Detection vs. Execution: Single-Bucket Probes Miss Half the Mamba-2 State Sink

Anonymous ACL submission

Abstract

Mechanistic interpretability often assumes that probes identifying a representational signature also identify the circuit executing the corresponding computation. We show that this assumption can fail systematically in Mamba-2. Studying the *state sink* (disproportionate Δ-gate activation on boundary tokens, analogous to the attention sink), we find that single-bucket probes recover only a small execution layer while missing a much larger detection layer with the same representational signature.

In Mamba-2, the state sink decomposes into two functional head sets. Single-bucket BOS-specialist heads (about 5% of heads at 2.7B) causally support both BOS-context and newline-target predictions across model scales and corpora. Dual heads (27–35$\%$ of heads, recovered by multi-class aggregation of the same probe) show stronger BOS–newline representational similarity but substantially weaker causal effects under ablation. Representational similarity does not imply functional equivalence.

This distinction matters for downstream behaviour: ablating BOS-specialist heads collapses RULER NIAH retrieval accuracy from 1.00 to 0.00 at 1024 context length in both Mamba-1 2.8B and Mamba-2 2.7B, while size-matched complements preserve baseline performance. A random channel-bucketing control rules out substrate granularity alone, implicating Mamba-2's head-shared Δ projection. Probe-derived specialty can identify execution circuits; at coarse granularity the same probe also recovers detection circuits, and separating them requires class-conditional ablation rather than class-conditional cosine.

1 Introduction

When mechanistic interpretability locates a circuit in a language model, it typically reads a probe and labels the highest-activating units as the circuit. We show this label-to-locus move is reliable at channel granularity (Mamba-1) but systematically undercounts the causal substrate by 3–7$\times$ in set size at head granularity (Mamba-2): the difference is architectural, not methodological. The implication is that interpretability findings localised at Mamba-2 heads under single-bucket probing warrant re-audit.

Decoder-only Transformer language models allocate disproportionate attention mass to the first token, the *attention sink* (Xiao et al., 2024; Gu et al., 2025; Cancedda, 2024). Recent work has made attention sinks both localisable and architecturally suppressible: Qiu et al. (2025) report that a head-specific sigmoid gate after SDPA reduces average first-token attention from 46.7% to 4.8%. We use *state sink* to refer to the analogous phenomenon in selective state-space language models (Gu and Dao, 2024; Dao and Gu, 2024): the disproportionate allocation of token-conditional gate Δ activation to boundary tokens (BOS, newline) that anchors the hidden state. Attention sink is measured in attention-score space; state sink is measured in Δ-gate space. Mechanistically the two sinks are distinct: attention sink emerges from the Softmax sum-to-one constraint that forces residual mass onto a low-information anchor (Xiao et al., 2024), while state sink arises from the SSM's selective gating, which independently elevates Δ at boundary tokens without any normalisation trade-off. Related SSM work supplies adjacent operationalisations (Chiang et al., 2025; Ye et al., 2025), including activation-outlier magnitude and hidden-state survival.

The labelling assumption that the top-activating units are the causal locus has not been tested at intervention scale on selective state-space substrates. The contribution is the way it breaks at Mamba-2 head granularity: the state sink decomposes into two functional head sets at the same depth (a functional partition, not a depth-wise architectural one) that single-bucket probes cannot see together. Single-bucket bos-specialists ($\sim 5\%$ of heads at 2.7B) are the execution layer; they carry the causal

mass for both BOS-context and newline-target predictions across all six (scale, corpus) cells we test (three M-2 scales: 130M / 1.3B / 2.7B). Dual heads are a 3–7× larger set (27–35% of heads), recovered from the same probe under a multi-class aggregation rule, and form a representational detection layer: their Δ_{pre} vectors are near-collinear on BOS and newline ($\cos = 0.89$ vs. 0.54 on bos-specialists), yet their ablation under-performs the bos-specialist set on newline targets in all six cells. At Mamba-1 channel granularity the two layers coincide and single-bucket labels track causal mass cleanly.

The intervention grid is pre-registered: 12,096 cells total, 6,060 evaluated (parity / Dyck-2 task loaders deferred per pre-registration). The grid is a six-factor sweep over four Mamba checkpoints (Mamba-1 130M / 2.8B, Mamba-2 1.3B / 2.7B), six interventions on the gate / time-step / input projection (`gate_zero`, `gate_one`, `gate_mean`, `delta_zero`, `delta_median`, `u_zero`), seven probe-derived channel or head scopes, six layer scopes, and two corpora (wikitext-2 (Merity et al., 2017), Pile-10k (Gao et al., 2020) for cross-dataset replication of the headline cells). The Mamba-2 head-granularity headline (both scales × both gate interventions on both corpora; 8 cells) additionally uses a 30-seed random-complement bank. The work extends the activation-patching and ablation tradition in mechanistic interpretability (Olsson et al., 2022; Conmy et al., 2023; Sharma et al., 2024) to selective state-space substrates. Headline cells are then ported to four Pythia attention checkpoints (Biderman et al., 2023) as a cross-architecture baseline, and to RULER Needle-in-a-Haystack (Hsieh et al., 2024) as a retrieval generalisation test.

Findings. Four claims, each cross-dataset, cross-scale, or cross-architecture replicated; full numbers in §4.

- **Granularity-conditional causality:** at Mamba-1 channel granularity, single-bucket bos-specialists carry the BOS sink mechanism with large positive differentials (cross-corpus matched to two significant figures at 130M); at Mamba-2 head granularity they remain positively causal but 2–3× smaller, and dual heads carry 4.8–6.4× the aggregate ablation effect (set-size driven; per-unit ∼1×).
- **Detection vs. execution (F1):** class-conditional ablation across three M-2 scales (130M / 1.3B / 2.7B, both corpora) shows bos-specialists are positively causal on both targets in all 6 cells; dual under-performs at newline in all 6, and dual × wt_bos is monotonic with scale. $\cos = 0.89$ on dual is representational, not functional (§4.3.4).
- **Architectural driver (T3):** random channel bucketing of Mamba-1 to $H/C = 20$ retains 29% bos-specialists rather than M-2's 5%, consistent with a head-shared Δ contribution rather than substrate granularity alone (§4.6).
- **Retrieval transfer (RULER NIAH):** at 1024 ctx, bos-specialist ablation drops needle-retrieval accuracy from 1.00 to 0.00 on Mamba-1 2.8B and Mamba-2 2.7B; complement ablation preserves 1.00. Pythia cross-architecture: four checkpoints show single-bucket selectivity without comparable dual-set dominance (§4.4).

Why granularity matters. Mamba-1 has thousands of channels per layer ($D_{\text{inner}} \in [1{,}536, 5{,}120]$) with channel-independent Δ projections; specialist sub-populations on different boundary classes co-exist without forcing multi-class activation. Mamba-2 has tens of heads per layer ($n_{\text{heads}} \in [24, 80]$) and the Δ projection is shared across the channels within each head, forcing the head's activation profile to multiplex across boundary classes (Gu and Dao, 2024; Dao and Gu, 2024). We interpret the detection layer (dual heads) as arising from this constraint; the single-bucket probe excludes it by definition.

2 Related Work

Attention sinks and BOS-token behaviour. Autoregressive Transformers can allocate disproportionate attention mass to initial tokens, the attention-sink phenomenon (Xiao et al., 2024); subsequent work characterises sink emergence during training (Gu et al., 2025), links it to spectral filters and "dark signals" (Cancedda, 2024), proposes causal-mask modifications (Yin et al., 2024), and connects sink-like behaviour to massive activations, head-level no-op/outlier patterns, and register tokens in vision Transformers (Sun et al., 2024; Bondarenko et al., 2023; Darcet et al., 2024). Qiu et al. (2025) show that a head-specific sigmoid gate after SDPA mitigates attention sinks, reducing average first-token attention from 46.7% to 4.8%. Work in this thread measures sink behaviour in attention-score space, typically by attention mass assigned

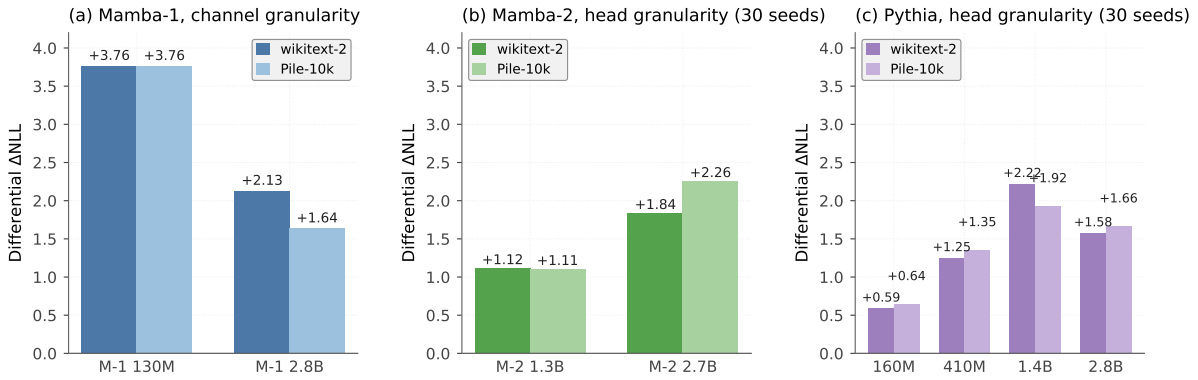

Figure 1: Granularity-conditional selectivity on `gate_zero` ablation of bos-specialist sets. Size-matched specialist–complement ΔNLL on first-eight-tokens-after-BOS for (a) Mamba-1 channels, (b) Mamba-2 heads (30-seed random-complement bank), and (c) Pythia attention heads. Bars: wikitext-2 (dark), Pile-10k (paler). Labels in nats.

to BOS or other initial positions; we test the analogous label-to-locus move in selective state-space models, where the token-conditional gate Δ has not been audited for sink-locus identification. Prior probes of sink-like behaviour operate at the activation distribution or hidden-state level (Chiang et al., 2025; Ye et al., 2025); we add unit-level resolution through channel- and head-targeted interventions.

