$	48B-A3B Base and Instruct
Qwen3.5	Gated DeltaNet	Three GDN layers followed by one gated full attention layer	35B-A3B Base and Instruct; 122B-A10B; 397B-A17B
Nemotron-H	Mamba-2	Mamba-2 and self-attention interleaved with separate MLP layers	8B, 47B, and 56B Base
Zamba2	Mamba-2	Mamba-2 backbone with periodically inserted shared Transformer blocks	1.2B, 2.7B, and 7B

accompanies stronger absolute activation, but the metrics differ: ISR measures persistence relative to adjacent PAS, whereas A_{ISP} measures absolute amplitude.

C MA Dynamics in Large-Scale Pretrained Hybrid Models

C.1 Model Families and Evaluated Checkpoints

To test whether the architecture-aligned MA organization extends beyond the controlled M-A-P suite, we evaluate 12 publicly available checkpoints from four large-scale hybrid model families. As summarized in Table 10, these models span 1.2B to 397B total parameters and cover linear attention and state-space mixers, periodic and nonuniform hybridization patterns, and both base and instruction-tuned checkpoints.

Kimi Linear. Kimi Linear replaces most full attention layers with Kimi Delta Attention (KDA), a linear attention mechanism based on the gated delta rule. Its 27-layer sequence-mixing stack contains 20 KDA layers and seven global Multi-Head Latent Attention (MLA) layers, with $\rho = 27/7$ by the all-mixer definition in §2. The approximately 3:1 KDA-to-MLA count is a different ratio; the schedule includes an additional MLA layer at the end. Matched 48B-A3B Base and Instruct checkpoints test PAS–ISP stability across post-training (Team et al., 2025a).

Qwen3.5. Qwen3.5 uses three GDN layers followed by one full softmax attention layer. This is a 3:1 linear-to-full pattern, but $\rho = 4$ under our all-mixer-to-full definition, not the M-A-P 3:1 configuration. The 35B-A3B, 122B-A10B, and 397B-A17B checkpoints contain 40, 48, and 60 sequence-mixing layers, including 10, 12, and 15 full attention layers, respectively. These full attention layers employ element-wise output gating, making Qwen3.5 particularly relevant to our controlled gating analysis. We evaluate 35B-A3B (Base and Instruct), 122B-A10B, and 397B-A17B checkpoints (Yang et al., 2025a).

Nemotron-H. Nemotron-H arranges Mamba-2, self-attention, and MLP modules as separate layers. The 8B and 56B configurations contain 24/4 and 54/10 Mamba-2/self-attention layers, respectively. The compressed 47B checkpoint retains a nonuniform schedule of 45 Mamba-2 and five self-attention layers. Nemotron-H therefore complements the fixed periodic designs above with sparser and partly irregular full attention placement (Blakeman et al., 2025).

Zamba2. Zamba2 interleaves a Mamba-2 backbone with hybrid blocks containing shared Transformer attention. The 1.2B, 2.7B, and 7B configurations contain 38, 54, and 81 layer positions, including six, nine, and thirteen hybrid blocks, respectively. These blocks occur approximately once every six layers. The 1.2B model repeatedly invokes one shared Transformer block, whereas the 2.7B and 7B models alternate between two shared blocks; layer-specific adapters specialize their repeated use across depth (Glorioso et al., 2024).

Together, these families cover linear attention and state-space hybrids, periodic and nonuniform full attention placement, shared and unshared attention modules, multiple parameter scales, and distinct post-training stages. This diversity allows us to test whether PAS and ISP reflect a general property of layer-interleaved hybrid architectures rather than a particular mixer or model family.

C.2 MA Dynamics across Models and Input Domains

Figure 19 presents the complete MA tracing results for the checkpoints summarized in Table 10. For each checkpoint, six trajectories cover the running example, *"Summer is warm. Winter is cold."*, and representative inputs from five domains: general prose from WikiText-103, scientific writing from Scientific Papers, mathematical reasoning from GSM8K, Python code from CodeSearchNet, and multilingual text from FLORES-200 (Merity et al., 2016; Cohan et al., 2018; Cobbe et al., 2021; Husain et al., 2019; Costa-Jussà et al., 2022). Each trajectory traces the maximum absolute hidden-state activation of the first token across model depth. Red dashed lines mark layers immediately preceding full attention, while green dotted lines mark full attention layers.

Post-Training and Model Scale. The matched base and instruction-tuned checkpoints of Kimi Linear and Qwen3.5 exhibit closely aligned PAS locations and ISP spans. Their activation magnitudes differ, but the correspondence between PAS–ISP organization and full attention placement remains largely intact in both matched pairs. Across Qwen3.5, Nemotron-H, and Zamba2 checkpoints of different sizes, PAS and ISP remain aligned with full attention placement even as their magnitude and prominence vary. Their magnitude does not vary monotonically with parameter count, whereas their correspondence with full attention placement persists.

Sequence Mixers and Input Domains. PAS and ISP recur in both linear attention hybrids, represented by Kimi Linear and Qwen3.5, and state-space hybrids, represented by Nemotron-H and Zamba2, despite their distinct sequence-mixing mechanisms. Across both classes, PAS occur immediately before full attention layers, while ISP extend across portions of the intervening linear-time sequence-mixing blocks. This cross-backbone consistency supports an association with layer-interleaved hybridization rather than a particular recurrent state-transition or update rule. Within each checkpoint, the six representative inputs show closely aligned PAS locations and ISP spans, with varying magnitudes. These plots provide representative first-token trajectories across checkpoints, with the interpretation of this anchor specified in Appendix B.2. The broader input-specific sink statistics and token controls are reported for M-A-P in Appendix B.6.

D Controlled Pretraining Studies of PAS and ISP

D.1 Experimental Setup and Analysis Overview

Model Architecture. We conduct controlled pretraining experiments using the Flash Linear Attention framework (Yang & Zhang, 2024). All models comprise 24 sequence-mixing layers and use Gated DeltaNet (GDN) as the linear attention backbone. We consider two parameter scales, 340M and 1.3B. To instantiate each hybrid configuration, we replace selected GDN mixers with full attention while holding all other architectural components fixed. Within each scale, this design isolates full attention placement and density from model depth, width, and FFN architecture.

Training Recipe. Following the optimization settings of prior GLA and GDN studies (Yang et al., 2023; 2025b), we train all models from scratch on FineWeb-Edu (Penedo et al., 2024). We use AdamW with a sequence length of 4,096, weight decay of 0.01, and gradient clipping at 1.0. The learning rate is linearly warmed up from 3×10^{-5} to 3×10^{-4} and then decayed to 3×10^{-5} using a cosine schedule. Unless otherwise specified, the 340M models are trained on 10B tokens with a global batch size of 0.5M tokens and a 0.5B-token warmup, whereas the 1.3B models are trained on 50B tokens with a global batch size of 2M tokens and a 1B-token warmup.

Analysis and Evaluation Overview. Our controlled study proceeds in three stages. First, we characterize the emergence of PAS across full attention placements and model scales (§D.3). Second, we trace ISP formation and consolidation in GDN models at both scales with a 3:1 hybridization ratio (§D.4). Third, we isolate the role of output gating through two complementary interventions: adding element-wise output gates to full attention layers following gated attention (Qiu et al., 2026), and removing the native output gates from GDN layers (§D.5). Alongside these morphology analyses, we evaluate the trained models on language modeling, commonsense reasoning, and in-context retrieval benchmarks following Yang et al. (2025b). These evaluations contextualize changes in PAS and ISP with downstream model capability. The complete evaluation protocol and benchmark descriptions are provided in §D.2.

D.2 Downstream Evaluation Protocol

We evaluate all models from our controlled pretraining experiments on language modeling, commonsense reasoning, and in-context retrieval benchmarks. Unless otherwise specified, we follow the evaluation protocol adopted in the Gated DeltaNet study (Yang et al., 2025b).

Language Modeling. We evaluate next-token prediction on WikiText-103 and report perplexity, with lower values indicating better language modeling performance (Merity et al., 2016).

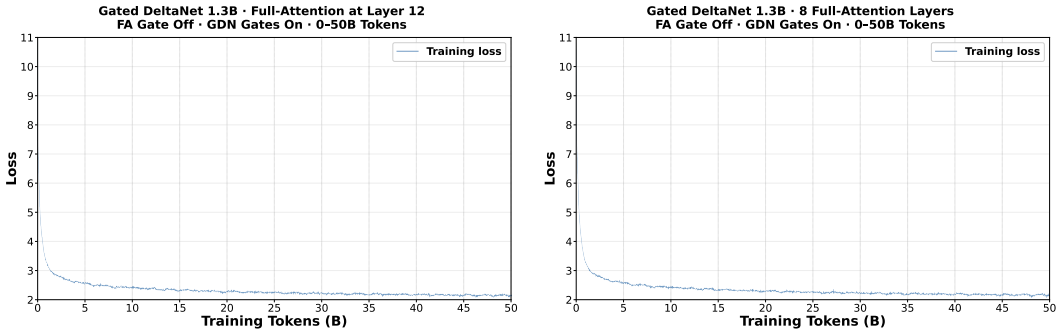
Commonsense Reasoning. We assess zero-shot performance on HellaSwag, PIQA, and ARC-Easy, which test plausible event completion, physical commonsense, and grade-school scientific reasoning, respectively (Zellers et al., 2019; Bisk et al., 2020; Clark et al., 2018). All evaluations are conducted using the Language Model Evaluation Harness (Gao et al., 2021). We report length-normalized accuracy for HellaSwag and accuracy for PIQA and ARC-Easy.

Real-World In-Context Retrieval. We evaluate recall-intensive performance on SWDE, FDA, SQuAD, and Natural Questions (NQ). SWDE assesses structured relation extraction from semi-structured web pages, while FDA assesses key–value retrieval from PDF documents (Lockard et al., 2019; Arora et al., 2024). SQuAD and NQ evaluate question answering over natural-language contexts (Rajpurkar et al., 2018; Kwiatkowski et al., 2019). Because our models are pretrained rather than instruction-tuned, we adopt the cloze-completion format of Arora et al. (2024), aligning these tasks with the next-token prediction objective. Inputs are truncated to a maximum context length of 2,048 tokens, and retrieval accuracy is reported for each benchmark.

Synthetic In-Context Retrieval. We additionally evaluate the three single-needle NIAH variants from RULER, denoted NIAH-1, NIAH-2, and NIAH-3 (Hsieh et al., 2024). NIAH-1 tests passkey-style retrieval from repeated distractors, NIAH-2 tests numerical-value retrieval from an essay context, and NIAH-3 increases the difficulty by replacing the target value with a UUID. Each task requires the model to recover the value associated with a target key embedded in a distractor context. All variants are evaluated at a fixed context length of 4,096 tokens using exact-match accuracy.

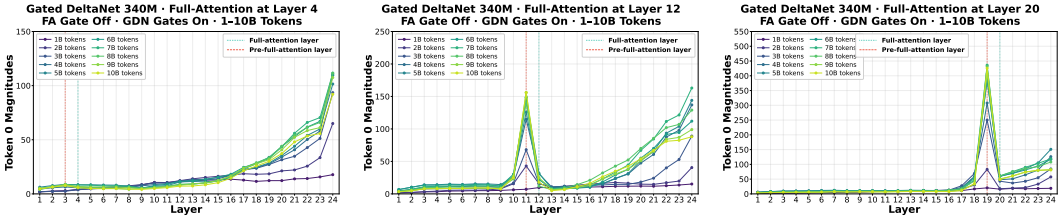
D.3 Training-Time Emergence of PAS

Experimental Setup. We study PAS formation by inserting a single full attention layer into an otherwise homogeneous 24-layer GDN model. At the 340M scale, we consider three insertion points—layers 4, 12, and 20—representing early, middle, and late full attention placements. All three models are trained from scratch on 10B tokens using the controlled pretraining recipe described in §D.1. To trace the emergence and evolution of PAS, we record the maximum absolute hidden-state activation of the first token across model depth at 1B-token intervals. At the final checkpoint, we



(a) PAS configuration: full attention only at layer 12. (b) ISP configuration with a 3:1 hybridization ratio.

Figure 7: **Training loss for representative PAS and ISP configurations.** Both curves are obtained from 1.3B GDN models trained on 50B tokens. The PAS configuration contains a single full attention layer at layer 12, whereas the ISP configuration interleaves eight full attention layers with 16 GDN layers under a 3:1 hybridization ratio.

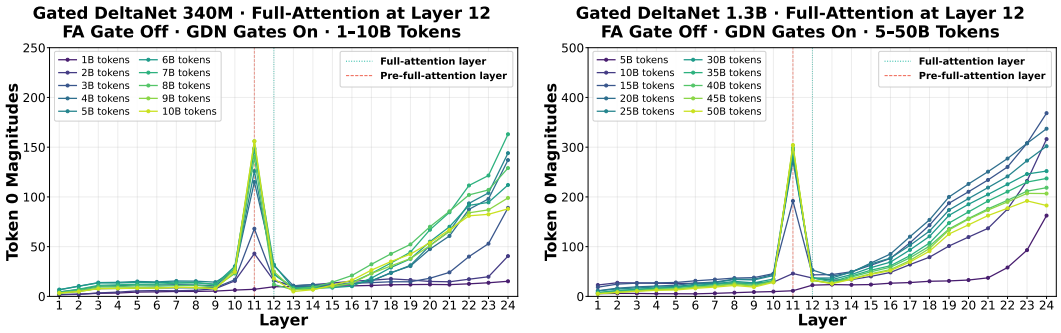


(a) Early placement: layer 4. (b) Middle placement: layer 12. (c) Late placement: layer 20.

Figure 8: **PAS formation across full attention placements.** Each panel traces the maximum absolute hidden-state activation of the first token across model depth at successive 1B-token checkpoints for a 340M GDN model containing a single full attention layer. PAS remains weak under early placement, becomes pronounced at the middle placement, and reaches its largest magnitude under late placement. Full attention placement therefore determines not only where PAS forms but also how prominently it develops.

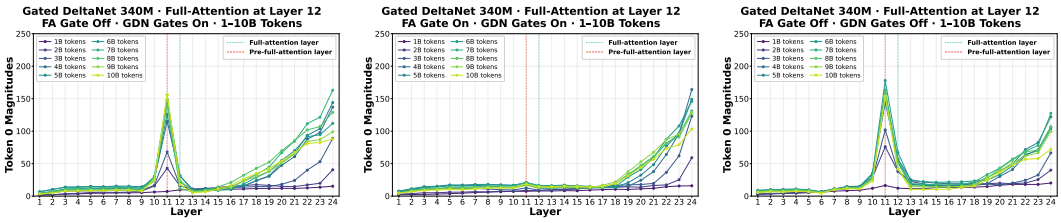
compute the sink–spike alignment rate and evaluate downstream performance. For layer 12, we also train a 1.3B counterpart on 50B tokens and record trajectories every 5B tokens, testing recurrence at a larger scale and training budget.

