SELECTIVE BUT NOT BROAD: RETRIEVAL CAPACITY AND ACCESSIBILITY IN STATE-SPACE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

State-space models (SSMs) such as Mamba solve selective retrieval (associative recall, MQAR) yet fail broad retrieval (uniform-over-positions discrimination) that requires every input position to stay recoverable (Position Recovery, PR). We argue these are two *different memory problems*: selective retrieval is *content-addressed* (recall by key), broad retrieval is *position-addressed* (keep every position discriminable), with distinct capacity requirements and distinct accessibility. Is the broad deficit a capacity limit or an optimization artifact? We model retrieval as distributing a fixed per-position signal-to-noise budget across the sequence. A single linear SSM channel provably cannot solve filler recall, which forces channel specialization: broad retrieval needs retrieval rank scaling with the sequence length, selective only with the number of keys, and the required timescales must be *placed* to match the queried ages. The theory predicts the structure of selective-retrieval collapse and the specialization signature of successful backbones; both appear in trained weights, including multi-scale RetNet tracking the theoretical $1/m$ line. A symmetric causal intervention then separates the two explanations: a curriculum unlocks broad retrieval on a standard SSM and transfers to selective, and unlocks selective at its own lower floor where standard SGD fails. At every scale tested, both deficits are accessibility-limited; the barrier persists through 122M parameters. A controlled Pareto frontier shows the objective is the lever: pure language modeling leaves broad at the random floor, and the PPL cost of adding broad is architecture-dependent. The theory holds for the linearized state map; every causal and theoretical claim is validated on this project's own checkpoints, and the frozen-probe absence of broad information replicates on 2.7B open pretrained checkpoints of both families.

1 INTRODUCTION

An empirical asymmetry motivates this work. On two retrieval tasks at matched scale ($d_{\text{model}} = 128$, 2 layers), the Mamba SSM (Gu & Dao, 2024) reaches MQAR accuracy 0.43 (selective content lookup, (Arora et al., 2024b)) but only 0.014 on Position Recovery (broad retrieval; random is $1/64 = 0.016$). The same model thus *retrieves selectively but not broadly*. Why? Two confounded explanations: (i) Mamba's *function class* cannot represent broad retrieval (a capacity claim), or (ii) SGD from random init does not *find* the broad-retrieval solution (an optimization claim). Throughout, *capacity* asks whether a retrieval solution exists in the function class; *accessibility* whether standard training reliably reaches it.

We resolve this with a capacity theory of retrieval and validate it on the model's own weights. The axis we isolate has been conflated with recall in prior work: MQAR is *content-addressed* (recall by key), Position Recovery *position-addressed* (every position must stay discriminable) — two memory problems with different floors, mechanisms, and accessibility; the second is the one recurrent-state compression bites first. The central result is a *symmetric causal test*: under the capacity hypothesis no training procedure reaches a task absent from the function class; a curriculum does, on both axes. A T-curriculum unlocks broad retrieval on a standard SSM and transfers to selective (MQAR ~ 0.80–0.92 vs. ~ 0.1, replicated on Mamba-3 through 122M); a K-curriculum unlocks selective at its $K\theta$ floor. At every scale tested, both deficits are accessibility-limited; the asymmetry

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.

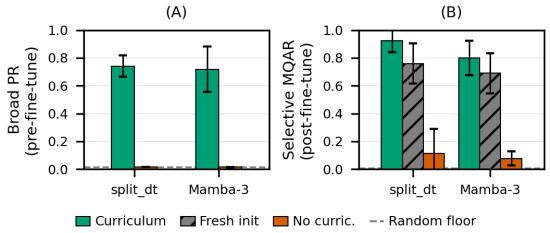

Figure 1: The symmetric causal intervention, the paper's central result (numbers and paired p-values in Table 3, §6). (A) A T-curriculum unlocks broad Position Recovery on standard Mamba-3 and the `split_dt` probe (0.72–0.74); no-curriculum training stays at the random floor (0.015). (B) Broad backbones transfer to selective MQAR (0.80–0.92); no-curriculum backbones stay near random (0.08–0.11); fresh inits land between (0.69–0.76): broad failure is an active trap. A curriculum reaching broad is impossible under the capacity hypothesis, so function-class impossibility is ruled out at this configuration. Bars: mean ± SD over 6 seeds.