Sink mass in selective state-space models. Chiang et al. (2025) identify SSM-input activation outliers (high-percentile of $|x|$) for post-training quantisation; Ye et al. (2025) study hidden-state memory decay and receptive-field limits in Mamba, using training-free token filtering to enlarge effective receptive fields. We complement these with two analytic baselines computed directly from Mamba's state: a right-tail $\Pr(\Delta_{\text{pre}} > 0.7)$ statistic for "reset" propensity, and a per-channel $-\log A$ baseline derived from the stored $A_{\log}$ parameter, the structural quantity governing state-decay in the SSM–attention duality (Dao and Gu, 2024). Whether thresholded sink sets agree across methods, and whether they predict causal effects, has not been audited at scale; we report a four-method cross-validation in Appendix A.

Mamba mechanistic interpretability and probes. Ali et al. (2025) reformulate Mamba's selective scan as a data-dependent attention matrix; Paulo et al. (2025) test whether selected Transformer interpretability methods transfer to RNNs/Mamba; Sharma et al. (2024) adapt the ROME (Meng et al., 2022) pipeline to Mamba and localise factual associations at mid-network SSM layers; and Ensign and Garriga-Alonso (2024) adapt IOI circuit analysis and positional edge attribution patching to Mamba. Linear probes are standard for representational analysis (Tenney et al., 2019; Belinkov, 2022; Hewitt and Liang, 2019; Voita and Titov, 2020; Ravichander et al., 2021; Elazar et al., 2021); causal-intervention work cautions that readable representations need not be behaviourally used and provides methods for testing that gap (Geiger et al., 2021; Vig et al., 2020; Heimersheim and Nanda, 2024; Belrose et al., 2023). Our intervention grid perturbs the token-conditional Δ gate that defines selective state-space models (Gu and Dao, 2024); the broader ablation, activation-patching, and automated-circuit-discovery tradition is established in (Wang et al., 2023; Olsson et al., 2022; Vig et al., 2020; Meng et al., 2022; Heimersheim and Nanda, 2024; Conmy et al., 2023; Huben et al., 2024; Templeton et al., 2024; Marks et al., 2025; Wu et al., 2025). Where prior probe-causality work showed that *what* a representation encodes is hard to read, we show that *where* a mechanism lives is hard in the same way, and that the difficulty is granularity-conditional.

Long-context retrieval and SSM limits. Pure SSMs have documented weaknesses relative to Transformers on copying, in-context-learning, and long-context reasoning tasks (Jelassi et al., 2024; Waleffe et al., 2024); we use single-key RULER (Kamradt, 2023; Hsieh et al., 2024) at 1024 context as a behavioural probe of the same causal axis, not as a benchmark contribution.

3 Phenomenology

3.1 Specialist classification protocol

For each (layer, channel) in Mamba-1 and (layer, head) in Mamba-2, we compute mean pre-softplus

activation $\bar{a}_{\ell,u,c}$ on tokens of class $c \in$ {BOS, newline, punctuation, whitespace_word} over a 30-document wikitext-2 sample, subtract the unit's mean activation on a reference (non-class) token set, and threshold the resulting differential $\Delta\bar{a}_{\ell,u,c}$ at $\tau = 0.5$ (pre-softplus units) to obtain a binary specialty matrix $S_{\ell,u,c} \in \{0, 1\}$. A unit is a *c-specialist* if exactly one class clears threshold; a *dual* specialist if two classes clear threshold; a *generalist* if three or more; *silent* otherwise. The threshold $\tau = 0.5$ is fixed across all 7 models and pre-registered to mark the gate's "on" state ($\mathrm{softplus}(0.5) \approx 0.97$); values below threshold map to the low-state mode of the bimodal pre-Δ distribution (§3.3). Threshold-sensitivity at $\tau \in \{0.3, 0.5, 0.7, 1.0\}$ is in Appendix A.

3.2 Per-token-class distributions across scale and architecture

The specialty fraction (bos-specialists / total units, averaged across layers) scales differently in the two architectures (Table 1). In Mamba-1, the fraction grows log-linearly with parameter count from 19.4% at 130M to 29.4% at 2.8B (least-squares fit $f = 7.4 \log_{10} N - 41.9$, $R^2 = 0.91$). In Mamba-2, the fraction caps at 4.9–8.2% across the same parameter range, non-monotonically (peak at 1.3B, 8.2%). The asymmetry tracks the absolute substrate budget: Mamba-1's hidden width grows from $D_{\text{inner}} = 1{,}536$ to 5,120, leaving room for an expanding specialist sub-population, whereas Mamba-2's head budget is bounded ($n_{\text{heads}} \in [24, 80]$ across the checkpoints we test) and the per-layer specialist count saturates well below the channel-granularity rate.

Within-model the fraction is highly heterogeneous across layers. At Mamba-1 130M, layer 0 alone is 76.0% bos-specialist while the median layer is 0.5%; at Mamba-1 2.8B, the most concentrated layer reaches 92.9% and layer 0 has fallen to 10.0%. The reorganisation is consistent with BOS handling localising at the bottom of the network in the smallest model and redistributing across deeper layers as scale grows (§4.2).

Pythia attention-head bos-specialists, defined in our operationalisation by mean attention-to-BOS strength > 0.10 and motivated by the attention-sink practice of measuring first-token attention mass (Xiao et al., 2024; Gu et al., 2025), account for 15% to 97% of heads per layer in mid-to-late layers, in line with prior reports of pervasive first-token sinks in Transformer LMs (Xiao et al., 2024).

Family	Model	units / layer	bos-spec	dual
Mamba-1	130M	1,536 ch	19.4%	13.8%
	370M	2,048 ch	20.2%	21.8%
	1.4B	4,096 ch	25.0%	23.8%
	2.8B	5,120 ch	29.4%	29.6%
Mamba-2	130M	24 heads	4.9%	13.4%
	1.3B	64 heads	8.2%	26.9%
	2.7B	80 heads	5.0%	35.4%

Table 1: Bos-specialist and dual specialty fractions (mean across layers) per model at $\tau = 0.5$ (pre-softplus units). The dual fraction converges to 24–35% on larger-substrate models in both architectures.

3.3 Sink-mass cross-validation and Δ-bimodality

We compute four per-unit sink-mass metrics (Chiang et al., 2025; Ye et al., 2025; Gu and Dao, 2024) and test whether LongMamba state-survival and right-tail Δ_{pre} reset-propensity are negatively correlated ($\rho_{L\Delta} < -0.3$, pre-registered). The prediction holds at *head* granularity in Mamba-2, with $\rho_{L\Delta} = -0.443$ at 1.3B, -0.437 at 2.7B, and a weaker -0.282 at 130M, and fails at *channel* granularity in Mamba-1, where $\rho_{L\Delta}$ stays in $[-0.02, +0.20]$ across the four scales we test. The head–vs–channel asymmetry converges with the dual-set dominance result of §4.3.

A second pre-registered prediction is that the selective-gate Δ distribution at sink-relevant positions exhibits bimodality. A 3-of-4 conjunction (Hartigan dip ≥ 0.05 (Hartigan and Hartigan, 1985), Silverman peak-count ≥ 2 (Silverman, 1981), GMM ΔBIC ≥ 10, KDE local-mode count ≥ 2) holds on 87.1–100.0% of (layer, bucket) cells across all four tested Mamba checkpoints, well above the pre-registered 50% floor. Per-model values, Quamba $\times$ A-log channel-level breakdown, top-10% Jaccard agreement, and the 4-of-4 strict-conjunction calibration (M-2 2.7B drops to 41.8% on the parametric ΔBIC arm) are in Appendix A and G.