Results. Figure 7a confirms stable optimization of the 1.3B layer-12 configuration through 50B tokens. Figure 8 compares PAS formation across early, middle, and late full attention placements, while Figure 9 provides the corresponding cross-scale comparison. Table 11 reports the final sink–spike alignment rates and downstream performance. The output-gating variants are analyzed in §D.5. First, PAS consistently localizes immediately before the inserted full attention layer, but its magnitude depends strongly on placement depth. As shown in Figure 8, placement at layer 4 produces only a weak local maximum, whereas layer 12 produces a pronounced spike and layer 20 yields the strongest PAS. Second, PAS emerges early and recurs at the larger scale. In the 340M layer-12 configuration, it is already visible after 1B training tokens and becomes progressively sharper and more stable thereafter. The 1.3B trajectory shows a weak precursor at 5B tokens, a pronounced PAS by 10B, and persistent localization immediately before full attention through 50B tokens. Both scales achieve 100% final sink–spike alignment despite differences in training budget and activation profile. Third, full attention placement substantially affects retrieval performance despite having little effect on the final sink–spike alignment rate. All three configurations achieve near-perfect alignment and comparable language-modeling and commonsense performance. Nevertheless, middle and late placements outperform early placement on most retrieval tasks, showing that alignment alone does not predict retrieval performance.



(a) 340M model trained on 10B tokens. (b) 1.3B model trained on 50B tokens.

Figure 9: **PAS formation across model scales.** Both models contain 24 sequence-mixing layers with a single full attention layer at layer 12. This comparison examines whether the layer-localized PAS organization observed at the 340M scale persists in the 1.3B model.



(a) Standard GDN. (b) GDN-GatedFA. (c) GDN-NoOutGate.

Figure 10: **PAS dynamics under output-gating interventions.** Each panel traces the maximum absolute hidden-state activation of the first token across model depth at successive 1B-token checkpoints for a 340M GDN model with a single full attention layer at layer 12. The standard model develops a pronounced PAS. Full attention output gating attenuates but does not eliminate PAS; removing GDN output gates moderately amplifies it.

D.4 Training-Time Emergence of ISP

Experimental Setup. We study ISP formation using 24-layer GDN models with a fixed 3:1 hybridization ratio, in which eight full attention layers are interleaved with 16 GDN layers. At the 340M scale, the model is trained from scratch on 10B tokens following the controlled pretraining recipe described in §D.1. To trace the emergence and evolution of ISP, we record the maximum absolute hidden-state activation of the first token across model depth at 1B-token intervals. At the final checkpoint, we compute the inter-spike retention score and evaluate downstream performance. We also train a 1.3B counterpart on 50B tokens, recording trajectories every 5B tokens to test recurrence at a larger scale and training budget.

Results. Figure 7b confirms stable optimization of the 1.3B ISP configuration through 50B tokens, while Figure 12a traces the 340M first-token activation trajectory throughout pretraining. Figure 11 provides the corresponding cross-scale comparison, and Table 12 reports the final inter-spike retention scores and downstream performance. The output-gating variants are analyzed in §D.5. First, ISP emerges early during pretraining and becomes progressively more pronounced and stable across subsequent checkpoints. An elevated activation region between adjacent PAS is already visible at the first recorded checkpoint after 1B training tokens and gradually consolidates into a sustained plateau. Second, the same organization recurs at 1.3B. In the larger model, a broad plateau becomes evident over the early checkpoints and remains aligned with the interleaved full attention schedule through 50B tokens, although its absolute magnitude evolves non-monotonically during training. Final ISR is 90.56% at 340M and 94.47% at 1.3B, confirming strong relative retention at both scales; the different training budgets preclude attributing this difference to scale alone. These trajectories establish ISP as a training-time morphology recurring at both scales.

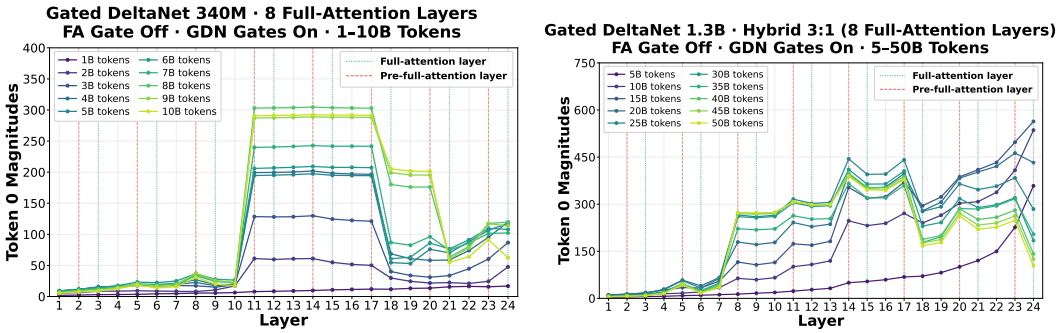
Table 11: **PAS and downstream performance under controlled pretraining.** FA Layer gives the one-based full attention position among 24 layers. Align. (sink–spike alignment) and downstream accuracies are percentages; PPL denotes perplexity.

(a) PAS alignment, language modeling, and real-world retrieval

Variant	Scale	FA Layer	Tokens	Align.	PPL	FDA	SWDE	NQ	SQuAD
GDN	340M	4/24	10B	99.53	14.00	8.02	20.21	14.95	25.30
GDN	340M	12/24	10B	100.00	13.78	60.43	40.49	19.61	32.81
GDN	340M	20/24	10B	100.00	13.84	53.71	35.80	21.10	35.42
GDN	1.3B	12/24	50B	100.00	9.79	66.58	45.36	23.57	36.96
GDN-NoOutGate	340M	12/24	10B	100.00	13.96	49.86	41.77	19.61	34.32
GDN-GatedFA	340M	12/24	10B	100.00	13.80	53.38	34.21	18.37	31.70

(b) Synthetic retrieval and commonsense reasoning

Variant	Scale	FA Layer	Tokens	NIAH-1	NIAH-2	NIAH-3	HellaSwag	PIQA	ARC-E
GDN	340M	4/24	10B	57.40	27.20	2.40	38.92	65.61	50.72
GDN	340M	12/24	10B	61.80	97.00	25.80	39.61	65.56	50.17
GDN	340M	20/24	10B	90.60	87.20	69.00	39.04	64.42	52.27
GDN	1.3B	12/24	50B	99.40	93.60	2.60	53.37	70.40	60.06
GDN-NoOutGate	340M	12/24	10B	80.40	94.40	39.40	38.87	65.78	49.33
GDN-GatedFA	340M	12/24	10B	41.60	16.80	30.80	39.73	65.13	50.63



(a) 340M model trained on 10B tokens.

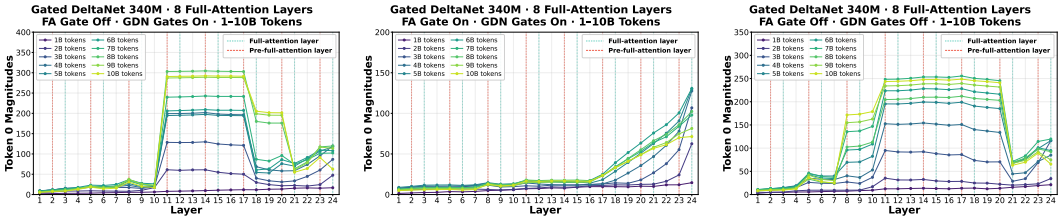
(b) 1.3B model trained on 50B tokens.

Figure 11: **ISP formation across model scales.** Both models contain 24 sequence-mixing layers and use a 3:1 hybridization ratio. Inter-spike persistence recurs at 1.3B despite differences in magnitude and training budget.

D.5 Effects of Output Gating on PAS and ISP

Experimental Setup. We examine how output gating regulates PAS and ISP using the representative 340M configurations from Sections D.3 and D.4. The PAS configuration is a 24-layer GDN model with a single full attention layer at layer 12, whereas the ISP configuration uses a 3:1 hybridization ratio, interleaving eight full attention layers with 16 GDN layers. All models are trained from scratch on 10B tokens under the same controlled recipe. For each configuration, we construct two variants while holding all other architectural components fixed. GDN-GatedFA adds element-wise output gates to the full attention layers following gated attention (Qiu et al., 2026), while retaining the native GDN gates. Conversely, GDN-NoOutGate removes all GDN output gates while leaving full attention ungated. These interventions isolate the module-specific effects of output gating. We record the maximum absolute hidden-state activation of the first token across model depth at successive 1B-token checkpoints and, at the final checkpoint, compute sink–spike alignment for PAS and inter-spike retention for ISP.

Results. Figures 10 and 12 compare the evolution of PAS and ISP under the standard configuration and the two gating interventions. Tables 11 and 12 report the corresponding morphology



(a) Standard GDN. (b) GDN-GatedFA. (c) GDN-NoOutGate.

Figure 12: **ISP dynamics under output-gating interventions.** Each panel traces the maximum absolute hidden-state activation of the first token across model depth at successive 1B-token checkpoints for a 340M GDN model with a 3:1 hybridization ratio trained on 10B tokens. The standard model develops a sustained ISP. Adding output gates to the full attention layers markedly reduces its absolute magnitude, although a weak inter-spike plateau remains visible and persists throughout training. Removing GDN output gates moderately amplifies the plateau.

Table 12: **ISP and downstream performance under controlled pretraining.** All configurations use 24 layers at 3:1. ISR and downstream accuracies are percentages; PPL denotes perplexity. ISR measures persistence relative to each variant's own PAS endpoints.

(a) ISP retention, language modeling, and real-world retrieval

Variant	Scale	Tokens	ISR	PPL	FDA	SWDE	NQ	SQuAD
GDN	340M	10B	90.56	13.74	54.62	39.68	21.63	38.40
GDN	1.3B	50B	94.47	9.80	64.82	45.92	24.90	40.65
GDN-NoOutGate	340M	10B	96.48	13.78	49.32	39.33	21.82	35.25
GDN-GatedFA	340M	10B	99.48	13.68	65.12	40.49	21.32	36.60

(b) Synthetic retrieval and commonsense reasoning

Variant	Scale	Tokens	NIAH-1	NIAH-2	NIAH-3	HellaSwag	PIQA	ARC-E
GDN	340M	10B	78.60	99.40	63.60	39.56	65.67	50.93
GDN	1.3B	50B	96.00	100.00	90.00	53.36	71.55	61.11
GDN-NoOutGate	340M	10B	57.20	48.20	14.20	39.57	66.16	50.25
GDN-GatedFA	340M	10B	61.00	98.20	5.00	39.38	66.00	50.34

metrics and downstream performance. We assess absolute ISP magnitude from the trajectories and relative persistence from ISR, which normalizes by adjacent PAS. First, adding output gates to full attention substantially attenuates the absolute magnitudes of PAS and ISP without eliminating their layerwise organization. Weak spikes and plateaus persist and become increasingly visible during pretraining: the trajectories retain a low-magnitude plateau, while ISR confirms continued inter-spike persistence. Consistently, Qwen3.5 models with native full attention output gating still exhibit both morphologies (Figures 19c–19f). Second, removing the native GDN output gates increases PAS and ISP magnitudes, with a more visible effect on the plateau, but the changes remain smaller than those induced by full attention gating. Thus, native GDN output gates modulate amplification and propagation but are not necessary for either morphology in these configurations. Third, gating the substantially fewer full attention layers produces larger changes in the absolute activation trajectories than removing gates from all GDN layers. This contrast motivates examining gate placement rather than gate count alone, while distinguishing the addition of full attention gates from the removal of GDN gates.

E Additional Systematic-Outlier Analyses of PAS and ISP

E.1 Fixed-Coordinate Systematic-Outlier Analysis of PAS

We examine the PAS immediately preceding the full attention layer $f = 12$ in a representative 1.3B GDN model with a 12:1 hybridization ratio. Following §4.1, let $t^\star$ denote the consensus sink token

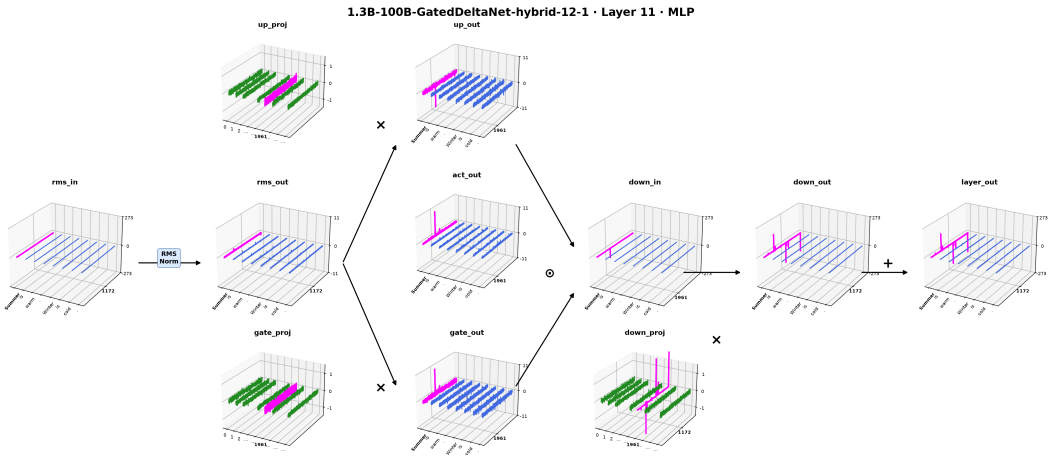


Figure 13: **Fixed-coordinate analysis of PAS formation.** At $(t^\star, j^\star)$, layer 11 produces an extreme update that writes a large outlier into the residual stream immediately before full attention.

and $j^\star = \arg\max_j |X^{(f-1)}_{t^\star,j}|$ its maximally activated feature at the corresponding pre-attention layer. We fix this token–feature coordinate and trace its signed activation and module-level updates through three stages: outlier formation, attention-sink coupling, and cancellation. Figures 13, 14, and 15 provide detailed evidence for these stages, respectively.

Outlier Formation. Figure 13 shows that the pre-attention layer produces an extreme update at the fixed coordinate, writing a large outlier into the residual stream immediately before full attention. This localization establishes that PAS arises from a concentrated update at a specific token–feature coordinate rather than a diffuse increase across tokens or features.

Attention-Sink Coupling. As shown in Figure 14, token $t^\star$ receives a disproportionate share of attention from subsequent query positions during full attention at layer 12, acting as the dominant attention sink. This links the fixed-coordinate outlier to full attention sink behavior.

Outlier Cancellation. Figure 15 tracks this coordinate through layer 12, where an opposite-signed update substantially cancels the incoming outlier and sharply reduces its residual magnitude. Following one coordinate links formation, attention-sink coupling, and cancellation within a coherent PAS lifecycle rather than unrelated layerwise extrema.

E.2 Cross-Model Systematic-Outlier Analyses

We extend §4 with layerwise systematic-outlier visualizations for M-A-P and large pretrained hybrids, tracing the first token of *"Summer is warm. Winter is cold."* Each figure pairs the MA envelope with signed residual input, mixer update, FFN update, and block output at each layer's dominant feature. The dominant feature is reselected at each layer, so these plots characterize layerwise write–cancel structure; Appendix E.1 separately follows one fixed coordinate for a representative PAS. In particular, an ISP envelope can persist while its dominant feature changes.

M-A-P Model Suite. We analyze 1.3B GDN, DeltaNet, GLA, HGRN, and RetNet models (Wang et al., 2025) at 24:1, 12:1, 6:1, and 3:1 (Figures 21–25). Figure 20 shows the common full attention reference once, since this limit is independent of the linear backbone. Across configurations, dominant outliers arise before full attention and are subsequently accompanied by opposite-signed updates. Denser full attention leaves larger portions of the inter-spike intervals elevated, consistent with delayed cancellation; at the full attention limit, spikes and plateaus merge into persistent MAs. These signed traces complement the quantitative morphology analysis in §3.3.