is the magnitude gap $K\theta \ll T\theta$. The theory behind this: a per-channel impossibility ($\mathrm{SNR} \leq 1/m$, Theorem 1) forces specialization into the asymmetric floors, and a symmetry obstruction with a curriculum-invariance corollary locate where any cure must act (Propositions 1–2, Corollary 1). Its signatures are visible in trained weights (§5b–§5c). The paper in one sentence: *capacity sets the floors ($K\theta \ll T\theta$); accessibility binds; the objective is the lever.* **Contributions.** *(i)* A two-axis decomposition: broad (position-addressed; PR) and selective (content-addressed; MQAR) retrieval are different memory problems with distinct rank floors $T\theta$ vs. $K\theta$, from a fixed per-position SNR budget (Theorem 1, Propositions 1–2). *(ii)* A symmetric causal test: curricula unlock both axes where standard SGD fails — both deficits are accessibility-limited through 122M (§6–§7). *(iii)* Signatures in trained weights (§5), and external validation: the frozen-probe absence of broad information replicates on 2.7B open checkpoints of both families (Appendix J). Figure 1 shows the causal result this analysis predicts.

2 Background and Related Work

SSMs. Mamba (Gu & Dao, 2024), Mamba-2 (Dao & Gu, 2024), Mamba-3 (Lahoti et al., 2026), RWKV (Peng et al., 2023), RetNet (Sun et al., 2023), Hyena (Poli et al., 2023), and hybrid architectures (Lieber et al., 2024) compress context into a fixed-size state — the model family under study.

Retrieval and recall limitations of fixed-state models. Arora et al. (2024a) introduce MQAR and the recall–throughput tradeoff as a single axis; Arora et al. (2024b) navigate that tradeoff with linear attention. Jelassi et al. (2024) prove a copying gap (SSMs need state growing with T to copy; transformers realize this through attention), and Merrill et al. (2024) show Mamba fails length-controlled retrieval and state tracking. These works locate *which* tasks separate SSMs from transformers. We contribute two splits. First, capacity vs. accessibility: at our scale both axes are representable, and SGD under-reaches them (§6). Second, broad vs. selective: PR is position-addressed (broad) and MQAR key-to-value (selective), with distinct floors $T\theta$ vs. $K\theta$ (§3). Our conclusions are orthogonal to the copy lower bound: there the state itself is too small; here it is representationally sufficient and the failure is that standard training does not find the allocation, a regime lower bounds cannot see (§6 rules it out causally). Copying also differs structurally: it outputs *all* of the window, whereas broad retrieval must keep *any queried* position discriminable. Wen et al. (2025) locate the gap in recurrent-state retrieval limitations, Okpekpe & Orvieto (2025) in optimization (narrow learning-rate ranges); our causal intervention isolates accessibility on the broad axis.

Memory capacity of attention and fast weights. Bietti et al. (2023) analyze how transformers form associative memories over key–value pairs in-context; the delta-rule family (Yang et al., 2025) and Titans (Behrouz et al., 2025) build explicit fixed-size associative memories with learned write/erase dynamics. These target *content-addressed* storage (write associations, recall by key); our budget applies to the complementary *position-addressed* axis — keeping every position dis-

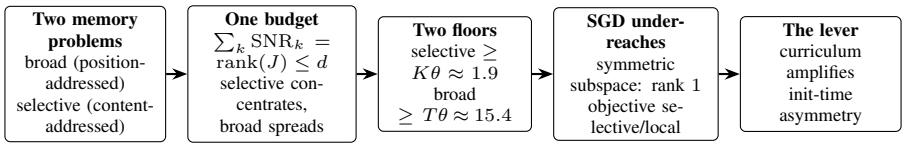

Figure 2: The paper in one picture. One budget, two floors; symmetric subspace and selective objective block standard SGD; a curriculum (which can only amplify init-time asymmetry, Corollary 1) unlocks both axes.

criminable in the pre-query state — where the $1/m$ ceiling and timescale covering bind, and where write/erase gates do not automatically help. Our theory also explains why that architecture line moves in the right direction: decay-style gating cannot escape the $1/m$ ceiling (exceeding it requires gating beyond decay, e.g. content- or position-dependent write gates; Theorem 1), no permutation-invariant training creates the required channel asymmetry on its own (Corollary 1), and learned, input-conditioned write dynamics (the delta rule) build it in by construction.

Expressivity vs. accessibility. Nishikawa & Suzuki (2024) prove SSMs (with nonlinear layers) match transformers on functions with input-dependent smoothness, so static expressivity largely does not separate the families. Our results are consistent with this and sharpen its reading: the broad-retrieval deficit we study is *not* an expressivity failure (curriculum reaches it in the same function class, §6) but an accessibility one; expressivity results alone cannot predict it.