4 Causal Analysis

4.1 Method

The intervention grid is a six-factor sweep: 4 models (Mamba-1 130M / 2.8B, Mamba-2 1.3B / 2.7B) $\times$ 6 interventions (`gate_zero` / `gate_one` / `gate_mean` / `delta_zero` / `delta_median` / `u_zero`) $\times$ 7 channel/head scopes $\times$ 6 layer scopes $\times$ 4 tasks $\times$ 3 seeds, totalling 12,096 cells of which 6,060 are evaluated (parity / Dyck-2 task loaders

deferred per pre-registration). Each cell uses a 30-document bootstrap CI ($n_{\text{boot}} = 5{,}000$). We report *differentials* (specialist mean − size-matched random complement mean); ratios are unstable when comp ≈ 0. The Mamba-2 head-granularity headline cells additionally use a 30-seed random-complement bank under stratified non-adjacent sampling because M-2 head differentials are 2–3× smaller than the Mamba-1 channel effect (§4.2) and require tighter CIs; the larger Mamba-1 channel effect (+3.76 nats at 130M, well above complement noise) is stable at the main-grid 3 seeds. Because our claims rest on pre-registered headline cells (Tables 2, 3) and the qualitative replications in §4.2 and §4.4, we report individual 95% bootstrap CIs without family-wise multiple-comparison correction across the full 6,060-cell grid. The exact seed protocol, sentinel reproducibility check, and 102-cell Pile-10k cross-dataset replication (Gao et al., 2020) are described in Appendix L.

Intervention labels name hook surfaces, not post-nonlinearity identities. In Mamba-1, `gate_one` sets pre-SiLU $z = 1$ (effective gate $\mathrm{SiLU}(1)$); in Mamba-2 it sets the gate multiplier to one. In Mamba-2, `delta_zero` zeros the pre-softplus timestep input, leaving the post-softplus timestep set by the learned bias.

Roadmap. §4.2 establishes the channel-granularity baseline (probe specialty tracks causal mass); §4.3 shows correspondence breaks at Mamba-2 head granularity, with §4.3.4 (F1) decomposing the break into detection and execution layers; §4.4 confirms Pythia attention lacks the detection layer; §4.5 transfers the execution claim to retrieval; §4.6 (T3) tests whether substrate granularity alone explains the asymmetry.

4.2 BOS coupling at channel granularity (Mamba-1)

Ablating the bos-specialist channel set across all layers (`gate_zero` × bos_specialist × `L_all`) on Mamba-1 130M produces a differential of +3.76 nats on both wikitext-2 and Pile-10k against the size-matched random complement, agreeing to two significant figures ($\Delta < 0.01$ nats). The differential persists at the larger scale, reaching +2.13 on wikitext-2 and +1.64 on Pile-10k at Mamba-1 2.8B (Figure 1a), roughly half the 130M effect, consistent with channel specialty being concentrated in early-layer state initialisation that scales sub-linearly with width. Setting the gate parameter to one (`gate_one`) is a dose-response variant that tests whether the coupling runs through selective gating rather than through static channel activity. On Mamba-1 130M the `gate_one` differential reaches +26.62 on wikitext-2 and +31.53 on Pile-10k; at 2.8B it is +19.94 and +19.88 (within 0.06 nats cross-dataset). The `gate_one` intervention amplifies the `gate_zero` effect by an order of magnitude, confirming that the coupling runs through the SSM selectivity mechanism (Gu and Dao, 2024) rather than through static channel statistics. Layer-wise, 46% of the all-layer effect concentrates at L_0 at 130M (+1.81 vs +3.97 `L_all`) but only 13% at 2.8B (+0.25 vs +1.87); BOS handling localises sharply at the bottom of the smallest network and redistributes across layers with scale, a 3.5× drop in L_0 localisation. This is also consistent with redundancy; we do not test sufficiency. A tempting reading as "BOS handling broadens with scale" is a task-averaging artefact once the wikitext-newline target is excluded (mid-layer ablation is null on wikitext-bos at both scales: +0.054 / +0.046; Appendix F).

4.3 BOS coupling and dual-head dominance at head granularity (Mamba-2)

At head granularity, single-bucket bos-specialty remains positively causal but with substantially smaller magnitude than at channel granularity, and a separate, larger head set (the dual heads, 3–7× more numerous than the single-bucket set) carries proportionally more aggregate BOS-causal mass under `gate_one` ablation; single-bucket probes by construction exclude this set. We support this in three steps: (i) the single-bucket M-2 bos-specialist effect is positively causal but small in absolute magnitude across all 8 cells under a 30-seed random-complement bank (§4.3.1); (ii) the finding survives a 4× context-length robustness check (§4.3.2); (iii) dual heads carry 4.8–6.4× the single-bucket differential under `gate_one` ablation on both M-2 scales and both corpora, with the ratio carrying no seed-driven uncertainty since both sets are deterministic (§4.3.3).

4.3.1 Single-bucket effect: positive but small

Takeaway. All 8 cells of the 30-seed grid (2 scales × 2 interventions × 2 corpora) show positive single-bucket bos-specialist differentials with bootstrap CI strictly excluding zero (Table 2). Cross-dataset agreement is tight on `gate_one`, with $|\Delta| \leq 0.022$ nats on both M-2 scales, and looser on gate-

Model	Interv	spec	$\bar{x}^{c}_{30}$	**diff**	CI_{95}
1.3B	g_zero wt-2	1.60	0.48	**1.12**	[0.95,1.28]
	g_zero Pile	1.56	0.45	**1.11**	[0.99,1.22]
	g_one wt-2	2.40	1.80	**0.60**	[0.18,0.93]
	g_one Pile	2.42	1.80	**0.62**	[0.16,0.97]
2.7B	g_zero wt-2	2.06	0.22	**1.84**	[1.74,1.93]
	g_zero Pile	2.39	0.13	**2.26**	[2.18,2.32]
	g_one wt-2	3.14	0.66	**2.48**	[2.30,2.65]
	g_one Pile	2.94	0.46	**2.48**	[2.34,2.61]

Table 2: Mamba-2 single-bucket bos-specialist effect under 30 non-adjacent random complement seeds on wikitext-bos at `L_all`. Columns: spec = specialist mean ΔNLL; $\bar{x}^{c}_{30}$ = 30-seed random-complement mean; **diff** = spec $-$ $\bar{x}^{c}_{30}$. CI is 5,000-bootstrap on the seed mean.

		M-2 1.3B		M-2 2.7B	
set	target	wt-2	Pile	wt-2	Pile
bos-spec†	wt_bos	$+1.12$	$+1.11$	$+1.84$	$+2.26$
dual	wt_bos	-0.40	-0.14	$+1.14$	$+1.00$
bos-spec	wt_nl	$+3.11$	$+0.75$	$+7.42$	$+1.57$
dual	wt_nl	-1.08	-0.21	-5.79	$+0.53$

Table 3: F1 functional decomposition: ΔNLL differentials (spec mean $-$ size-matched 30-seed random complement mean) on wikitext-2 and Pile-10k at $\tau = 0.5$, `gate_zero`, `L_all`. Per-scope means and bootstrap CIs in Appendix I. †Replicated from §4.3.1 for direct comparison; comp baseline is the 30-seed bank from Table 2. 130M cells in Figure 2.

zero: 0.011 nats at 1.3B and 0.42 nats at 2.7B, where Pile-10k yields the larger magnitude ($+2.26$ vs. $+1.84$) but both CIs still exclude zero.

The 30-seed M-2 differentials are roughly 2–3$\times$ smaller than Mamba-1 2.8B at a comparable scale (cf. §4.2), reflecting the smaller absolute substrate available to a single-bucket label at head granularity.

4.3.2 Robustness at $4\times$ context length

At 2048 ctx, 7 of 8 cells reproduce the 512-ctx direction with bootstrap CI excluding zero (the one cell that weakens, M-2 1.3B / `gate_one` / Pile-10k, was already the widest-CI cell at 512 ctx). The dual-set aggregate effect (§4.3.3) is robust at both context lengths; full breakdown in Appendix D.

4.3.3 Dual-head aggregate dominance

At `gate_one` ablation on wikitext-bos at `L_all`, the dual set (heads active on exactly two classes, disjoint from single-bucket bos-specialists) carries 4.8–6.4$\times$ the single-bucket differential on both M-2 scales and both corpora: 4.84$\times$ and 4.77$\times$ at 1.3B (wt-2 and Pile-10k respectively), 6.41$\times$ and 5.70$\times$ at 2.7B. Both sets are deterministic, so the ratio carries no seed-driven variance. The aggregate ratio is set-size-driven and τ-sensitive (range 1.02–6.41$\times$ across $\tau \in \{0.3, 0.5, 0.7, 1.0\}$, per-unit $\sim 1\times$); `gate_zero` and wikitext-newline give different multipliers (Appendix A). The load-bearing claim is therefore the F1 functional decomposition in §4.3.4, not the specific multiplier.