Large-Scale Open-Source Hybrid Models. Figures 26 and 27 examine three Qwen3.5 and three Nemotron-H checkpoints, respectively. Qwen3.5 interleaves GDN with output-gated full attention;

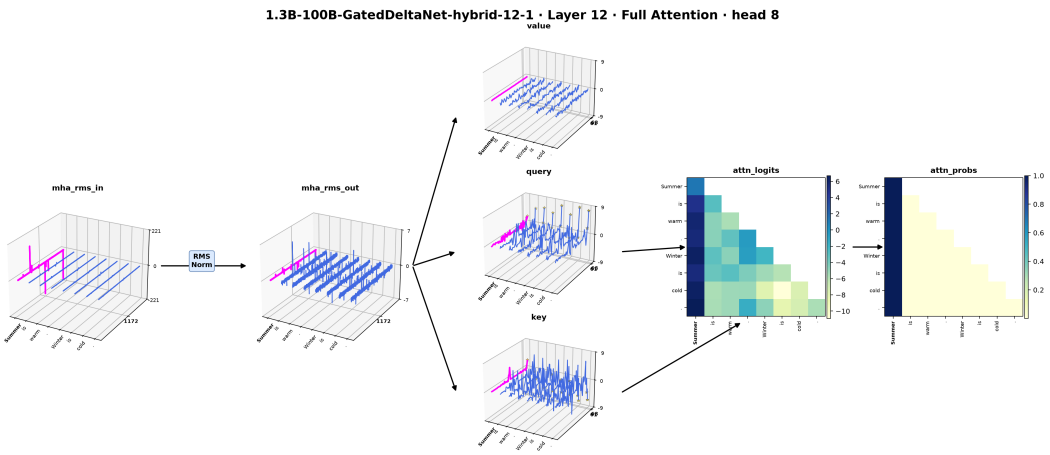


Figure 14: **Coupling between PAS and full attention sink behavior.** During full attention at layer 12, token $t^\star$ receives a disproportionate share of attention from subsequent query positions and acts as the dominant attention sink.

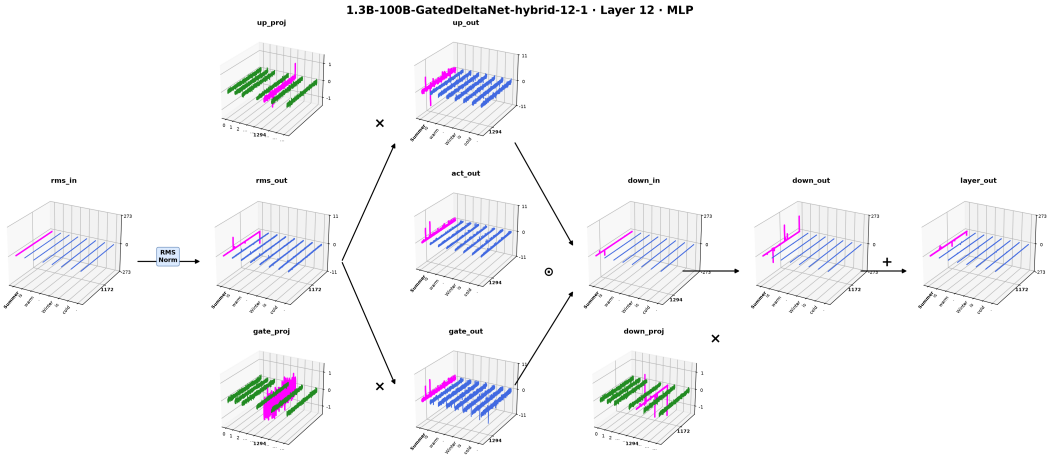


Figure 15: **Fixed-coordinate analysis of PAS cancellation.** At $(t^\star, j^\star)$, an opposite-signed layer-12 update substantially cancels the incoming outlier.

Nemotron-H combines Mamba-2 with full attention. Both families show amplification near full attention boundaries and extended high-activation regions between them, despite different mixers, scales, and schedules. Thus, the layerwise pattern also occurs with native full attention output gating and in state-space hybrids, broadening the cancellation-timing interpretation beyond M-A-P.

F Functional Intervention Protocols and Results

F.1 Intervention Definition and Evaluation Scope

Checkpoints and coordinates. We use official M-A-P GDN checkpoints, distinct from the controlled-pretraining models in §3.5. Both scales have 24 blocks; $\rho = 12, 6, 3$ corresponds to two, four, and eight full attention layers. Both interventions use the first prompt token as a fixed anchor, following §3.2. As in §4.3, $p_i = f_i - 1$ indexes the residual state entering full attention and $z^{(\ell)} = X^{(\ell)}_{1,:}$ its normal reference trajectory. We select $C_i = \mathrm{TopK}(|z^{(p_i)}|, k)$ and use the union $C = \bigcup_i C_i$ for connection.

Deletion and connection. Deletion sets $\widetilde{X}_{1,j}^{(p_i)} = 0$ for every $j \in C_i$ and every i. Coordinates selected in the normal pass remain fixed throughout deletion. Connection acts on C and sets, for $p_i < \ell < p_{i+1}$,

$$\widetilde{X}_{1,j}^{(\ell)} = s_{i,j} \max\Big(s_{i,j} z_j^{(\ell)}, |z_j^{(p_i)}|\Big), \quad s_{i,j} = \begin{cases} 1, & z_j^{(p_i)} \geq 0, \\ -1, & z_j^{(p_i)} < 0, \end{cases} \qquad j \in C. \tag{11}$$

At every endpoint, $\widetilde{X}_{1,j}^{(p_i)} = z_j^{(p_i)}$ for $j \in C$; unselected entries follow the intervened computation. Equation 11 uses the normal reference trajectory throughout, allowing signs to change at endpoints. The budget k applies *per input layer*, so the union C may contain more than k coordinates. Every example is evaluated, including identity cases already satisfying the target and weak-activation cases. Coordinate selection and reference values depend exclusively on activations.

Checks and scoring. The implementation checks coordinate-level target attainment and exact preservation of the selected endpoint values. Pinning endpoints without changing interior values must reproduce the normal output. Accelerated graph execution is checked against native forwards, and reused normal results must match the frozen baseline. Models run in BF16 with greedy retrieval decoding; frozen manifests record the execution seed and numerical configuration. For autoregressive retrieval, the intervention edits the original first-token state during prompt processing, retaining its token identity during generation. Model weights remain frozen. We use the benchmark formats and metrics in Appendix D.2. PIQA, ARC-Easy, and HellaSwag contain 1,838, 2,376, and 10,042 examples; HellaSwag uses length-normalized accuracy. FDA, SWDE, SQuAD, and NQ contain 1,101, 1,067, 2,977, and 3,157 cloze-completion examples, with up to 2,048 prompt and 48 generated tokens. SQuAD measures retrieval accuracy, not extractive F1. Each NIAH variant has 500 examples and a 4,096-token budget (up to 3,968 prompt plus 128 generated tokens), scored by exact match. WikiText-103 uses 2,048-token windows with stride 1,024: 326 windows cover 334,660 scored target tokens. Perplexity exponentiates their mean NLL.

Aggregation and interpretation. Accuracy is reported as a percentage. For each task, Table 3 averages the six models equally. Relative change is $100(\bar{S}_{\text{int}}/\bar{S}_{\text{normal}} - 1)$: the ratio of model means, not the mean of relative changes or a percentage-point difference. The ten-task mean weights accuracy tasks equally, excluding perplexity. Each model's perplexity exponentiates aggregate token NLL; averaging these perplexities is descriptive, not pooled likelihood. Appendix F.2 retains every score, including regressions, and reports exploratory paired uncertainty for retrieval. Deletion precedes normalization and can alter the full normalized representation. The results quantify dependence on the selected PAS coordinates. Connection jointly edits inter-spike activations and preserves reference endpoint values, so the measured gains concern this combined rule. The endpoint-only identity check verifies implementation consistency. An additional reference pass and trajectory storage are required.

F.2 Complete Model-Specific Results

Tables 14–19 retain all per-model results; relative changes use each model's own normal score.

The retrieval mean weights FDA, SWDE, SQuAD, and NQ equally, retaining negative task changes. All eighteen model–k settings improve this mean by 1.06%–12.60% relative to normal inference. These settings share checkpoints and evaluation examples.

Exploratory Paired Uncertainty. Using existing paired win/loss/tie counts, we bootstrap binary correctness differences within each task and recompute the four-task mean (100,000 resamples; seed 20260919). Pointwise 95% percentile intervals are positive for all eighteen means; Table 13 reports $k = 4$, with all budgets in the supplementary data. Two-sided exact McNemar tests, Holm-corrected over 72 model–budget–task contrasts, detect 24 positive changes and no negative changes at $\alpha = 0.05$; non-rejection does not establish equivalence. This post-hoc analysis assumes independent items within tasks, without document/site clustering or training-seed uncertainty.

Appendix F.1 specifies scoring and context limits. The following tables retain the complete task-level results, including regressions and non-monotonic deletion responses.

Table 13: **Connection gains and paired uncertainty.** $k = 4$. Scores average FDA, SWDE, SQuAD, and NQ equally. Relative change is measured against normal inference; Δ and its exploratory item-paired 95% bootstrap interval are in percentage points (pp). Intervals are pointwise and do not account for document/site clustering or training-seed variation.

Scale	Ratio	Normal (%)	Connect (%)	Relative (%)	Δ [95% CI] (pp)
340M	12:1	23.16	23.88	+3.11	+0.72 [0.41, 1.04]
	6:1	23.36	24.37	+4.29	+1.00 [0.61, 1.39]
	3:1	30.95	34.85	+12.60	+3.90 [3.23, 4.57]
1.3B	12:1	42.82	43.31	+1.14	+0.49 [0.18, 0.79]
	6:1	45.04	47.15	+4.68	+2.11 [1.54, 2.68]
	3:1	46.36	48.95	+5.57	+2.58 [2.04, 3.14]

Table 14: **340M GDN,** 12:1**.** Accuracy (%) and PPL; relative changes in parentheses.

Benchmark	Normal	PAS suppression			Spike connection		
		$k = 1$	$k = 2$	$k = 4$	$k = 1$	$k = 2$	$k = 4$
PIQA	67.41	66.27 (−1.69%)	66.92 (−0.73%)	67.36 (−0.08%)	67.46 (+0.08%)	67.74 (+0.48%)	67.57 (+0.24%)
ARC-Easy	58.54	57.07 (−2.52%)	56.86 (−2.88%)	56.61 (−3.31%)	58.63 (+0.14%)	58.84 (+0.50%)	58.63 (+0.14%)
HellaSwag	42.17	40.22 (−4.63%)	40.21 (−4.65%)	40.02 (−5.10%)	42.08 (−0.21%)	42.17 (+0.00%)	42.04 (−0.31%)
FDA	22.98	5.63 (−75.49%)	6.36 (−72.33%)	7.08 (−69.17%)	23.34 (+1.58%)	22.98 (+0.00%)	22.71 (−1.19%)
SWDE	22.21	8.72 (−60.76%)	11.81 (−46.84%)	12.46 (−43.88%)	22.40 (+0.84%)	22.40 (+0.84%)	22.59 (+1.69%)
SQuAD	31.94	19.42 (−39.22%)	25.16 (−21.24%)	28.08 (−12.09%)	34.97 (+9.46%)	35.27 (+10.41%)	34.50 (+7.99%)
NQ	15.52	12.35 (−20.41%)	12.45 (−19.80%)	12.80 (−17.55%)	15.74 (+1.43%)	15.77 (+1.63%)	15.74 (+1.43%)
NIAH-1	36.80	15.60 (−57.61%)	13.60 (−63.04%)	13.40 (−63.59%)	36.80 (+0.00%)	36.60 (−0.54%)	37.00 (+0.54%)
NIAH-2	50.40	15.80 (−68.65%)	20.00 (−60.32%)	22.00 (−56.35%)	51.60 (+2.38%)	51.80 (+2.78%)	51.20 (+1.59%)
NIAH-3	12.00	2.60 (−78.33%)	2.00 (−83.33%)	1.60 (−86.67%)	12.40 (+3.33%)	12.40 (+3.33%)	11.20 (−6.67%)
WikiText-103 PPL	14.0183	15.8006 (+12.71%)	15.6746 (+11.82%)	15.8038 (+12.74%)	14.0152 (−0.02%)	14.0127 (−0.04%)	14.0109 (−0.05%)
Ten-task mean	36.00	24.37 (−32.31%)	25.54 (−29.06%)	26.14 (−27.38%)	36.54 (+1.51%)	36.60 (+1.66%)	36.32 (+0.89%)

Table 15: **1.3B GDN,** 12:1**.** Accuracy (%) and PPL; relative changes in parentheses.

Benchmark	Normal	PAS suppression			Spike connection		
		$k=1$	$k=2$	$k=4$	$k=1$	$k=2$	$k=4$
PIQA	73.29	72.96 (−0.45%)	73.07 (−0.30%)	70.62 (−3.64%)	73.23 (−0.07%)	73.12 (−0.22%)	73.07 (−0.30%)
ARC-Easy	72.64	71.76 (−1.22%)	71.59 (−1.45%)	65.78 (−9.44%)	72.18 (−0.64%)	72.39 (−0.35%)	72.43 (−0.29%)
HellaSwag	60.99	60.77 (−0.36%)	60.74 (−0.41%)	55.90 (−8.36%)	60.83 (−0.26%)	60.75 (−0.39%)	60.85 (−0.23%)
FDA	59.22	60.22 (+1.69%)	59.67 (+0.77%)	4.36 (−92.64%)	59.58 (+0.61%)	59.49 (+0.46%)	59.58 (+0.61%)
SWDE	45.17	45.45 (+0.62%)	43.49 (−3.73%)	13.59 (−69.92%)	44.80 (−0.83%)	45.08 (−0.21%)	45.27 (+0.21%)
SQuAD	41.42	44.41 (+7.22%)	43.84 (+5.84%)	31.44 (−24.09%)	43.40 (+4.79%)	43.53 (+5.11%)	43.10 (+4.06%)
NQ	25.47	25.34 (−0.50%)	25.47 (+0.00%)	14.86 (−41.67%)	25.31 (−0.62%)	25.37 (−0.37%)	25.28 (−0.75%)
NIAH-1	82.60	81.40 (−1.45%)	76.60 (−7.26%)	2.40 (−97.09%)	82.60 (+0.00%)	82.60 (+0.00%)	82.60 (+0.00%)
NIAH-2	100.00	100.00 (+0.00%)	100.00 (+0.00%)	0.00 (−100.00%)	100.00 (+0.00%)	100.00 (+0.00%)	100.00 (+0.00%)
NIAH-3	64.20	63.80 (−0.62%)	56.80 (−11.53%)	0.00 (−100.00%)	65.40 (+1.87%)	65.00 (+1.25%)	64.80 (+0.93%)
WikiText-103 PPL	8.5199	8.5109 (−0.11%)	8.5187 (−0.01%)	11.8805 (+39.44%)	8.5097 (−0.12%)	8.5094 (−0.12%)	8.5091 (−0.13%)
Ten-task mean	62.50	62.61 (+0.18%)	61.13 (−2.20%)	25.89 (−58.57%)	62.73 (+0.37%)	62.73 (+0.38%)	62.70 (+0.32%)

Table 16: **340M GDN,** 6:1**.** Accuracy (%) and PPL; relative changes in parentheses.