SSM capacity and training. Li et al. (2021); Wang & Li (2024) analyze memory/optimization tradeoffs in recurrent models; our budget constraint is the classical leverage-score identity (§3).

3 A THEORY OF RETRIEVAL CAPACITY

We predict *which* retrieval tasks a recurrent-state model can solve and *why* broad and selective retrieval are asymmetric in both capacity and accessibility. The section runs in five moves — budget, impossibility, floors, obstruction, predictions; Figure 2 sketches the story. Statements here; proofs in Appendix B.

The leverage budget (move i). A recurrent state $s_T \in \mathbb{R}^d$ distributes a fixed per-position *leverage budget* $\sum_k \mathrm{SNR}_k = \mathrm{rank}(J) \le d$ across positions, by the leverage-score identity (Hoaglin & Welsch, 1978; Drineas et al., 2011), exact for any differentiable state map (Proposition 3):

$$\sum_{k=1}^{T} r_k^\top \Sigma^+ r_k = \mathrm{rank}(J) \le d, \qquad r_k := \partial s_T / \partial x_k, \quad \mathrm{SNR}_k := r_k^\top \Sigma^+ r_k, \tag{1}$$

with $J = [r_1, \dots, r_T]$, $\Sigma = JJ^\top$. We read SNR_k as a per-position *leverage-SNR proxy* (the identity is exact; the SNR reading adds a linearized noise model, Appendix B.1). **Retrieval criterion (modeling assumption).** A position k is reliably recoverable from s_T when $\mathrm{SNR}_k \ge \theta$, a per-position detection threshold ($\theta = O(1)$; value fit from behavior in §5b). In the full-row-rank regime $\mathrm{rank}(J) = d$, $p_k = \mathrm{SNR}_k / d$ is a distribution over positions: selective recall needs it *concentrated* ($H \approx \log K$), broad discrimination *uniform* ($H \approx \log T$); $T > K$, so broad demands higher allocation entropy, the structural seed of the asymmetry. (The identity holds for any differentiable state map; retrieval factors through the pre-query state $x \to S_{\mathrm{pre}} \to y$, giving a memory and an accessible rank (below). A Transformer's external key–value cache grows with T, so its memory is not d-bounded.)

A per-channel impossibility forces specialization (move ii). One channel with fixed decay α processing $[\textsc{key}, f_1, \dots, f_m, \textsc{query}]$ has key-vs-filler $\mathrm{SNR}(\alpha, m) = \alpha^{2m}(1-\alpha^2)/(1-\alpha^{2m})$ (a per-channel object, not the per-position SNR_k of move i).

Theorem 1 (Per-channel impossibility)**.** *For all $\alpha \in (0,1)$, $m \ge 1$: $\mathrm{SNR}(\alpha, m) \le 1/m$, with equality only as $\alpha \to 1$.*

For $m \ge 2$ no single decay lets the key beat accumulated filler noise, so a linear channel provably cannot solve filler recall: the model must specialize. Input-dependent *decay* does not escape this

(Theorem 2(b): arbitrary $\{A_t\}$ still gives $\max \mathrm{SNR} = 1/m$); exceeding the bound requires gating beyond decay, e.g. a content- or position-dependent *write* gate — the specialization selective retrieval uses.

Two floors (move iii). Summing the budget over the positions a task must discriminate gives the floor.

Proposition 1 (Capacity asymmetry, under the retrieval criterion)**.** *Under the retrieval criterion ($\mathrm{SNR}_k \geq \theta$), broad retrieval needs $\mathrm{rank}(J) \geq T\theta$; selective needs $\mathrm{rank}(J) \geq K\theta$. For $T \gg K$ broad is the harder axis.*

At our scale the necessary floors are broad $\gtrsim T\theta \approx 15.4$ and selective $\gtrsim K\theta \approx 1.9$ (evaluated at the cliff edge $\theta \approx 0.24$; $\theta \in (0.20, 0.25)$ bracketed from the §5b cliff, 8 seeds) — an $8\times$ gap. These are necessary, not sufficient: the model must also distribute capacity uniformly.

Specialization pricing. Reading the budget as a rank also prices *specialization*: coincident-timescale channels add nothing (identical decays give rank 1, Proposition 4), and the required timescales must moreover be *placed* to match ages: a bank resolves ages near k only with a pole pair of timescale $\approx k$ and relative separation $\approx c^*/k$, with adjacent-age resolution decaying as $\Theta(1/k)$ even when optimally placed (Lemmas 2, 1, Appendix B.2).