4.3.4 Functional decomposition by ablation target (F1)

The 4.8–6.4$\times$ aggregate ratio is set-size driven and conflates two distinct causal roles. To decompose, we re-evaluate gate-zero on each head set against both target-token classes: tokens predicted under BOS-context (wikitext-bos) and tokens where the target is a newline (wikitext-newline). For each cell (head set, target, scale, corpus) we draw 30 size-matched random complement seeds and report the spec–comp differential with bootstrap 95% CI (Table 3; Appendix I).

Detection vs. execution. The dual-head BOS–newline cosine alignment (cos $= 0.89$; §5, Appendix E) is selection-driven: dual heads are defined by activity on exactly two classes, so elevated cross-class cosine is expected by construction. The principal evidence is therefore the *functional* ablation of Table 3: dual heads share boundary-class activation in Δ-space but do not execute the newline-target state reset, which is carried by the bos-specialist heads that also execute the BOS-context reset. Single-bucket bos-specialists are the load-bearing execution layer (necessary for both targets; we do not test sufficiency); dual heads are the representational detection layer that single-bucket classification excludes by definition. The 4.8–6.4$\times$ aggregate dominance (§4.3.3) is the detection layer's set-size signature; the per-unit $\sim 1\times$ ratio reflects no additional execution mass in the dual set. The dual $\times$ wt_bos differential is monotonic across the three M-2 scales, moving from $-2.89/-1.59$ at 130M through $-0.40/-0.14$ at 1.3B to $+1.14/+1.00$ at 2.7B (wt-2 / Pile; Figure 2). This is consistent with a scale-emergent stratification of the dual role: at small scale the dual set has not yet cleanly separated from the execution layer and partially interferes with BOS-context prediction. The negative dual $\times$ wt_nl differential at M-2 2.7B wt-2 (-5.79) is additionally consistent with the cross-task swap of Appendix F.

4.4 Cross-architecture baseline (Pythia attention)

To separate head granularity per se from architecture-specific dual allocation, we run the same specialist classification and ablation pipeline on four Pythia attention-head checkpoints. Pythia heads are classified as `bos_specialist` by mean attention-to-BOS strength above 0.10 from positions of each boundary class (Xiao et al. 2024; Gu et al. 2025); this threshold has different semantics from Mamba's pre-softplus Δ threshold. We zero out classified specialist heads across all layers and score first-eight-tokens-after-BOS NLL on wikitext-2 and Pile-10k, comparing against 30-seed size-matched non-specialist complements (same protocol as §4.3.1).

All four Pythia checkpoints show clean positive selectivity on both corpora, all eight CIs strictly excluding zero (cross-dataset agreement 0.05–0.30 nats per cell, same direction; Appendix B). Three architecture-granularity combinations (Mamba-1 channels, Mamba-2 heads, Pythia heads) therefore exhibit positive single-bucket selectivity in the same direction. Mamba-2 head granularity additionally exhibits dual-head dominance, which Mamba-1 channels do not. Applying the same multi-class rule (exactly 2 classes above threshold) to Pythia yields a sparse dual set covering only 2–12% of heads per model, against 27–35% on Mamba-2. Ablation effects sit at the noise level: $\leq$ 0.16 nats with the dual-to-bos-spec ratio between $-4.6\times$ and $0.6\times$ (Appendix B), in contrast to Mamba-2's multi-nat dual differentials. Pythia's attention-sink mechanism is therefore distributed across generalist heads rather than concentrated at a clean dual set. The Pythia-1.4B differential of $+2.22$ nats is comparable to Mamba-1 2.8B's $+2.13$ nats; the Pythia drop from $+2.22$ at 1.4B to $+1.58$ at 2.8B parallels the same saturating-with-scale pattern Mamba-1 channels show ($+3.76 \rightarrow +2.13$).

4.5 Long-context retrieval transfer (RULER NIAH)

We test whether the bos-specialist set transfers to a qualitatively different task by running RULER Needle-in-a-Haystack (Hsieh et al., 2024) at 1024 and 2048 context on the two flagship Mamba checkpoints. At 1024 ctx, both architectures exhibit clean three-way separation: baseline retrieval succeeds (1.00 accuracy at all three needle depths), bos-specialist ablation eliminates it (0.00 everywhere), and size-matched complement ablation preserves it (1.00). The complement is size-matched random. A stronger control replaces it with a magnitude-matched non-specialist set (per layer, the top-K non-bos-specialist heads ranked by mean $|\Delta_{\text{pre}}|$ across all tokens, with K matching the bos-specialist count). At Mamba-2 2.7B the magnitude-matched complement still preserves retrieval at both context lengths (mean accuracy 1.00 at 1024 ctx, 0.97 at 2048 ctx), confirming that BOS-specialty rather than raw activation magnitude drives the collapse; at Mamba-1 2.8B the magnitude-matched complement collapses retrieval at both context lengths (mean accuracy 0.00 at 1024 ctx, 0.00 at 2048 ctx), consistent with the broader channel-granularity causal substrate of §4.2 where multiple sufficient subsets exist. The per-token NLL gap on the answer span under bos-specialist ablation is $\sim$8.5 nats on M-1 2.8B and $\sim$ 9.7 nats on M-2 2.7B, substantially larger than the corresponding wikitext-bos differentials. At 2048 ctx the three-way separation holds on M-2 2.7B: baseline retains 1.00 at all depths, bos-specialist ablation collapses retrieval to 0.00 (mean NLL 9.51), and the size-matched complement preserves 1.00 (mean NLL 0.30). These longer-ctx runs require an H800 80 GB card; the 24-GB consumer-GPU budget that capped the 1024-ctx grid is documented in Appendix K. M-1 2.8B, which was trained at 2048, shows partial baseline degradation at depth 0.5 (accuracy 0.80; depths 0.15 and 0.85 remain 1.00) but bos-specialist ablation still drops retrieval to $\sim$0.0. The Mamba-2 2.7B dual ablation produces partial degradation at both context lengths (mean accuracy 0.40 at 1024 ctx, 0.27 at 2048 ctx; intermediate between baseline 1.00 and bos-specialist 0.00), converging with the wikitext-bos 4.8–6.4$\times$ dominance ratio (§4.3); the dual set carries less than full causal mass on its own on NIAH, indicating that bos-specialist and dual heads have partially overlapping but non-identical retrieval substrates. At 4096+ ctx, both checkpoints' baseline retrieval has collapsed (both pretrained at 2048, and baseline NIAH accuracy is already near zero at 4096, consistent with the receptive-field analyses of (Ye et al., 2025)); long-context intervention experiments therefore lack a meaningful baseline. Full per-condition accuracy / NLL breakdown at both ctx and the long-context compute constraint are in Appendix C and K.

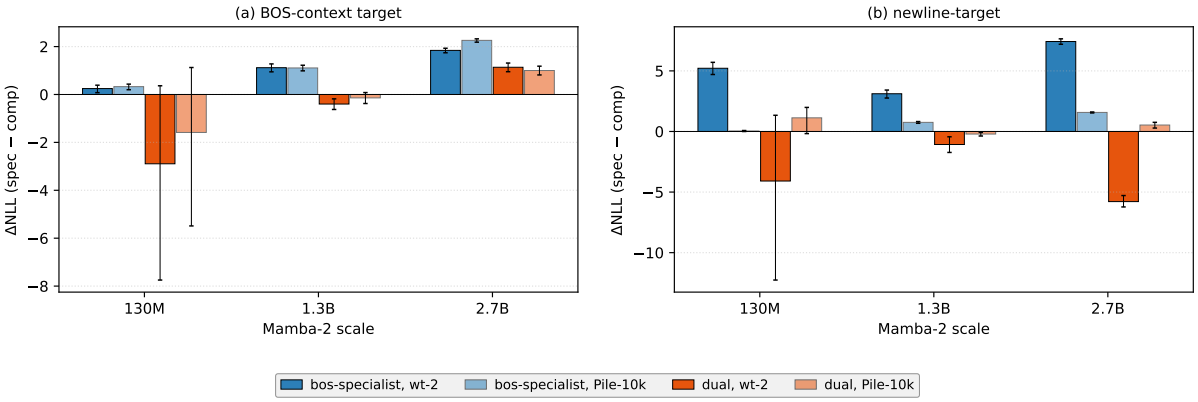

Figure 2: F1 functional decomposition across three M-2 scales: spec − size-matched random complement differentials (`gate_zero`, `L_all`, 30-seed bootstrap 95% CI). (a) BOS-context target. (b) Newline-target. Per-cell breakdown in §4.3.4 and Appendix I.

w	1	4	8	16	32	64	M-2 2.7B
H/C	1280	320	160	80	40	20	**20**
bos-spec	.294	.302	.299	.296	.293	.292	**.050**
dual	.296	.285	.279	.277	.278	.280	**.354**

Table 4: T3 bucket-width sweep on Mamba-1 2.8B ($K = 10$ random bucketings; mean fractions). Remaining $\sim 40\%$ of buckets are generalist or silent (full breakdown in Appendix J).