Benchmark	Normal	PAS suppression			Spike connection		
		$k=1$	$k=2$	$k=4$	$k=1$	$k=2$	$k=4$
PIQA	66.59	64.36 (−3.35%)	62.46 (−6.21%)	62.40 (−6.29%)	66.70 (+0.16%)	66.54 (−0.08%)	66.54 (−0.08%)
ARC-Easy	59.43	55.81 (−6.09%)	50.21 (−15.51%)	47.98 (−19.26%)	59.13 (−0.50%)	58.71 (−1.20%)	58.84 (−0.99%)
HellaSwag	41.65	38.24 (−8.18%)	37.22 (−10.62%)	36.39 (−12.63%)	41.70 (+0.14%)	41.67 (+0.07%)	41.72 (+0.19%)
FDA	20.07	5.99 (−70.14%)	4.54 (−77.38%)	10.45 (−47.96%)	18.80 (−6.33%)	19.26 (−4.07%)	19.35 (−3.62%)
SWDE	23.34	6.19 (−73.49%)	5.06 (−78.31%)	11.06 (−52.61%)	24.18 (+3.61%)	24.37 (+4.42%)	23.71 (+1.61%)
SQuAD	32.72	25.56 (−21.87%)	27.31 (−16.53%)	28.69 (−12.32%)	39.17 (+19.71%)	38.80 (+18.58%)	37.55 (+14.78%)
NQ	17.33	13.18 (−23.95%)	12.39 (−28.52%)	12.73 (−26.51%)	16.98 (−2.01%)	17.04 (−1.65%)	16.85 (−2.74%)
NIAH-1	40.60	23.00 (−43.35%)	16.00 (−60.59%)	9.20 (−77.34%)	41.00 (+0.99%)	41.00 (+0.99%)	41.00 (+0.99%)
NIAH-2	51.00	21.40 (−58.04%)	14.80 (−70.98%)	11.00 (−78.43%)	53.00 (+3.92%)	52.80 (+3.53%)	52.80 (+3.53%)
NIAH-3	4.00	3.20 (−20.00%)	3.20 (−20.00%)	1.40 (−65.00%)	4.00 (+0.00%)	4.20 (+5.00%)	4.00 (+0.00%)
WikiText-103 PPL	14.6162	16.5302 (+13.10%)	16.9186 (+15.75%)	17.2137 (+17.77%)	14.6422 (+0.18%)	14.6407 (+0.17%)	14.6333 (+0.12%)
Ten-task mean	35.67	25.69 (−27.97%)	23.32 (−34.63%)	23.13 (−35.16%)	36.47 (+2.23%)	36.44 (+2.15%)	36.24 (+1.58%)

Table 17: **1.3B GDN,** 6:1**.** Accuracy (%) and PPL; relative changes in parentheses.

Benchmark	Normal	PAS suppression			Spike connection		
		$k=1$	$k=2$	$k=4$	$k=1$	$k=2$	$k=4$
PIQA	72.80	70.67 (−2.91%)	64.36 (−11.58%)	70.51 (−3.14%)	72.85 (+0.07%)	72.85 (+0.07%)	72.96 (+0.22%)
ARC-Easy	72.52	68.94 (−4.93%)	58.96 (−18.69%)	66.25 (−8.65%)	71.13 (−1.92%)	71.13 (−1.92%)	72.05 (−0.64%)
HellaSwag	60.44	54.22 (−10.28%)	44.89 (−25.72%)	57.48 (−4.89%)	60.17 (−0.44%)	60.17 (−0.44%)	60.36 (−0.13%)
FDA	61.13	61.49 (+0.59%)	40.96 (−32.99%)	52.59 (−13.97%)	62.40 (+2.08%)	61.94 (+1.34%)	63.22 (+3.42%)
SWDE	48.08	41.05 (−14.62%)	13.96 (−70.96%)	19.49 (−59.45%)	50.23 (+4.48%)	49.95 (+3.90%)	50.33 (+4.68%)
SQuAD	41.99	46.83 (+11.52%)	46.09 (+9.76%)	46.83 (+11.52%)	46.09 (+9.76%)	46.32 (+10.32%)	46.12 (+9.84%)
NQ	28.98	27.87 (−3.83%)	24.11 (−16.83%)	26.10 (−9.95%)	28.67 (−1.09%)	28.57 (−1.42%)	28.95 (−0.11%)
NIAH-1	95.60	87.80 (−8.16%)	54.20 (−43.31%)	56.00 (−41.42%)	94.60 (−1.05%)	94.60 (−1.05%)	94.80 (−0.84%)
NIAH-2	99.40	97.40 (−2.01%)	70.20 (−29.38%)	88.20 (−11.27%)	99.40 (+0.00%)	99.20 (−0.20%)	99.20 (−0.20%)
NIAH-3	54.80	49.20 (−10.22%)	29.40 (−46.35%)	13.00 (−76.28%)	58.20 (+6.20%)	58.00 (+5.84%)	59.80 (+9.12%)
WikiText-103 PPL	8.5247	8.9381 (+4.85%)	12.3677 (+45.08%)	9.3092 (+9.20%)	8.4998 (−0.29%)	8.5004 (−0.28%)	8.4969 (−0.33%)
Ten-task mean	63.57	60.55 (−4.76%)	44.71 (−29.66%)	49.64 (−21.91%)	64.37 (+1.26%)	64.27 (+1.10%)	64.78 (+1.90%)

Table 18: **340M GDN,** 3:1**.** Accuracy (%) and PPL; relative changes in parentheses.

Benchmark	Normal	PAS suppression			Spike connection		
		$k=1$	$k=2$	$k=4$	$k=1$	$k=2$	$k=4$
PIQA	67.25	65.45 (−2.67%)	65.07 (−3.24%)	63.87 (−5.02%)	67.03 (−0.32%)	67.19 (−0.08%)	66.92 (−0.49%)
ARC-Easy	57.91	55.51 (−4.14%)	54.08 (−6.61%)	52.44 (−9.45%)	57.83 (−0.15%)	58.00 (+0.15%)	58.16 (+0.44%)
HellaSwag	41.18	40.46 (−1.74%)	38.72 (−5.97%)	39.31 (−4.52%)	41.43 (+0.60%)	41.45 (+0.65%)	41.32 (+0.34%)
FDA	41.24	7.27 (−82.38%)	1.63 (−96.04%)	0.64 (−98.46%)	43.23 (+4.85%)	43.05 (+4.41%)	42.69 (+3.52%)
SWDE	27.65	9.75 (−64.75%)	1.31 (−95.25%)	0.84 (−96.95%)	33.36 (+20.68%)	33.46 (+21.02%)	33.93 (+22.71%)
SQuAD	35.77	32.82 (−8.26%)	18.98 (−46.95%)	8.87 (−75.21%)	41.08 (+14.84%)	40.88 (+14.27%)	44.21 (+23.57%)
NQ	19.13	14.89 (−22.19%)	7.13 (−62.75%)	3.61 (−81.13%)	18.15 (−5.13%)	18.06 (−5.63%)	18.56 (−2.98%)
NIAH-1	71.80	15.20 (−78.83%)	6.00 (−91.64%)	0.00 (−100.00%)	73.00 (+1.67%)	73.20 (+1.95%)	72.80 (+1.39%)
NIAH-2	82.60	10.40 (−87.41%)	5.20 (−93.70%)	0.00 (−100.00%)	84.40 (+2.18%)	84.40 (+2.18%)	84.40 (+2.18%)
NIAH-3	21.40	0.00 (−100.00%)	0.00 (−100.00%)	0.00 (−100.00%)	21.40 (+0.00%)	21.80 (+1.87%)	21.80 (+1.87%)
WikiText-103 PPL	14.6004	17.3538 (+18.86%)	26.8624 (+83.98%)	51.6760 (+253.94%)	14.6175 (+0.12%)	14.5953 (−0.03%)	14.5937 (−0.05%)
Ten-task mean	46.59	25.17 (−45.97%)	19.81 (−57.48%)	16.96 (−63.60%)	48.09 (+3.22%)	48.15 (+3.34%)	48.48 (+4.05%)

Table 19: **1.3B GDN,** 3:1**.** Accuracy (%) and PPL; relative changes in parentheses.

Benchmark	Normal	PAS suppression			Spike connection		
		$k=1$	$k=2$	$k=4$	$k=1$	$k=2$	$k=4$
PIQA	73.45	72.96 (−0.67%)	72.74 (−0.96%)	65.94 (−10.22%)	73.34 (−0.15%)	73.50 (+0.07%)	73.56 (+0.15%)
ARC-Easy	72.26	71.55 (−0.99%)	71.09 (−1.63%)	49.37 (−31.68%)	72.18 (−0.12%)	72.22 (−0.06%)	72.14 (−0.17%)
HellaSwag	60.25	59.16 (−1.80%)	59.43 (−1.36%)	45.17 (−25.02%)	60.12 (−0.21%)	60.16 (−0.15%)	60.13 (−0.20%)
FDA	67.57	60.04 (−11.16%)	54.31 (−19.62%)	16.80 (−75.13%)	69.57 (+2.96%)	69.21 (+2.42%)	68.76 (+1.75%)
SWDE	46.95	44.05 (−6.19%)	42.17 (−10.18%)	38.89 (−17.17%)	50.33 (+7.19%)	50.42 (+7.39%)	50.52 (+7.58%)
SQuAD	42.86	50.22 (+17.16%)	54.11 (+26.25%)	48.20 (+12.46%)	50.25 (+17.24%)	49.21 (+14.81%)	47.67 (+11.21%)
NQ	28.06	27.56 (−1.81%)	24.93 (−11.17%)	14.76 (−47.40%)	28.92 (+3.05%)	28.73 (+2.37%)	28.86 (+2.82%)
NIAH-1	98.60	99.40 (+0.81%)	99.40 (+0.81%)	35.00 (−64.50%)	98.40 (−0.20%)	98.60 (+0.00%)	98.60 (+0.00%)
NIAH-2	99.20	71.00 (−28.43%)	85.40 (−13.91%)	24.40 (−75.40%)	99.20 (+0.00%)	99.20 (+0.00%)	99.20 (+0.00%)
NIAH-3	65.20	52.20 (−19.94%)	40.20 (−38.34%)	9.60 (−85.28%)	60.20 (−7.67%)	59.80 (−8.28%)	61.20 (−6.13%)
WikiText-103 PPL	9.4945	9.5874 (+0.98%)	9.5905 (+1.01%)	14.7169 (+55.00%)	9.4952 (+0.01%)	9.4986 (+0.04%)	9.4987 (+0.04%)
Ten-task mean	65.44	60.81 (−7.07%)	60.38 (−7.74%)	34.81 (−46.80%)	66.25 (+1.24%)	66.11 (+1.01%)	66.06 (+0.95%)

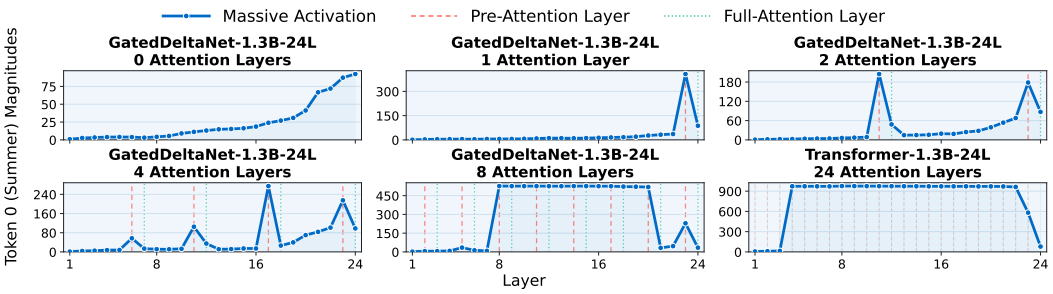
F.3 Exploratory Extension to Qwen3.5-27B

We evaluate reference-conditioned Top-4 spike-to-plateau connection on Qwen3.5-27B without updating model weights. The evaluation covers scientific reasoning, mathematics, and code generation. Table 20 reports paired comparisons between normal inference and reference-conditioned connection on the same evaluation examples.

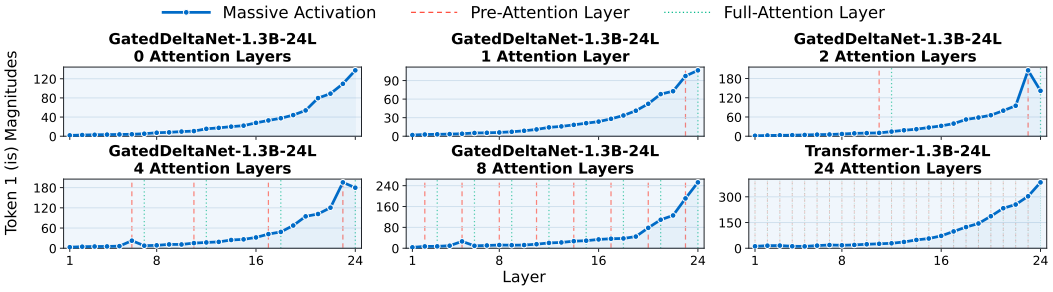
Table 20: **Qwen3.5-27B with Top-4 spike-to-plateau connection.** Scores are percentages; changes are relative to normal inference.

Benchmark	Normal (%)	Connect (%)	Relative Δ (%)
GPQA-Diamond	77.78	78.79	+1.30
MATH500	98.20	98.40	+0.20
LiveCodeBench	87.11	88.09	+1.12
AIME2025	80.00	80.00	+0.00
HumanEval	96.95	98.17	+1.26

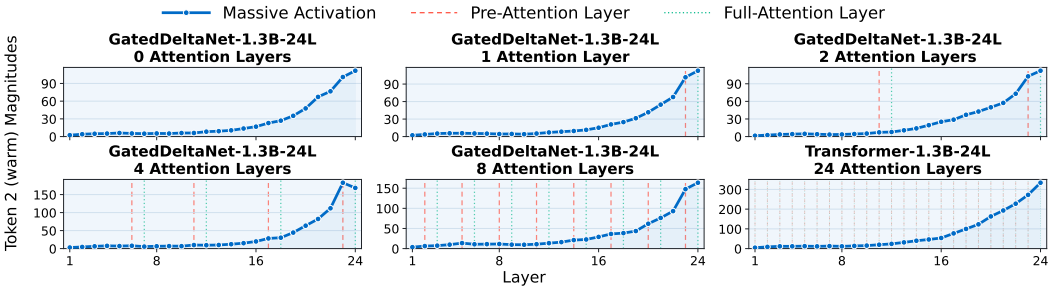
Connection yields small net improvements on GPQA-Diamond, MATH500, LiveCodeBench and HumanEval, with unchanged AIME2025 accuracy.



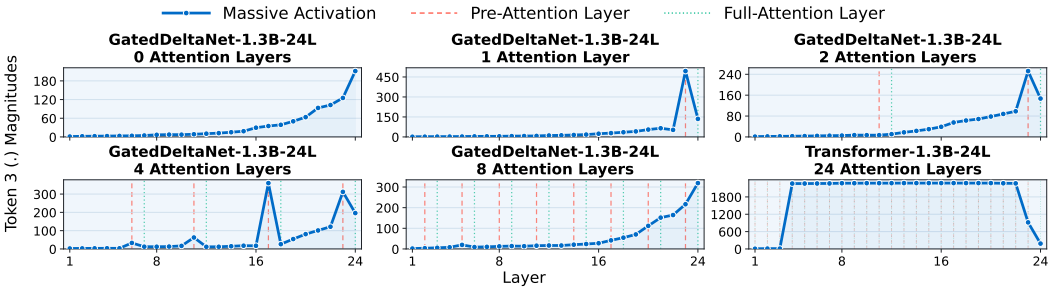
(a) Token 0: “Summer”



(b) Token 1: “is”

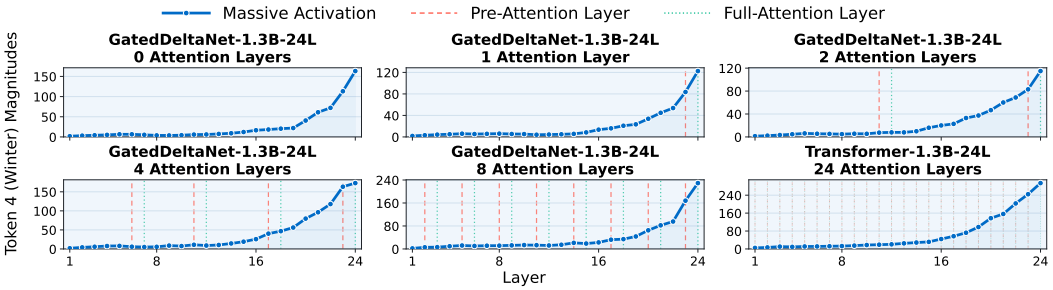


(c) Token 2: “warm”

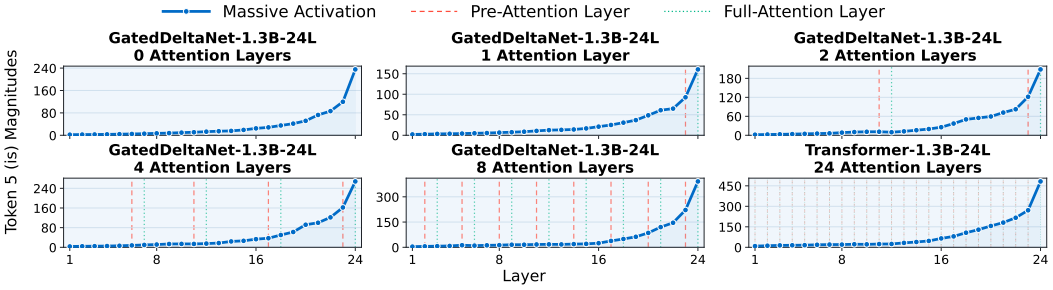


(d) Token 3: “.”

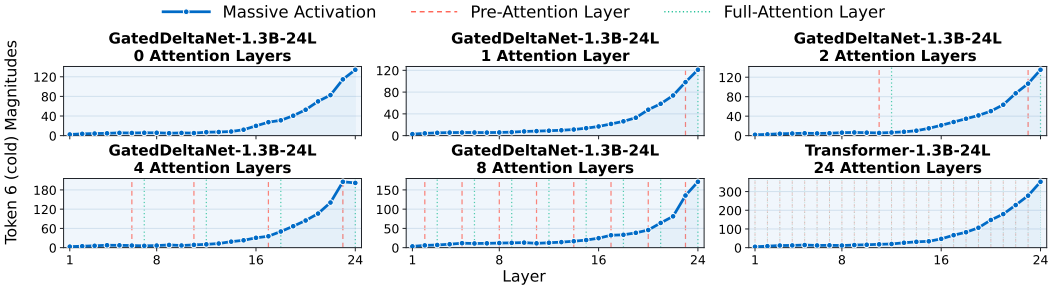
Figure 16: **MA dynamics across token positions in GDN.** Each panel traces the maximum absolute hidden-state activation of one token in *“Summer is warm. Winter is cold.”* across pure linear attention, the 24:1, 12:1, 6:1, and 3:1 hybrid configurations, and full attention. The attention-sink positions, “Summer” and the first period, exhibit pronounced PAS and increasingly connected ISP as full attention becomes denser, whereas non-sink positions exhibit neither morphology. This figure extends the first-token analysis in §3.2; detailed token-wise discussion is provided in Appendix B.3.



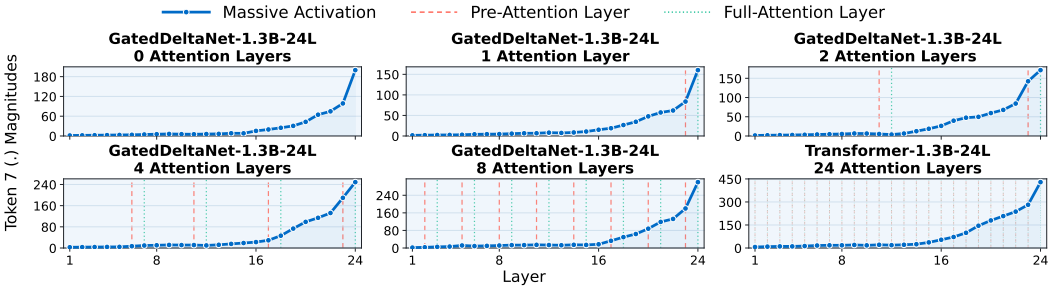
(e) Token 4: “Winter”



(f) Token 5: “is”

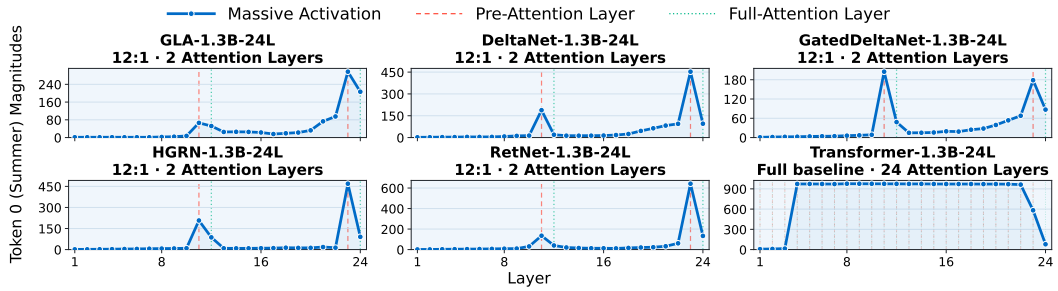


(g) Token 6: “cold”

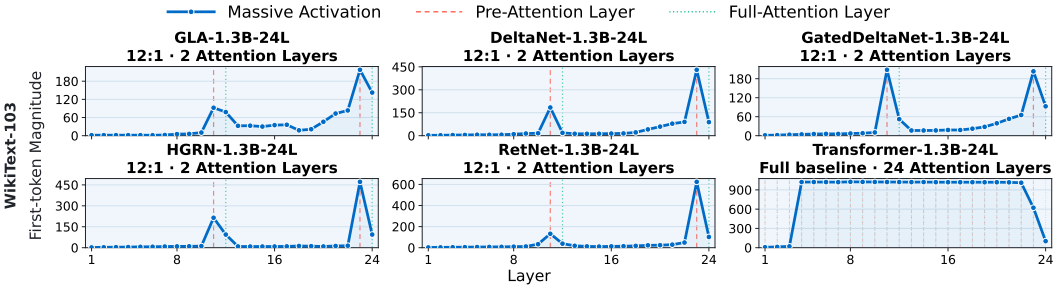


(h) Token 7: “.”

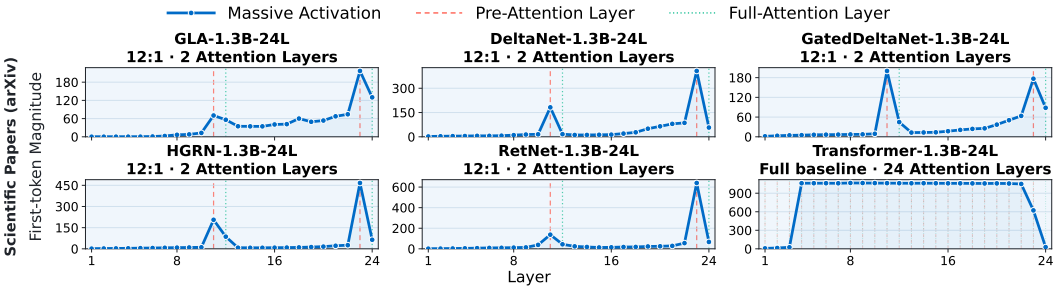
Figure 16: **MA dynamics across token positions in GDN (continued).**



(a) Running example: *"Summer is warm. Winter is cold."*

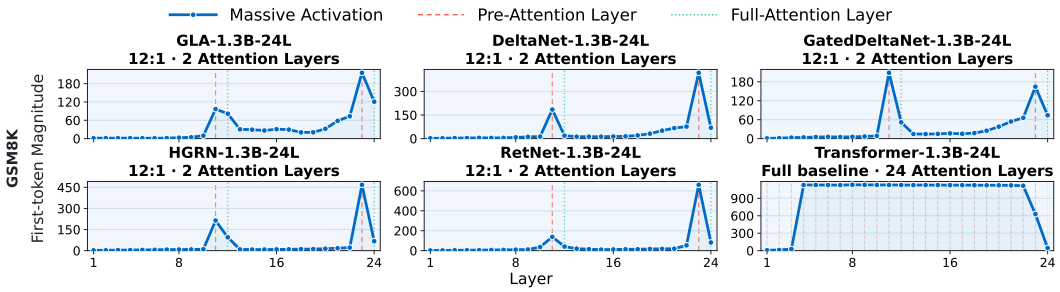


(b) General prose from WikiText-103.

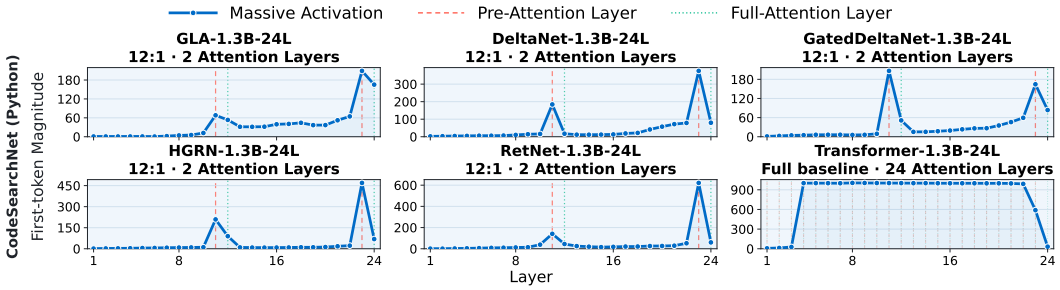


(c) Scientific writing from Scientific Papers.

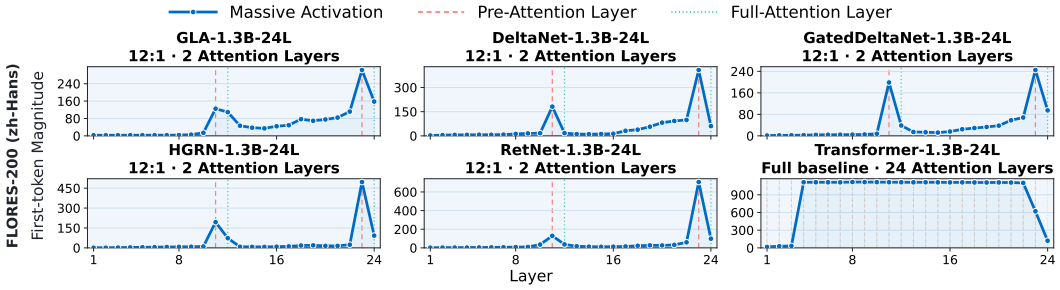
Figure 17: **MA dynamics across input domains in the M-A-P model suite.** Each panel traces the maximum absolute hidden-state activation of the first token across depth for five 1.3B HLA architectures under a fixed 12:1 hybridization ratio, with a full attention Transformer included as a reference. Across the running example and five representative domain-specific inputs, PAS consistently emerge immediately before full attention layers. This figure extends the cross-domain analysis in §3.2; detailed discussion is provided in Appendix B.4.



(d) Mathematical reasoning from GSM8K.

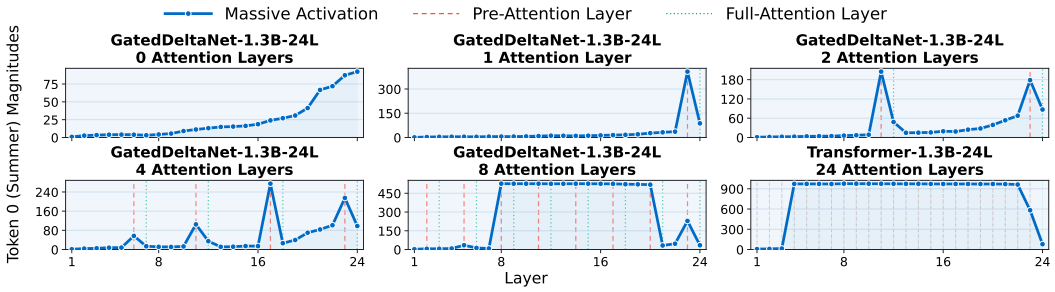


(e) Python code from CodeSearchNet.

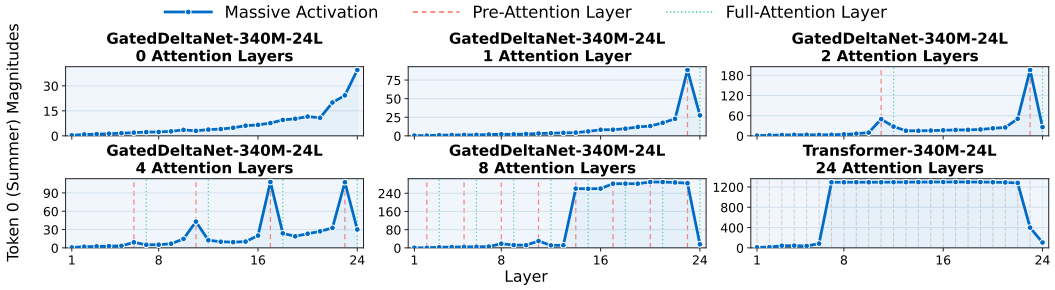


(f) Multilingual text from FLORES-200.