4.6 Substrate granularity alone does not drive the asymmetry

The 5% (Mamba-2 head) vs. 29% (Mamba-1 channel) bos-specialist gap (Table 1) admits a natural substrate-capacity hypothesis: that Mamba-2's small per-layer unit budget ($H/C = 20$ at 2.7B, where H is the per-layer unit count (heads in M-2, buckets in bucketed M-1 below) and $C = 4$ classes) forces units to multiplex across boundary classes, suppressing single-bucket specialty and elevating dual specialty. We test this directly by random channel bucketing of Mamba-1 2.8B: partition each layer's 5,120 channels into $\lfloor 5120/w \rfloor$ random groups of w channels, classify each bucket using bucket-mean $\Delta\bar{a}$ per class at the same $\tau = 0.5$, and report per-category fractions for $w \in \{1, 4, 8, 16, 32, 64\}$, the latter matching M-2 2.7B's $H/C = 20$ (Table 4; $K = 10$ random bucketings; Appendix J).

Random channel aggregation cannot recover M-2's distribution. This is consistent with a role for Mamba-2's head-shared Δ projection, though head dimension, state width, and training data remain unisolated. A Phase B ablation grid on bucketed Mamba-1 at $w = 8$ yields sub-nat differentials, an order of magnitude below M-2 head-level differentials (Appendix J).

5 Discussion

The detection / execution dichotomy. The Mamba-2 state sink decomposes into a representational detection layer (dual heads) and a functional execution layer (single-bucket bos-specialists) excluded from single-bucket classification; reading $\cos = 0.89$ on dual heads as functional equivalence inverts the causal structure (§4.3.4). Unlike attention-sink heads, whose value norms are suppressed (Xiao et al., 2024), BOS-specialists carry causal mass for both BOS targets and retrieval (§4.2, §4.5); the state sink is an active anchor, not a null no-op. The asymmetry is architectural: Mamba-2's head-shared Δ projection forces multiplexing across boundary classes; Mamba-1's channel-independent projections do not (§4.6), and Pythia at comparable H/C shows neither (§4.4).

Implications. Single-signal readings of attention sinks (Xiao et al., 2024; Gu et al., 2025) and Mamba interpretability (Chiang et al., 2025; Ye et al., 2025) suffice at fine granularity but are incomplete at Mamba-2 heads; prior single-bucket localisations on head-shared substrates (Sharma et al., 2024; Ali et al., 2025) warrant re-audit under multi-class aggregation. SDPA sink heads are *replaceable* (Qiu et al., 2025); the Mamba-2 state sink is *decomposed*.

6 Limitations

Datasets and models. The intervention grid covers four Mamba checkpoints (Mamba-1 130M /

2.8B, Mamba-2 1.3B / 2.7B) on wikitext-2 (Merity et al., 2017) and Pile-10k (Gao et al., 2020); the 30-seed cross-dataset confirmation covers all 8 M-2 head-granularity headline cells. Mamba-2 130M is added for the F1 scale ladder (§4.3.4) and the phenomenology cross-method correlations only. Three further wikitext-2-specific findings do not replicate on Pile-10k (Appendix F). Pythia (Biderman et al., 2023) 160M–2.8B is the cross-architecture baseline; hybrid SSM–attention models such as Zamba (Glorioso et al., 2024) are out of scope. Primary scoring is first-eight-tokens-after-boundary NLL on wikitext-bos / wikitext-newline plus RULER NIAH retrieval at 1024 and 2048 ctx; formal-task evaluations (parity, Dyck-2, S_3-permutation) are deferred. Untested classes, including punctuation and `whitespace_word`, may form additional execution layers.

Effect size. Mamba-2 single-bucket differentials are small ($+0.25$–$+2.48$ on wikitext-bos vs. $+3.76$ for M-1 130M); the M-2 1.3B `gate_one` CI is wide but excludes zero. At M-2 130M the dual set is small enough that its ablation CIs are wide (point estimates remain monotonic with scale; §4.3.4). T3 rules out substrate granularity alone, not head dimension, state width, or training-data alternatives.

Ethics Statement

Data and models. All experiments use publicly released language models (Mamba-1 (Gu and Dao, 2024), Mamba-2 (Dao and Gu, 2024), Pythia (Biderman et al., 2023)) under their respective public release terms, and publicly released text corpora (wikitext-2 (Merity et al., 2017), Pile-10k (Gao et al., 2020)). No human subjects, no personally identifying information, and no proprietary data are involved. We comply with the data release terms on attribution and redistribution.

Compute and environmental impact. The full intervention grid, 30-seed banks, and retrieval-transfer experiments together consume approximately 25–30 GPU-hours on a mix of consumer RTX 3090 and cloud H800 80 GB instances. Compute is modest relative to model-training runs; we minimise repeated forward passes by caching specialty classifications and re-using the same chunk preparation across cells.

Reproducibility. The grid is pre-registered with an amendment trail documenting design changes between v0 and v6 (Appendix H). Random complement protocols use stratified non-adjacent seeds (master seed 13, gap $\geq$ 61 across $[0, 10\,000)$), $5\,000$-sample bootstrap on the seed mean, and deterministic specialty classification. We release the full code base, per-cell raw ΔNLL values (for re-aggregation), and intermediate artefacts on acceptance.

Potential applications and risks. Interpretability of selective state-space language models has primarily safety- and transparency-positive applications: it enables informed head-level diagnosis, principled pruning, and audit tooling for production systems. Misuse to selectively damage downstream capability would require white-box access to the deployed model, which is the same threat surface as standard model deployment. We are not aware of risks beyond those already present in publishing model weights and standard interpretability analyses.

References

Ameen Ali, Itamar Zimerman, and Lior Wolf. 2025. The hidden attention of Mamba models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1516–1534.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*.

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

Nicola Cancedda. 2024. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4792–4808.

Hung-Yueh Chiang, Chi-Chih Chang, Natalia Frumkin, Kai-Chiang Wu, and Diana Marculescu. 2025. Quamba: A post-training quantization recipe for selective state space models. In *International Conference on Learning Representations (ICLR)*.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tri Dao and Albert Gu. 2024. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*.

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

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguist.*, 9:160–175.

Danielle Ensign and Adrià Garriga-Alonso. 2024. Investigating the indirect object identification circuit in Mamba. *arXiv preprint arXiv:2407.14008*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Paolo Glorioso, Quentin Anthony, Yury Tokpanov, James Whittington, Jonathan Pilault, Adam Ibrahim, and Beren Millidge. 2024. Zamba: A compact 7b SSM hybrid model. *arXiv preprint arXiv:2405.16712*.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations (ICLR)*. ArXiv:2410.10781.

John A. Hartigan and Pamela M. Hartigan. 1985. The dip test of unimodality. *The Annals of Statistics*, 13(1):70–84.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *Conference on Language Modeling (COLM)*. ArXiv:2404.06654.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. 2024. Repeat after me: Transformers are better than state space models at copying. In *International Conference on Machine Learning (ICML)*.

Greg Kamradt. 2023. Needle in a haystack — pressure testing LLMs. `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`. GitHub repository.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations (ICLR)*. ArXiv:2403.19647.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Gonçalo Paulo, Thomas Marshall, and Nora Belrose. 2025. Do transformer interpretability methods transfer to RNNs? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27565–27572.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2505.06708.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 3363–3377.

Arnab Sen Sharma, David Atkinson, and David Bau. 2024. Locating and editing factual associations in Mamba. In *Conference on Language Modeling (COLM)*. ArXiv:2404.03646.

Bernard W. Silverman. 1981. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(1):97–99.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. In *Conference on Language Modeling (COLM)*. ArXiv:2402.17762.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C Daniel Freeman, and 7 others. 2024. Scaling monosemanticity: Extracting interpretable features from Claude 3 sonnet. *Transformer Circuits Thread*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, Garvit Kulshreshtha, Vartika Singh, Jared Casper, Jan Kautz, Mohammad Shoeybi, and Bryan Catanzaro. 2024. An empirical study of mamba-based language models. *arXiv preprint arXiv:2406.07887*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*.