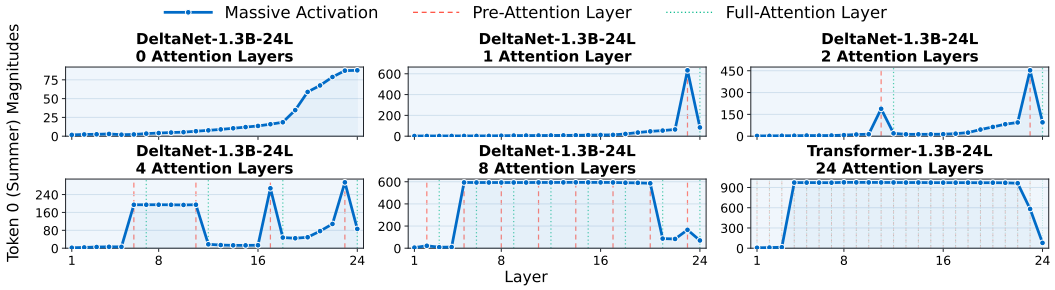
Figure 17: **MA dynamics across input domains in the M-A-P model suite (continued).**



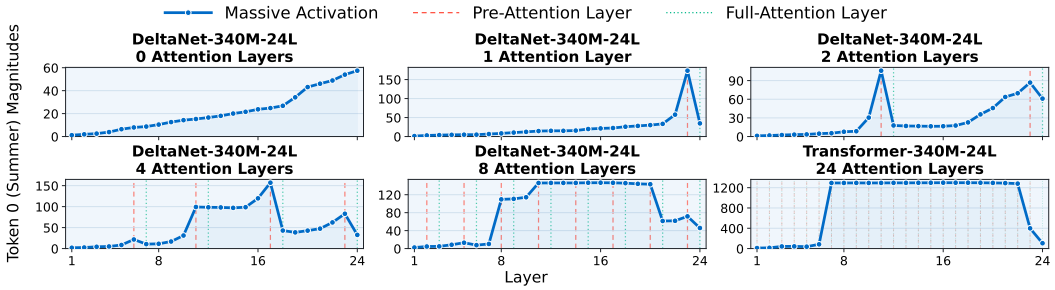
(a) GDN, 1.3B



(b) GDN, 340M

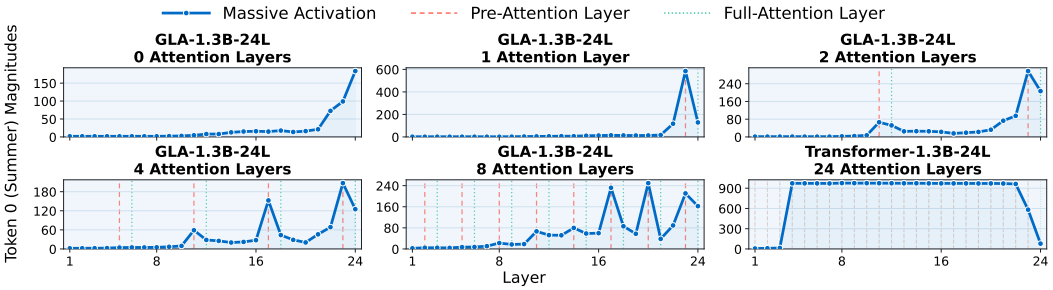


(c) DeltaNet, 1.3B

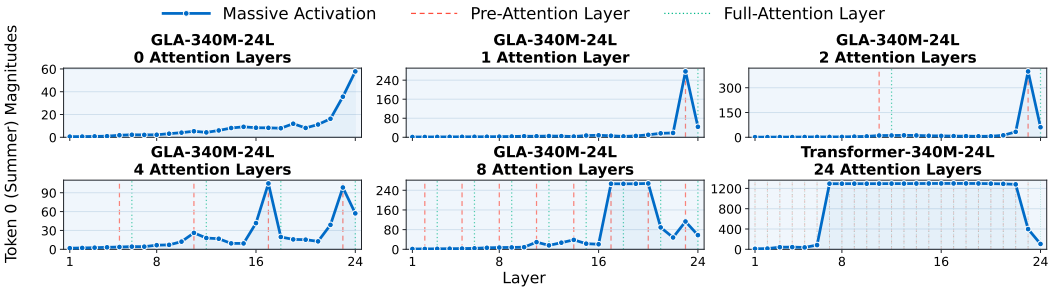


(d) DeltaNet, 340M

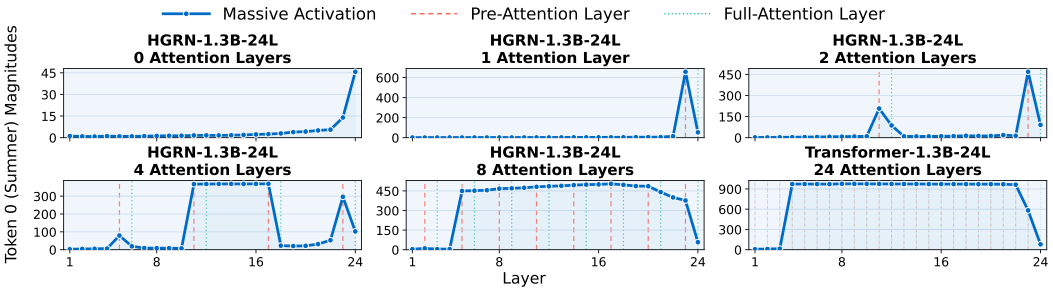
Figure 18: **MA dynamics across hybridization ratios in the M-A-P model suite.** Each panel traces the maximum absolute hidden-state activation of the first token across pure linear attention, the 24:1, 12:1, 6:1, and 3:1 hybrid configurations, and full attention. Across architectures and model scales, isolated PAS become progressively connected through ISP as full attention becomes denser. Further discussion appears in §3.3 and Appendix B.5.



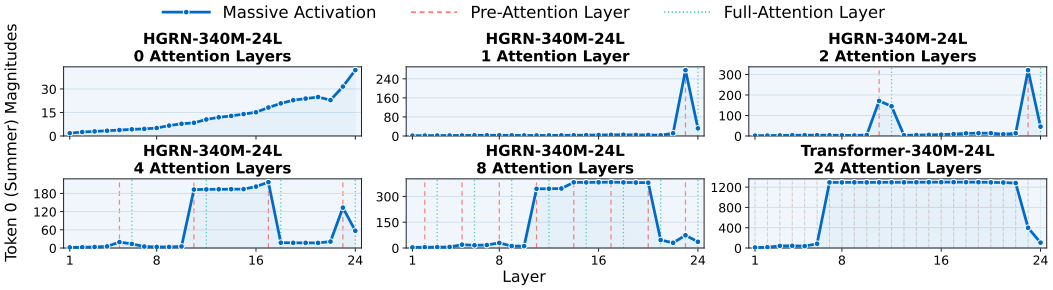
(e) GLA, 1.3B



(f) GLA, 340M

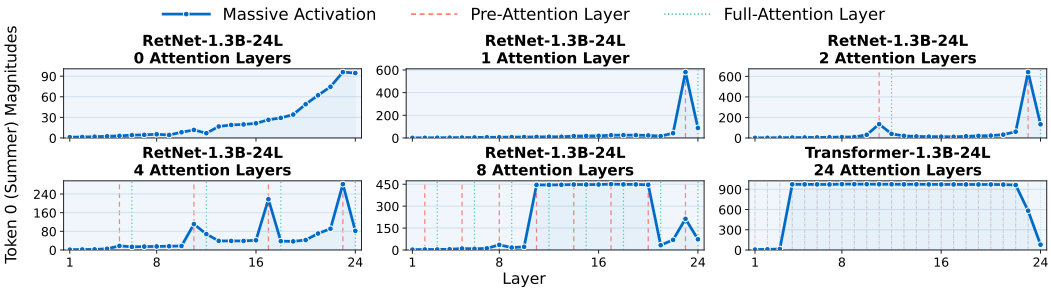


(g) HGRN, 1.3B

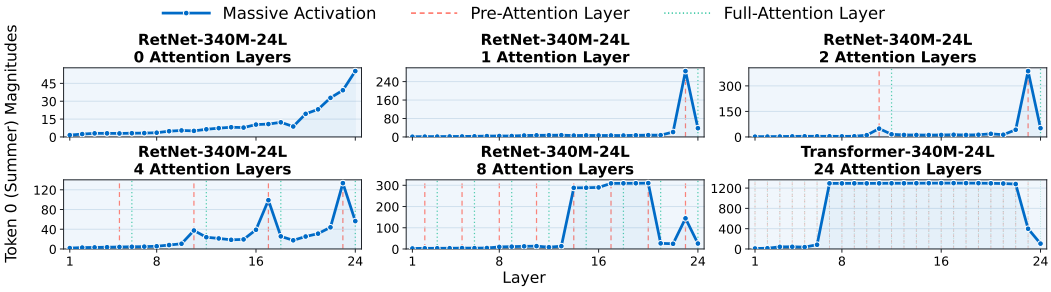


(h) HGRN, 340M

Figure 18: **MA dynamics across hybridization ratios in the M-A-P model suite (continued).**

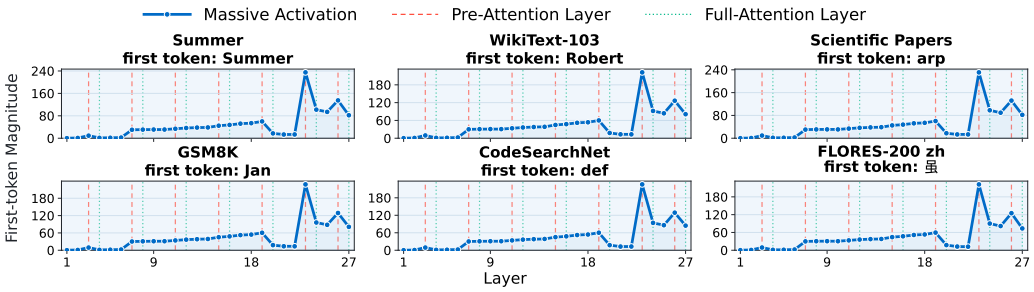


(i) RetNet, 1.3B

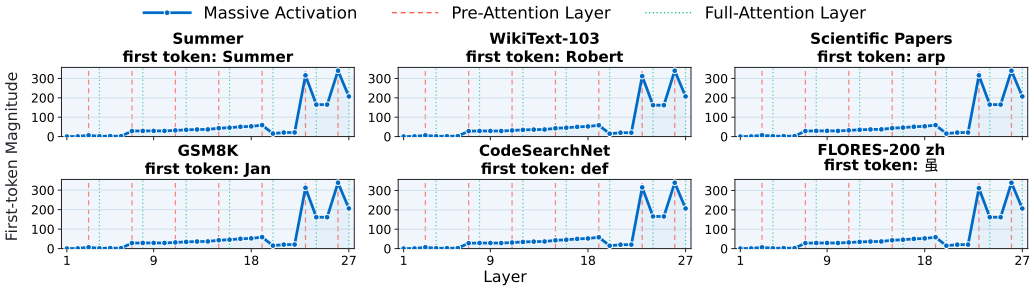


(j) RetNet, 340M

Figure 18: **MA dynamics across hybridization ratios in the M-A-P model suite (continued).**

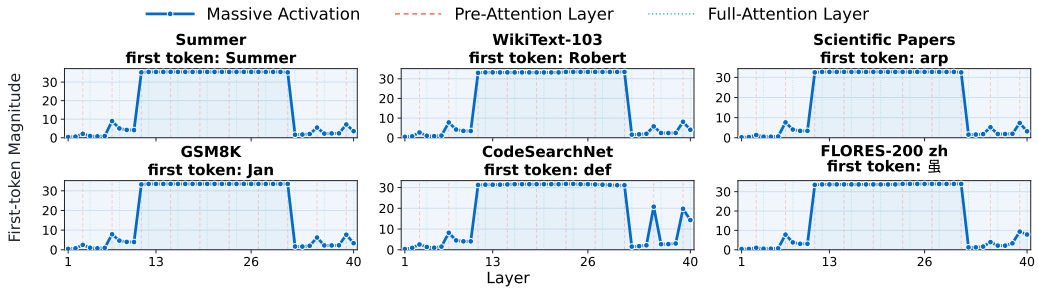


(a) Kimi-Linear-48B-A3B-Base

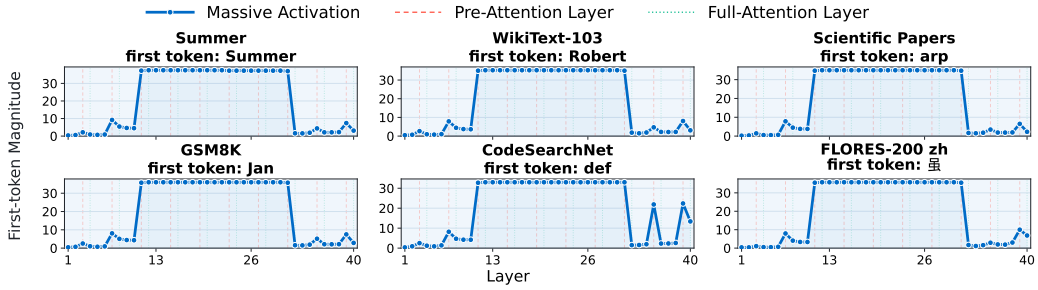


(b) Kimi-Linear-48B-A3B-Instruct

Figure 19: **MA dynamics in large-scale pretrained hybrid models.** Each checkpoint panel presents first-token activation trajectories for the running example and five domain-specific inputs. Results are grouped by family to compare matched base/instruction-tuned checkpoints, parameter scales, and linear attention versus state-space backbones. PAS and ISP remain aligned with the hybrid layer arrangement across these settings, while their absolute magnitudes vary across checkpoints and inputs. Detailed cross-model analysis is provided in Appendix C.2.

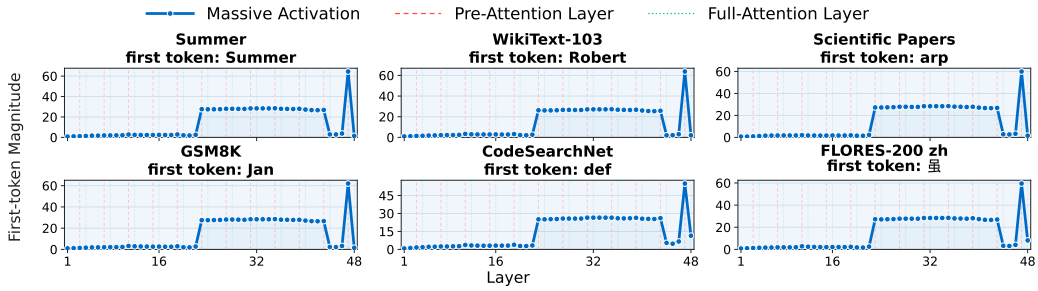


(c) Qwen3.5-35B-A3B-Base

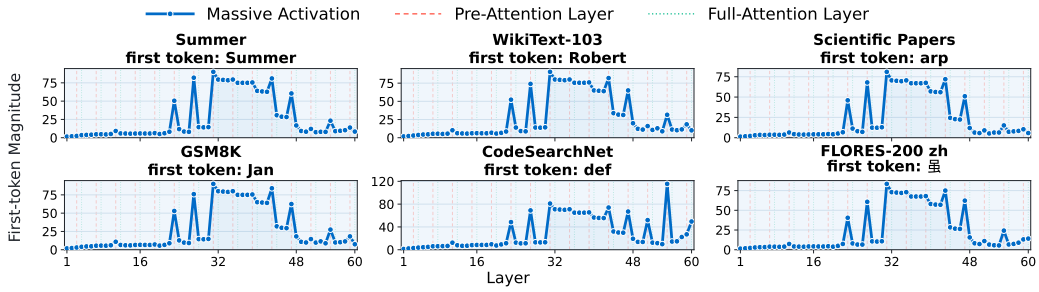


(d) Qwen3.5-35B-A3B

Figure 19: **MA dynamics in large-scale pretrained hybrid models (continued).** The matched Qwen3.5-35B Base and instruction-tuned checkpoints exhibit similar PAS locations and plateau spans, despite differences in absolute activation magnitude.