Model	$\rho_{L\times\Delta}$	Verdict
Mamba-1 130M	$+0.198$	FAIL
Mamba-1 370M	$+0.182$	FAIL
Mamba-1 1.4B	-0.023	WEAK
Mamba-1 2.8B	$+0.085$	FAIL
Mamba-2 130M	-0.282	WEAK
Mamba-2 1.3B	$\mathbf{-0.443}$	**PASS**
Mamba-2 2.7B	$\mathbf{-0.437}$	**PASS**

Table 5: LongMamba $\times$ Δ_{pre}-right-tail Spearman ρ averaged across all layers per model; pre-registered prediction $\rho < -0.3$ (§3.3).

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2025. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations (ICLR)*. ArXiv:2404.15574.

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

Zhifan Ye, Kejing Xia, Yonggan Fu, Xin Dong, Jihoon Hong, Xiangchi Yuan, Shizhe Diao, Jan Kautz, Pavlo Molchanov, and Yingyan Lin. 2025. LongMamba: Enhancing Mamba's long-context capabilities via training-free receptive field enlargement. In *International Conference on Learning Representations (ICLR)*.

Qingyu Yin, Xuzheng He, Xiang Zhuang, Yu Zhao, Jianhua Yao, Xiaoyu Shen, and Qiang Zhang. 2024. StableMask: Refining causal masking in decoder-only transformer. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 57033–57052.

A Phenomenology cross-method tables and Figure 2

The main body §3.3 reports the headline LongMamba $\times$ Δ_{pre}-right-tail Spearman correlation result; the per-model values, the parallel Quamba $\times$ A-log polarisation pair, and the cross-method agreement breakdown with probe-derived specialist sets are collected here. Two prior-work metrics (Chiang et al., 2025; Ye et al., 2025) and two analytic baselines (right-tail $\Pr(\Delta_{\text{pre}} > 0.7)$ and $-\log A$ from the stored $A_{\log}$ parameter) are evaluated per channel/head per layer for all 7 Mamba checkpoints.

Cross-method agreement with probe specialty labels. The pre-registered prediction was that ≥ 3 of 4 sink-mass metrics should agree with probe-derived bos-specialist sets at top-10% Jaccard > 0.4. Observed Jaccard is in $[0.07, 0.15]$ across all 7 models and all 4 lines, substantially above chance

Model	3-of-4 / total	rate
Mamba-1 130M	96 / 96	**100.0**%
Mamba-1 2.8B	253 / 256	**98.8**%
Mamba-2 1.3B	174 / 192	**90.6**%
Mamba-2 2.7B	223 / 256	**87.1**%

Table 6: Per-bucket pre-Δ bimodality rate under the 3-of-4 method conjunction (§3.3).

($\sim$ 5%) but well below the 0.4 prediction. Best agreement on Mamba-2: LongMamba and A-log polarisation (Jaccard $\sim$ 0.14–0.15); worst: Δ_{pre} right-tail (Jaccard $\sim$ 0.03–0.07). Prior sink-mass methods capture partially complementary signals rather than a single canonical sink-mass quantity.

Quamba $\times$ A-log polarisation systematic anti-correlation on Mamba-1. Quamba (activation outlier) and A-log polarisation (architectural decay) are consistently negatively correlated on all four Mamba-1 scales ($\rho \in [-0.25, -0.37]$) but uncorrelated or weakly positive on Mamba-2 ($\rho \in [-0.01, +0.19]$). At channel granularity, activation-based and spectral-based sink scores capture distinct aspects of channel behaviour; at head granularity, the distinction collapses.

Per-bucket bimodality 3-of-4 conjunction (Table 6).

Threshold sensitivity of the dual-dominance ratio. We re-ran the `gate_one` $\times$ wikitext-bos $\times$ `L_all` ablation at three alternative thresholds $\tau \in \{0.3, 0.7, 1.0\}$ (in addition to the paper default $\tau = 0.5$). At each τ, we re-classified (layer, head) units from cached per-bucket mean differentials, then ran the ablation on the new bos-specialist and dual sets. The aggregate ratio is non-monotonic in τ and peaks near the default $\tau = 0.5$ (Table 7); it ranges across 1.02×–6.41× depending on τ and cell. The $|D|/|S|$ set-size dominance is robust across all sampled τ ($\geq$ 2.16×), and the per-unit ratio is consistently $\leq$ 1.5×. The headline aggregate ratio 4.8–6.4× should therefore be read as the value at the pre-registered default threshold, not as a τ-invariant claim; the load-bearing mechanistic claim is the qualitative BOS–newline alignment documented in §E.

B Pythia per-scale differentials

Pythia heads classified as `bos_specialist` by attention-to-BOS sink strength under our fixed threshold ($>$ 0.10); ablation zeros the output of

			wt-2		Pile	
τ	$\lvert S\rvert$	$\lvert D\rvert$	spec	dual	spec	dual
M-2 1.3B (ratio under each (spec, dual) pair)						
0.3	142	698	+1.42	+6.04	+1.18	+7.53
			4.26×		6.36×	
0.5*	251	826	+2.13	+10.33	+2.42	+11.54
			4.84×		**4.77**×	
0.7	351	860	+2.55	+7.31	+3.31	+7.61
			2.87×		2.30×	
1.0	434	939	+4.29	+4.56	+5.30	+5.41
			1.06×		1.02×	
M-2 2.7B						
0.3	154	1457	+1.77	+6.60	+1.98	+6.81
			3.72×		3.43×	
0.5*	254	1810	+2.66	+17.07	+2.94	+16.74
			6.41×		**5.70**×	
0.7	370	1886	+3.56	+12.37	+3.76	+12.24
			3.48×		3.25×	
1.0	514	1843	+4.64	+10.87	+4.78	+11.39
			2.34×		2.38×	

Table 7: Threshold sensitivity of the dual / bos-specialist aggregate ratio under `gate_one` ablation at `L_all` on wikitext-bos. $\tau = 0.5^*$ is the pre-registered default; discussion of the τ-dependence and the set-size dominance $|D|/|S| \geq 2.16\times$ in §A.

all classified specialist heads across all layers; first-eight-tokens-after-BOS NLL on wikitext-bos and on Pile-10k, 30 stratified random size-matched non-specialist complement seeds per cell (same protocol as Table 2; master seed 13, min seed gap 61).

C NIAH retrieval per-condition breakdown

Per-condition retrieval accuracy and per-token NLL on RULER NIAH (Hsieh et al., 2024) at 1024 and 2048 context, averaged across three needle depths $\{0.15, 0.5, 0.85\}$.

D 2048-context robustness check for Mamba-2 head granularity

We re-run the 8 Mamba-2 head-granularity headline cells at 2048 ctx (4× the H1 grid's 512 ctx), under the same 30-seed random-complement protocol. Both corpora yield 30 chunks at 2048 ctx; on wikitext-2 this requires concatenating the first 500 documents rather than 300 (the 300-doc budget caps at 23–25 chunks at this context length).

E Dual-head BOS–newline alignment (Δ_{pre} test)

We test the boundary-reset hypothesis (§5) by comparing class-conditional cosine similarity on the Δ_{pre} activations of (a) Mamba-2 dual heads vs.

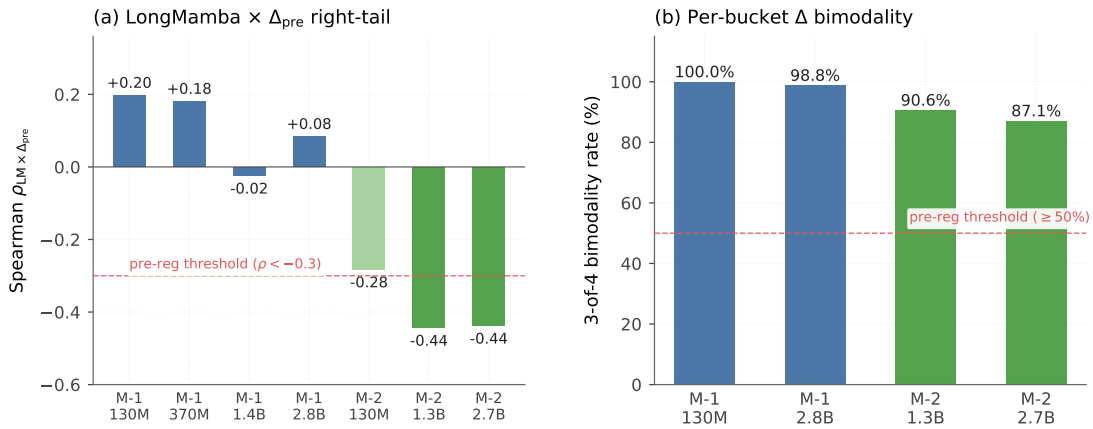

Figure 3: Cross-method phenomenology evidence. (a) LongMamba–Δ_{pre}-right-tail Spearman ρ averaged across layers. (b) Per-bucket pre-Δ distribution bimodality rate under the 3-of-4 method conjunction (Hartigan dip, Silverman peaks, GMM ΔBIC, KDE local modes). Findings discussed in §3.3.

	wt-2		Pile		cross	
Model	**diff**	CI_{95}	**diff**	CI_{95}	$\lvert\Delta\rvert$	dir
Pythia-160m	$+\mathbf{0.59}$	[+0.51,+0.68]	$+\mathbf{0.64}$	[+0.57,+0.71]	0.05	=
Pythia-410m	$+\mathbf{1.25}$	[+1.23,+1.28]	$+\mathbf{1.35}$	[+1.33,+1.37]	0.10	=
Pythia-1.4b	$+\mathbf{2.22}$	[+2.20,+2.25]	$+\mathbf{1.92}$	[+1.90,+1.95]	0.30	=
Pythia-2.8b	$+\mathbf{1.58}$	[+1.56,+1.60]	$+\mathbf{1.66}$	[+1.64,+1.68]	0.08	=

Table 8: Pythia cross-architecture baseline under the 30-seed stratified complement protocol on both corpora. The largest cross-dataset shift (Pythia-1.4b, 0.30 nats) is dominated by the deterministic specialist cell (+2.35 wt-2 vs. +2.00 Pile, $\Delta = 0.36$ nats) rather than by complement sampling (+0.13 vs. +0.07); the bank-mean variance from re-seeding the 30-seed complement is ~ 0.03 nats, well below the observed shift, so this reflects a genuine model–corpus interaction. Per-scale discussion in §4.4.