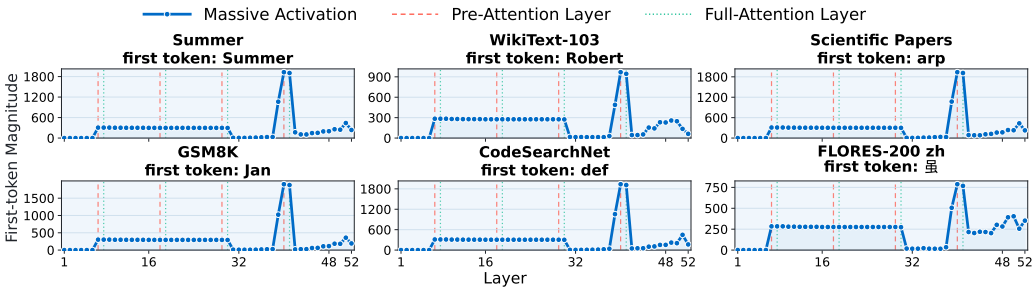


(e) Qwen3.5-122B-A10B

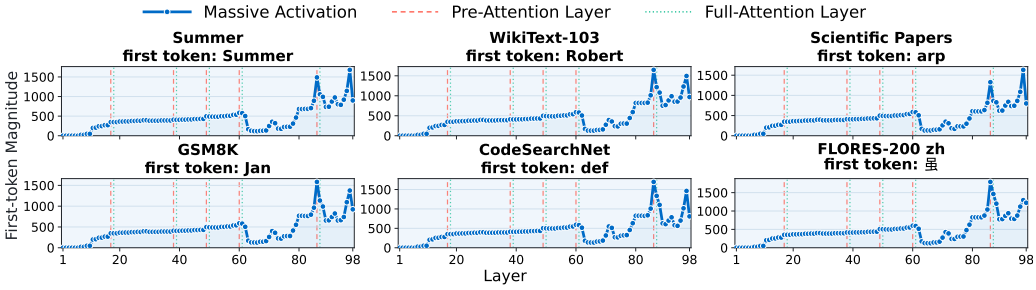


(f) Qwen3.5-397B-A17B

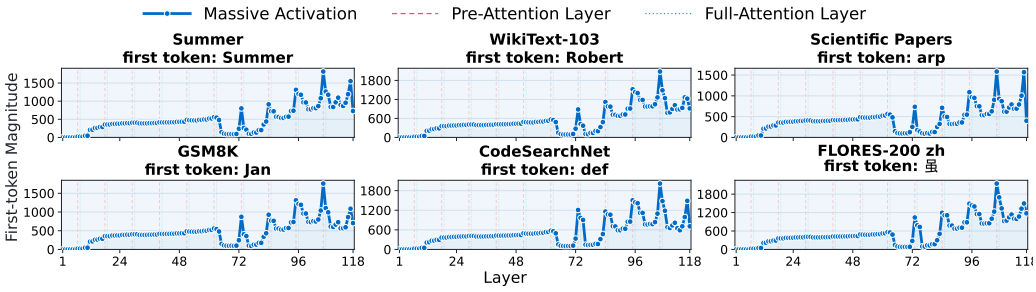
Figure 19: **MA dynamics in large-scale pretrained hybrid models (continued).** Qwen3.5 checkpoints at the 122B and 397B scales preserve the architecture-aligned PAS–ISP organization, although their activation magnitudes and plateau profiles differ.



(g) Nemotron-H-8B-Base-8K

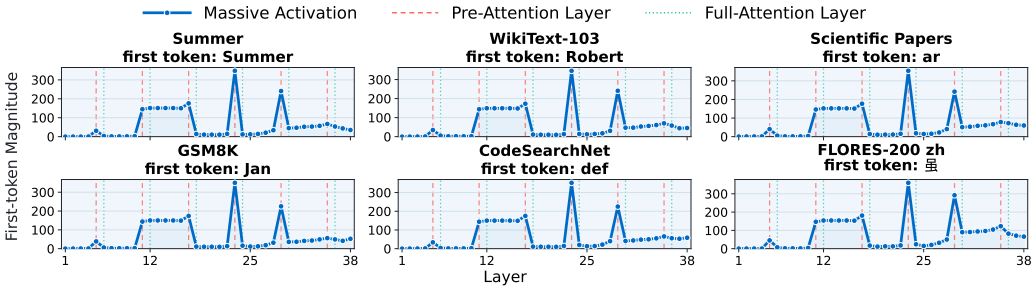


(h) Nemotron-H-47B-Base-8K

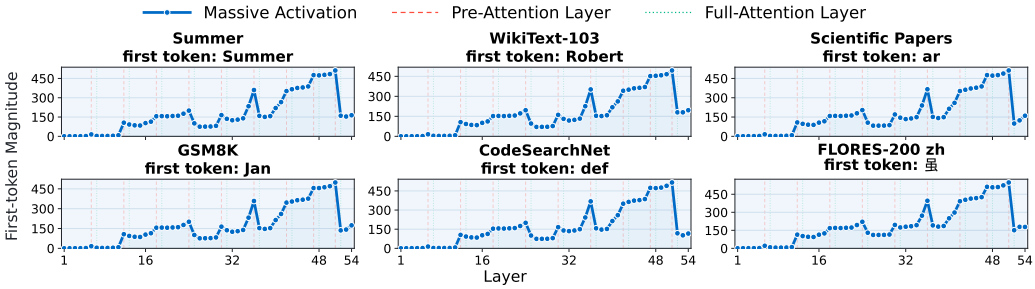


(i) Nemotron-H-56B-Base-8K

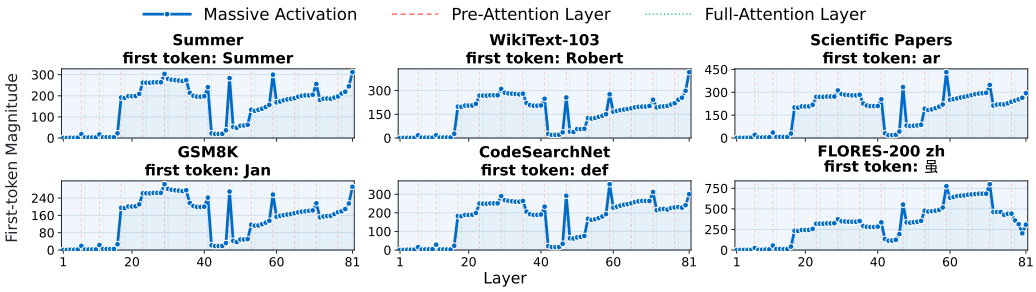
Figure 19: **MA dynamics in large-scale pretrained hybrid models (continued).** Nemotron-H checkpoints at the 8B, 47B, and 56B scales exhibit PAS- and ISP-like activation patterns aligned with the placement of full attention layers, extending the observed organization to state-space hybrids.



(j) Zamba2-1.2B



(k) Zamba2-2.7B



(l) Zamba2-7B

Figure 19: **MA dynamics in large-scale pretrained hybrid models (continued).** Zamba2 checkpoints from 1.2B to 7B parameters further demonstrate that architecture-aligned spike and plateau morphologies recur across model scales within this state-space hybrid family.

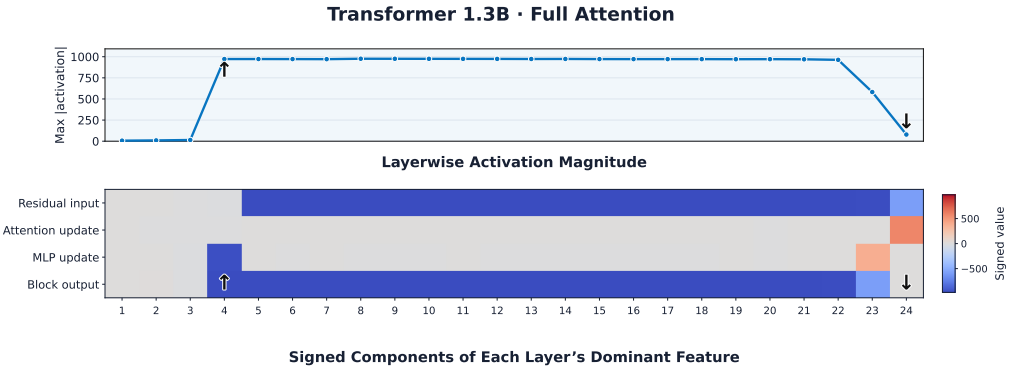
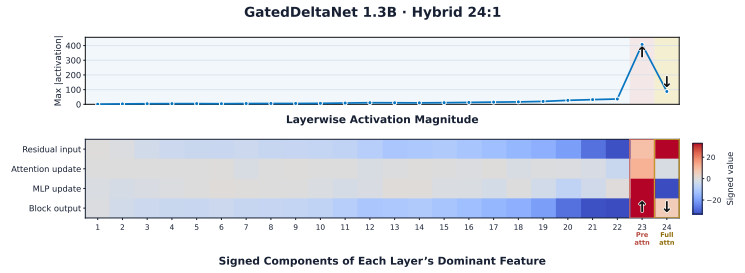
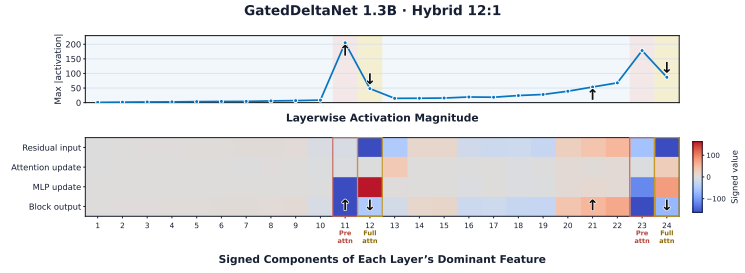


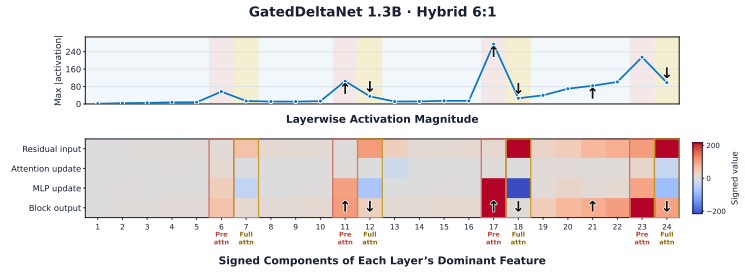
Figure 20: **Layerwise systematic-outlier pattern in the 1.3B full attention Transformer from the M-A-P suite.** MAs persist across much of the model depth, providing the full attention reference for the PAS–ISP continuum.



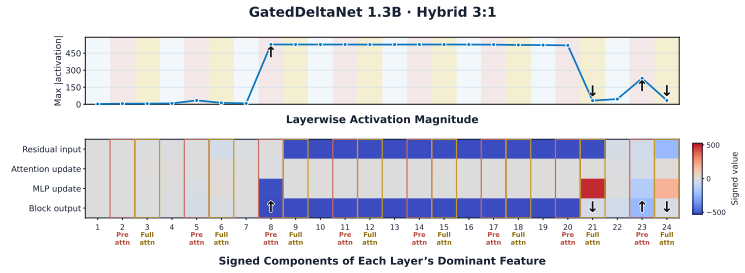
(a) GDN, 24:1.



(b) GDN, 12:1.

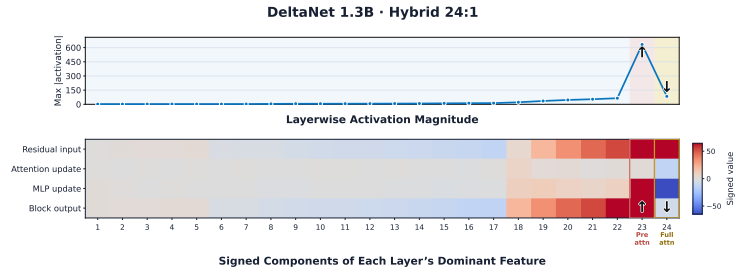


(c) GDN, 6:1.

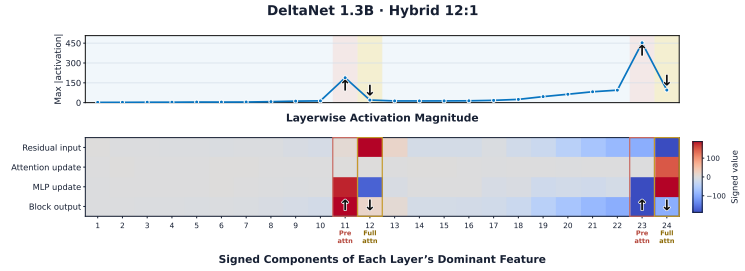


(d) GDN, 3:1.

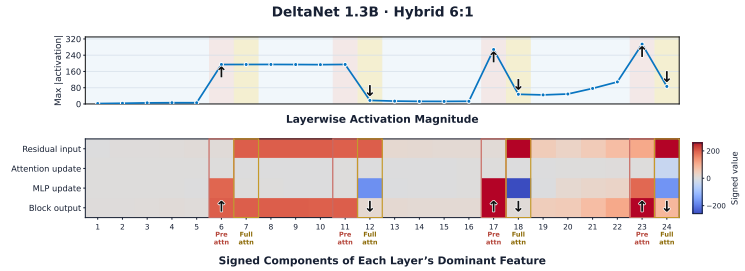
Figure 21: **Layerwise systematic-outlier patterns across hybridization ratios in 1.3B GDN models from the M-A-P suite.** This figure extends the hybridization-ratio analysis in §3.3; additional results and discussion are provided in Appendices B.5 and E.2.



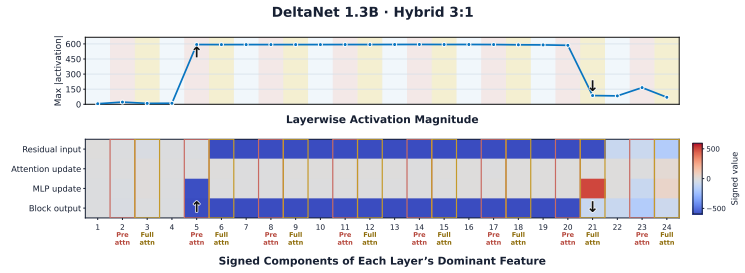
(a) DeltaNet, 24:1.



(b) DeltaNet, 12:1.

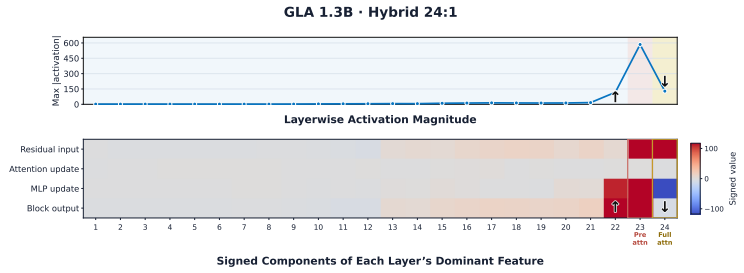


(c) DeltaNet, 6:1.

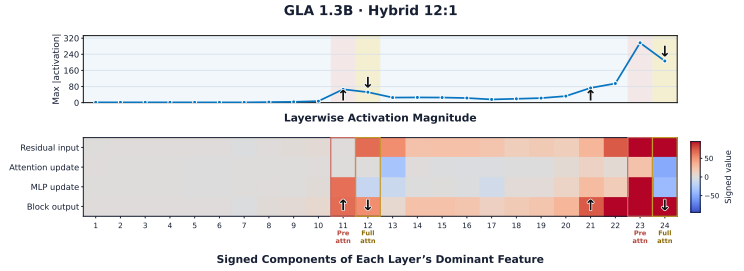


(d) DeltaNet, 3:1.