(b) Mamba-2 single-bucket bos-specialists, on 100 wikitext-2 documents at 512 ctx. For each (layer, head) in the head set, we compute the mean Δ_{pre} conditioned on the next-token class $\in$ {BOS, newline, punctuation, whitespace-word, other-word}; stack head-means into a class-vector and take pairwise cosine.

Size-matched control. The dual set is 3–7× larger than the bos-specialist set; averaging over more units could in principle smooth the class-mean estimate and inflate the cosine. We recompute $\cos(\Delta_{pre}^{BOS}, \Delta_{pre}^{newline})$ on $K = 200$ random subsamples of the dual set down to $|S|$ (matched to the bos-specialist set size). The subsampled mean is 0.9000 at M-2 1.3B (95% CI $[0.881, 0.919]$) and 0.8879 at M-2 2.7B (95% CI $[0.849, 0.910]$), essentially identical to the full-set values (0.9004, 0.8880) and decisively above the single-bucket cosines (0.53–0.54). The BOS-newline alignment is therefore a representational property of the dual head set and not a sampling smoothing artefact.

We additionally fit a linear probe on the per-position concatenated Δ_{pre} vector to predict the 4-way next-token boundary class (BOS, newline, punctuation, other-word; stratified 80/20 split with class balance via `class_weight="balanced"`, sklearn `LogisticRegression`). Both head sets achieve near-perfect overall accuracy (dual: 0.997 M-2 2.7B, 0.999 M-2 1.3B; bos-spec: 0.996 on both); the probe is too easy to discriminate between the two head sets (1810 / 826 dual features vs. 254 single-bucket features both saturate the linear separability of boundary class). The discriminating mechanism evidence comes from the class-conditional cosine in Table 11, not the probe accuracy.

F Wikitext-2-Specific Phenomena

Three findings from our intervention grid appeared on wikitext-2 but did not replicate on Pile-10k. We document them here as *informative observations*: they reveal that probe-derived specialty labels capture mechanism that is partially dataset-conditioned. Each is paired with a candidate mechanism hypothesis.

F.1 Newline sign-reversal scale-emergent

`gate_mean` × `newline_specialist` × `L_all` on `wikitext-newline`:

- wikitext-2: $+0.58/-0.05/-0.15/-0.58$ (4 models, monotonic crossover with endpoint magnitude convergence)

Model	ctx	condition	acc ↑	NLL ↓
M-1 2.8B	1024	baseline	**1.00**	0.10
	1024	bos_specialist	0.00	8.67
	1024	complement_bos	**1.00**	0.18
	1024	mag_matched_comp	0.00	9.44
	2048	baseline	0.93	0.48
	2048	bos_specialist	0.03	8.29
	2048	complement_bos	0.93	0.43
	2048	mag_matched_comp	0.00	10.06
M-2 2.7B	1024	baseline	**1.00**	0.03
	1024	bos_specialist	0.00	9.70
	1024	complement_bos	**1.00**	0.02
	1024	mag_matched_comp	**1.00**	0.54
	1024	dual	0.40	4.18
	2048	baseline	**1.00**	0.05
	2048	bos_specialist	0.00	9.51
	2048	complement_bos	**1.00**	0.30
	2048	mag_matched_comp	**0.97**	0.76
	2048	dual	0.27	5.59

Table 9: RULER NIAH retrieval accuracy and per-token NLL at 1024 and 2048 context, averaged across needle depths $\{0.15, 0.5, 0.85\}$. `mag_matched_comp` is the magnitude-matched non-specialist complement (§4.5). M-1 2.8B at 2048 ctx shows partial baseline degradation at depth 0.5 (0.80) but 1.00 at depths 0.15 and 0.85; bos-specialist ablation still drives retrieval to $\sim$0.0. M-2 2.7B at 2048 ctx reproduces the 1024-ctx three-way separation plus a slightly stronger dual degradation (0.27 vs. 0.40).

- Pile-10k: $+0.003/-0.002/+0.03/+0.04$ (collapse to noise floor; 3/4 endpoints sign-flip)

Mechanism hypothesis. Wikitext-2 article-boundary structure (consistent double-newline between articles, paragraph-level newline patterns) provides a stable signal that newline-specialist channels can specialise on, yielding a gate-mean ablation effect with predictable scale-emergent sign reversal. Pile-10k contains code, web text, and books with heterogeneous newline semantics (function delimiters, list markers, poetry line breaks); the absence of a single dominant newline distribution prevents the model from developing a clean newline circuit, so the gate-mean ablation has no consistent effect.

F.2 Cross-task swap on wikitext-2 M-2 2.7B

`gate_zero` × `bos_specialist` × `L_all` on M-2 2.7B:

- wikitext-2: BOS = $+1.73$, newline = $+7.32$ ($4.22\times$ ratio, bos-spec damages newline more than BOS)

Cell	512 ctx diff	2048 ctx diff	2048 CI_{95}
1.3B `g_zero` wt-2	$+1.12$	$\mathbf{+1.60}$	[+1.44,+1.74]
1.3B `g_zero` Pile	$+1.11$	$\mathbf{+1.23}$	[+1.07,+1.40]
1.3B `g_one` wt-2	$+0.60$	$\mathbf{+0.61}$	[+0.22,+0.92]
1.3B `g_one` Pile	$+0.62$	$+0.12$	[-0.32,+0.56] $*$
2.7B `g_zero` wt-2	$+1.84$	$\mathbf{+1.84}$	[+1.75,+1.93]
2.7B `g_zero` Pile	$+2.26$	$\mathbf{+1.98}$	[+1.90,+2.06]
2.7B `g_one` wt-2	$+2.48$	$\mathbf{+2.37}$	[+2.19,+2.55]
2.7B `g_one` Pile	$+2.48$	$\mathbf{+3.00}$	[+2.86,+3.13]

Table 10: Mamba-2 bos-specialist differential at 2048 ctx versus the 512-ctx baseline; cells in bold have bootstrap CI excluding zero. $*$M-2 1.3B / `gate_one` / Pile-10k loses significance at 2048 ctx; the 512-ctx CI on this cell was already the widest of all eight ($[+0.16, +0.97]$). `gate_zero` cells are uniformly robust to context length (effect localised to the gated timestep), whereas `gate_one` cells slightly attenuate at longer ctx (effect distributed across $\propto L$ positions in the chunked-scan accumulator).

Model	class pair	dual	bos_spec
M-2 1.3B	cos(BOS, newline)	$\mathbf{+0.90}$	$+0.53$
	cos(BOS, punct.)	$+0.70$	$+0.72$
	cos(newline, punct.)	$+0.73$	$+0.68$
	cos(BOS, other-word)	$+0.77$	$+0.76$
M-2 2.7B	cos(BOS, newline)	$\mathbf{+0.89}$	$+0.54$
	cos(BOS, punct.)	$+0.47$	$+0.45$
	cos(newline, punct.)	$+0.57$	$+0.85$
	cos(BOS, other-word)	$+0.66$	$+0.59$

Table 11: Class-conditional Δ_{pre} cosine similarity on Mamba-2 dual vs. bos-specialist head sets, 100 wt-2 docs at 512 ctx. The BOS–newline row is the discriminating signal; other class pairs are similar across head sets because Δ_{pre} values are predominantly positive in L_2 space. Full 5×5 matrix in `results/dual_head_mechanism/`.