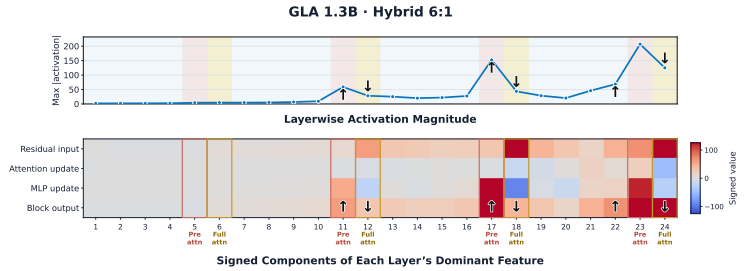
Figure 22: **Layerwise systematic-outlier patterns across hybridization ratios in 1.3B DeltaNet models from the M-A-P suite.** This figure extends the hybridization-ratio analysis in §3.3; additional results and discussion are provided in Appendices B.5 and E.2.



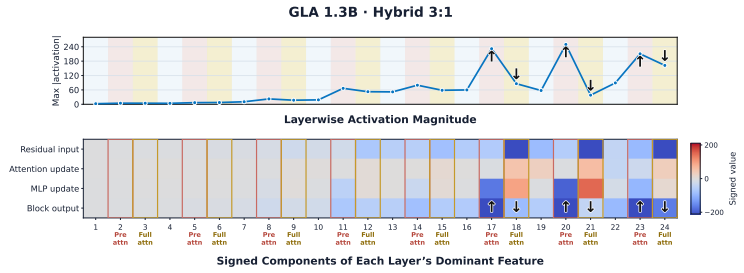
(a) GLA, 24:1.



(b) GLA, 12:1.

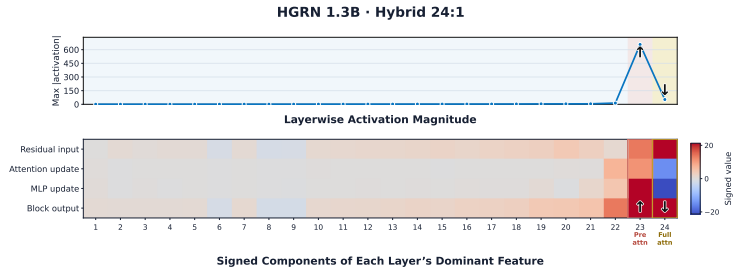


(c) GLA, 6:1.

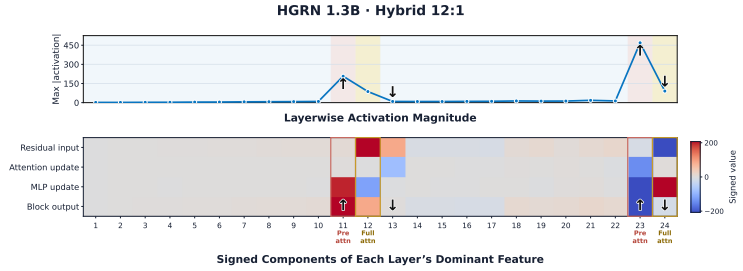


(d) GLA, 3:1.

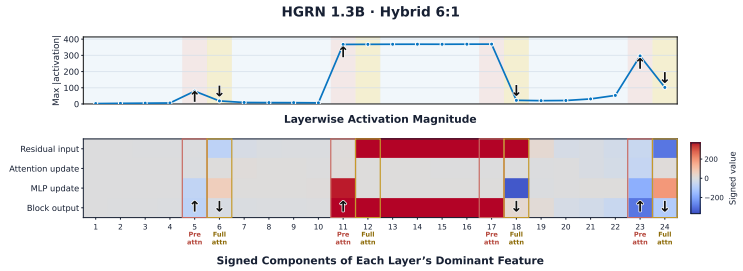
Figure 23: **Layerwise systematic-outlier patterns across hybridization ratios in 1.3B GLA models from the M-A-P suite.** This figure extends the hybridization-ratio analysis in §3.3; additional results and discussion are provided in Appendices B.5 and E.2.



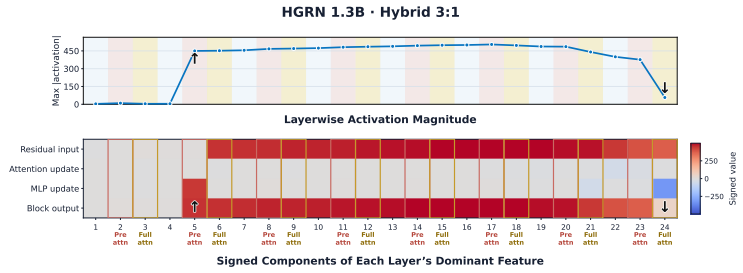
(a) HGRN, 24:1.



(b) HGRN, 12:1.

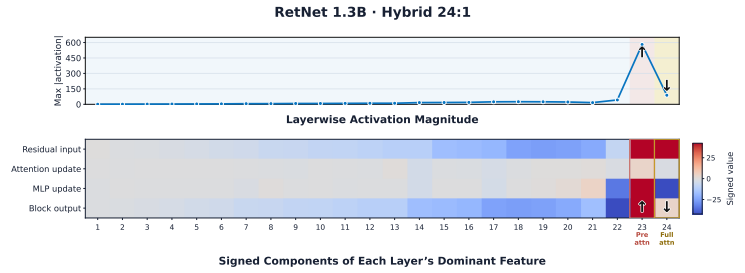


(c) HGRN, 6:1.

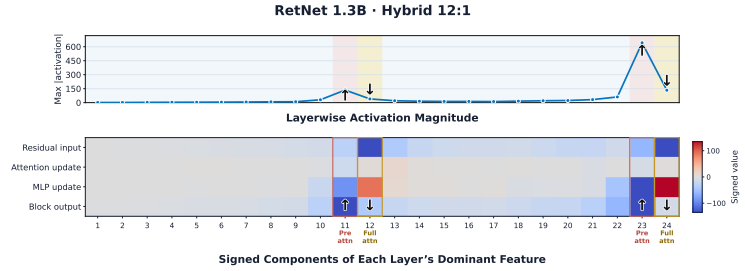


(d) HGRN, 3:1.

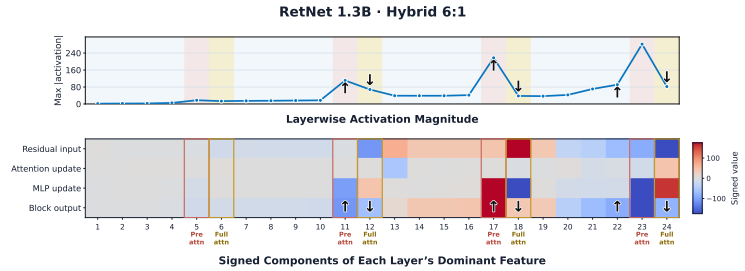
Figure 24: **Layerwise systematic-outlier patterns across hybridization ratios in 1.3B HGRN models from the M-A-P suite.** This figure extends the hybridization-ratio analysis in §3.3; additional results and discussion are provided in Appendices B.5 and E.2.



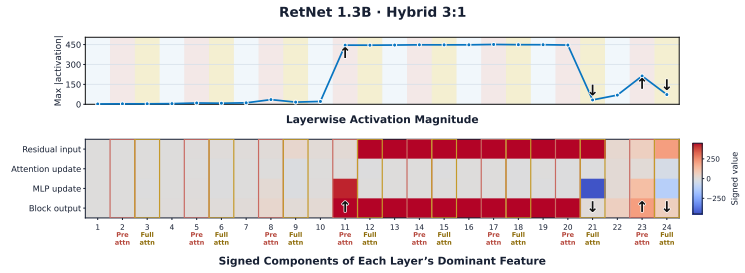
(a) RetNet, 24:1.



(b) RetNet, 12:1.

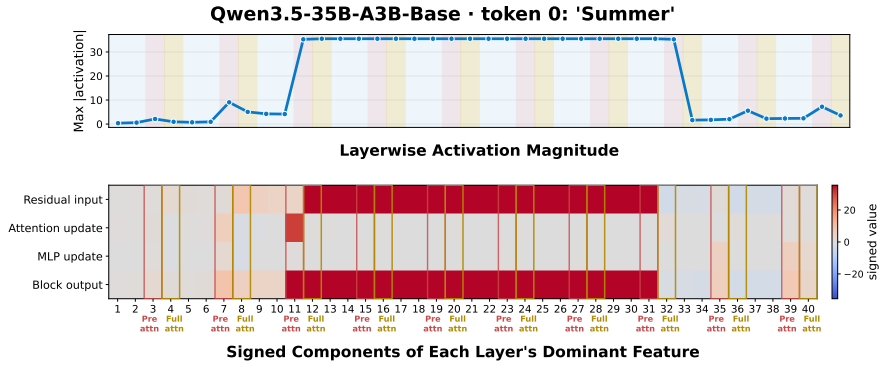


(c) RetNet, 6:1.

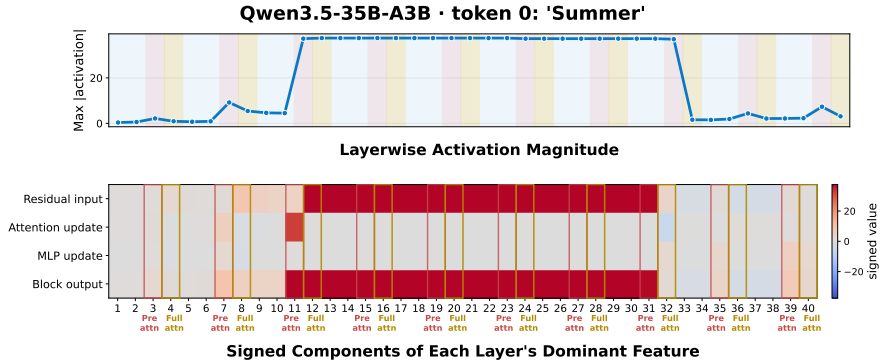


(d) RetNet, 3:1.

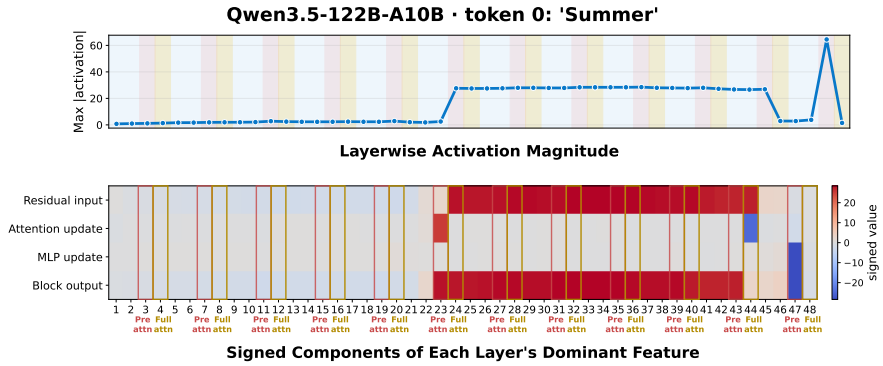
Figure 25: **Layerwise systematic-outlier patterns across hybridization ratios in 1.3B RetNet models from the M-A-P suite.** This figure extends the hybridization-ratio analysis in §3.3; additional results and discussion are provided in Appendices B.5 and E.2.



(a) Qwen3.5-35B-A3B-Base

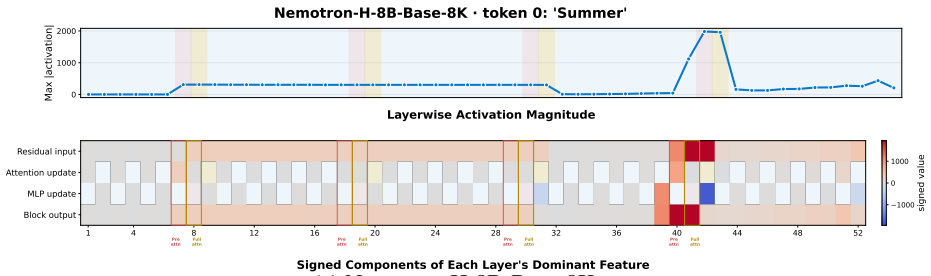


(b) Qwen3.5-35B-A3B

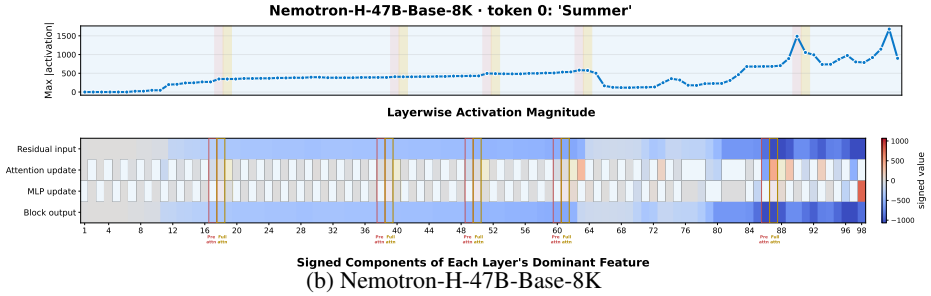


(c) Qwen3.5-122B-A10B

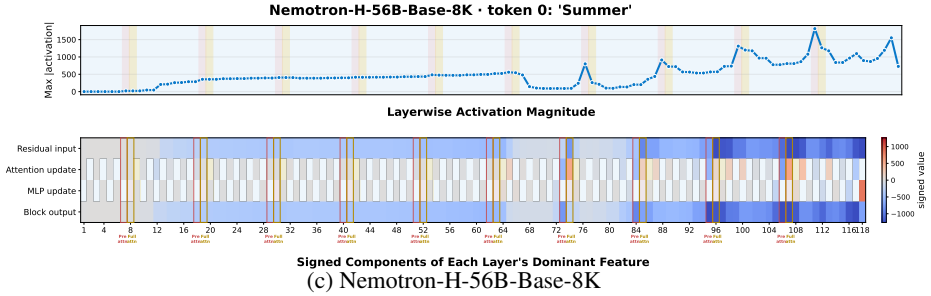
Figure 26: **Layerwise systematic-outlier patterns in large-scale Qwen3.5 linear attention hybrid models.** Each panel traces the first-token MA trajectory for *"Summer is warm. Winter is cold."* together with its signed module-level decomposition. Across the evaluated scales and post-training stages, the dominant MA transitions remain systematically aligned with full attention placement. PAS and ISP persist despite the native output gating in Qwen3.5. This figure extends §3.4; additional results are provided in Appendices C.2 and E.2.



(a) Nemotron-H-8B-Base-8K



(b) Nemotron-H-47B-Base-8K



(c) Nemotron-H-56B-Base-8K

Figure 27: **Layerwise systematic-outlier patterns in large-scale Nemotron-H state-space hybrids.** Each panel traces the first-token MA trajectory for *"Summer is warm. Winter is cold."* together with its signed module-level decomposition. Across the evaluated checkpoints and layer schedules, dominant outlier amplification and opposing updates recur near full attention boundaries, extending the observed architecture-aligned pattern beyond linear attention backbones to hybrids interleaving Mamba-2 and full attention. This figure complements the analysis in §3.4; additional results across input domains are provided in Appendix C.2.