- Pile-10k: BOS = $+2.39$, newline = $+1.64$ ($0.69\times$ ratio, normal direction restored)

Mechanism hypothesis. On wikitext-2, BOS and newline tokens systematically co-occur (BOS introduces an article, newlines structure paragraphs within); the bos-specialist head set in M-2 2.7B may have learned a co-distribution that ties BOS handling to newline prediction. On Pile-10k, BOS and newline are not as tightly co-distributed (BOS appears at the start of arbitrary chunks, newlines have varied roles), so the cross-task effect dissolves.

F.3 Within-task sign reversal at boundary layers, M-2 2.7B wt-2

`gate_zero` × `dual` × `L_boundary`:

- M-2 2.7B / wt-2: BOS $= +2.30$, newline $= -1.79$ (sign flip)
- M-2 2.7B / Pile: BOS $= +2.00$, newline $= +1.77$ (no flip)
- M-2 1.3B: never flips on either dataset.

Mechanism hypothesis. On wikitext-2 in M-2 2.7B specifically, the dual-head set at boundary layers may participate in an article-boundary mechanism where ablating dual hurts BOS prediction (causal for BOS) but helps newline prediction by removing a competing signal (anti-causal for newline). This is a model-and-dataset-specific mechanism, not an architectural property, with single-cell evidence within the M-2 family.

F.4 Mid-layer effect is wikitext-newline-only

`gate_zero` $\times$ `bos_specialist` $\times$ `L_mid_third` per-task split:

- M-1 130M / wikitext-bos: spec $+0.055$ vs comp $+0.001$, differential $+0.054$ (null, expected)
- M-1 2.8B / wikitext-bos: spec $+0.059$ vs comp $+0.013$, differential $+0.046$ (null on bos task)
- M-1 2.8B / wikitext-bos / Pile: spec -0.019, differential -0.022 (null)

Mechanism hypothesis. On the BOS-target metric, mid-layer ablation is essentially null at both 130M and 2.8B scales on both wikitext-2 and Pile-10k; the non-null mid-layer effect is driven entirely by the wikitext-newline target, consistent with the corpus-conditioned newline phenomena above. A reading as "BOS handling broadens with scale" would be a task-averaging artefact.

F.5 Why wikitext-2 specificity matters

These three phenomena are not paper failures: they suggest that *specialty labels derived from a particular corpus capture mechanism that is partially corpus-conditioned*. For the paper-main claims (BOS channel coupling, dual-head dominance, and `gate_one` selectivity), we report only cross-dataset replicated cells. For follow-up work, the wikitext-2-specific phenomena suggest investigating dataset-structure-dependent specialisation in selective state-space models as an independent research thread: when does specialty generalise across corpora, when does it not, and what corpus features predict generalisation?

Model	silver	ΔBIC	3-of-4	4-of-4
M-1 130M	100.0	96.9	100.0	96.9
M-1 2.8B	97.7	96.1	98.8	94.9
M-2 1.3B	85.9	58.3	90.6	53.6
M-2 2.7B	81.6	47.3	87.1	41.8

Table 12: Bimodality positive rates (%) per model. Hartigan dip ≥ 0.05 and KDE local-mode count ≥ 2 both saturate at 100% across all four models and are omitted from the table; Silverman peaks ≥ 2 and GMM ΔBIC ≥ 10 are the discriminating criteria, with the parametric ΔBIC the most restrictive on Mamba-2 (lowest rate 47.3% on M-2 2.7B). The strict 4-of-4 conjunction drops below the 50% pre-registration threshold on Mamba-2 2.7B (41.8%); the 3-of-4 conjunction (specified pre-registration to retain power against the unimodal null when one method is borderline) exceeds 87% on all four models.

G Bimodality detection per method and strict conjunction

For each (layer, bucket) cell of the four Mamba models tested in §3.3, we record per-method positive rates and both the pre-registered 3-of-4 conjunction and the strict 4-of-4 conjunction (Table 12).

H Pre-registration and Amendments

Pre-registration (anonymised OSF link in camera-ready) was filed prior to the intervention grid; framing amendments are timestamped in the project repository's `PRE_REGISTRATION_AMENDMENT.md`. All findings reported in the main body were pre-registered; the cross-architecture Pythia baseline and the RULER NIAH retrieval transfer were added as registered extensions on 2026-05-05.

I F1 Functional Decomposition Details

The F1 grid (§4.3.4) runs `gate_zero` $\times$ `L_all` on Mamba-2 1.3B and 2.7B against the cross product of head sets {bos-specialist, dual, complement_bos, complement_dual} and targets {wikitext-bos, wikitext-newline} on wikitext-2 and Pile-10k. Spec sets are deterministic; complement sets are size-matched with 30 non-adjacent stratified seeds (master 13). Each cell is bootstrapped with 5,000 replicates on the per-document ΔNLL distribution ($n_{\text{docs}} = 30$, max-len 512). Table 3 in the body reports headline differentials (spec mean minus seed-averaged complement mean) with 95% bootstrap CI; the full per-scope absolute ΔNLL means and per-seed bootstrap CIs are in the released

f1_dual_function_test artefact, one summary per (scale, corpus) cell. The dual × wt_newline differential at M-2 2.7B wt-2 (-5.79 nats) has a 95% CI of $[-6.25, -5.33]$, well-excluding zero in the negative direction; the corresponding M-2 1.3B wt-2 differential (-1.08) has CI $[-1.71, -0.46]$; M-2 1.3B Pile-10k differential (-0.21) has CI $[-0.38, -0.08]$.

J T3 Bucketing Sweep Details

T3 (§4.6) partitions each Mamba-1 2.8B layer's 5,120 channels into $\lfloor 5120/w \rfloor$ random groups of w channels and re-classifies each bucket using bucket-mean $\Delta\bar{a}$ per class at $\tau = 0.5$, for $w \in \{1, 4, 8, 16, 32, 64\}$. Phase A reports per-category fractions across $K = 10$ random bucketings (Table 4). Generalist and silent fractions (omitted from the body table for compactness) range 0.330–0.382 and 0.046–0.073 across bucket widths respectively. Phase B runs the F1-style ablation grid on the $w = 8$ bucketing (seed 42): bos-specialist-bucket and dual-bucket sets are deterministic; complement_bos_bucket and complement_dual_bucket are size-matched random sets drawn from the layer's non-(spec ∪ dual) channels with 30 seeds. Headline differentials are $+0.574$ / $+0.695$ at wt_bos and -0.428 / $+0.481$ at wt_newline (bos / dual respectively), all bootstrap CIs in the released t3_m1_bucketing/bw8 artefact. The $w = 64$ run (matching M-2 head granularity exactly) suffers within-layer bucketing saturation (random 64-channel buckets collapse to layer-mean statistics on Mamba-1, producing all-or-nothing per-layer classifications), so Phase B is run at the smaller $w = 8$ where within-layer heterogeneity is preserved.

K NIAH Long-Context Compute Constraint

The Mamba-2 intervention path uses a custom monkey-patched _patched_mamba2_torch_forward routine that materialises a $G_{\text{intermediate}}$ tensor of shape $[B, n_{\text{chunks}}, \text{chunk_size}, \text{chunk_size}, n_{\text{heads}}, d_{\text{state}}]$ during the SSD chunked scan (B_c, C_c are broadcast from n_{groups} to n_{heads} prior to the diagonal-block contraction). At context length 2048 for Mamba-2 2.7B ($n_{\text{heads}} = 80$, $d_{\text{state}} = 128$, chunk_size $= 256$), the total forward peak measured by torch.cuda.max_memory_allocated is 42.9 GB, which exceeds the 24 GB consumer GPU on which the original 1024-ctx NIAH grid was run but comfortably fits an H800 80 GB card. The 2048-ctx Mamba-2 2.7B intervention results in Table 9 were obtained on the latter hardware. At 4096 ctx the same allocation reaches ∼40 GB without the intervention overhead and exceeds 80 GB once the patched path is engaged; baseline retrieval at 4096 ctx already collapses to 0.03 accuracy (averaged over depths) so intervention experiments at 4096+ would lack a meaningful baseline to ablate even without the compute constraint. A chunk-streaming reimplementation of the intervention path is left to future work.

L Cross-Dataset Replication Protocol

Cross-dataset replication uses the NeelNanda/pile-10k subset of The Pile (Gao et al., 2020): 30 documents per cell, filtered to contain at least two newline tokens per chunk for parity with the wikitext-2 sampling. Seeds, bootstrap CI methodology ($n_{\text{boot}} = 5{,}000$), and per-token NLL scoring on first-eight-tokens-after-BOS are identical to the wikitext-2 grid. We replicated 102 cells targeting the headline mechanisms: 24 BOS-coupling cells (§4.2, §4.3), 12 gate_one selectivity cells (§4.3), 12 dual-head dominance cells, 6 cross-task swap cells, 16 layer-locus / mid-layer cells, and 32 endpoint cells from the newline mirror-symmetry analysis that ultimately demoted to wikitext-2-specific (Appendix F).