What the floors are not. The floors constrain $\mathrm{rank}(J)$, not a hyperparameter, and the two architectural ranks of the retrieval map differ in status. Retrieval factors through the pre-query state, $M = (\partial y/\partial S_{\mathrm{pre}})(\partial S_{\mathrm{pre}}/\partial x)$, so $\mathrm{rank}(M) \leq \min(r_{\mathrm{mem}}, r_{\mathrm{acc}})$: the readout side $r_{\mathrm{acc}} \leq d_{\mathrm{model}}$ is a formal bound (a linear head onto $h_{\mathrm{query}} \in \mathbb{R}^{d_{\mathrm{model}}}$; tested by the readout-rank ablation, §5d), while the memory side is empirical: d_{state} acts as a soft *covering* factor (broad reaches 0.19 at $d_{\mathrm{state}} = 2 \ll T\theta$, Appendix H), so $d_{\mathrm{state}} \geq T\theta$ is not a feasibility floor. Empirically neither rank alone sets the horizon; the horizon factorial, the realized coverage window, and the one clean d_{state} effect (from-scratch bootstrapping) are detailed in Appendix I.

Why SGD under-reaches broad (move iv). Two structural facts. First, a *symmetry obstruction* (Proposition 2 below): broad is unrepresentable on the permutation-invariant symmetric subspace, where $\mathrm{rank}(J) = 1$.

Proposition 2 (Symmetric obstruction)**.** *The retrieval loss is permutation-invariant; the symmetric subspace is gradient-flow-invariant, and along it $\mathrm{rank}(J) = 1$, so broad retrieval is unrepresentable there. Gradient flow from a symmetric init never reaches broad.*

This extends from a single loss to *any curriculum*: no sequence of permutation-invariant objectives can leave the symmetric subspace from a symmetric init (Corollary 1, Appendix B.5): a curriculum can help only by amplifying *pre-existing* asymmetry, never by creating it. Real Mamba's init is only approximately symmetric and the subspace need not attract nearby trajectories, so this does not say curricula cannot work (§6 shows they do); it says *where the lever must be*, consistent with the rising channel spread under curriculum (Figure 5). Second, from *approximately* symmetric inits, three established facts carry the under-reach without any unproven mechanism: the *objective* supplies no pressure (§6.1's Pareto frontier shows PPL alone leaves broad at the random floor); failed broad training is an *active trap* (fresh inits beat no-curriculum backbones, §6); and from-scratch training at full length hits a *bootstrap barrier* ($T_{\mathrm{boot}} \propto d_{\mathrm{state}}$, Appendix I). A parameterization-level candidate (an implicit low-rank bias of the optimizer) was tested three ways and returned no support; we retired it (Appendix B.9; the open-mechanism status is in Scope below).

The objective is the allocation lever. On the natural-text distribution we study, next-token prediction is empirically selective/local ($K_{\mathrm{eff}} \approx 4$–$8 \ll T$, Appendix K), so PPL training leaves the model selective and broad unexercised; a broad objective (PR) forces the high-rank allocation broad requires (§6); the SSM-specific difficulty of learning PR is rooted in sequential decay-chain coupling (Appendix B.10).

Predictions (move v). Four consistency checks confirm the theory recovers the motivating asymmetry (P1–P3, §5a: the architecture table's asymmetry, the flat d_{state} trajectory, RetNet's escape; P4, the d_{state}-only bootstrap limit, Appendix I); five theory-derived predictions carry the paper's evidence:

- **P5** (§5c): successful backbones specialize channels (multi-timescale for broad, input-dependent gating for selective); failures stay uniform.
- **P6** (§5b): non-selective channels collapse when $1/(K-1) < \theta$ (the cliff's $1/m$ structure).
- **P7** (§6): a T-curriculum unlocks broad; broad $\rightarrow$ selective transfers.
- **P8** (§7): the asymmetry persists across scale.
- **P9** (§5d, §6): readout rank $< T\theta$ abolishes broad; K-curriculum unlocks selective at $\text{rank} \approx K\theta$.

Scope. Theorem 1 and Propositions 1–2 hold for the linearized state map (proofs, Appendix B); the causal account of the under-reach (§6) is likewise free of unproven ingredients, with the microscopic mechanism left open (Appendix B.9). θ is a behavioral bracket from the K-sweep cliff, $\theta \in (0.20, 0.25)$ (midpoint 0.225, §5b; floors quoted at the cliff-edge value ≈ 0.24); the theory predicts the cliff's $1/m$ structure and the broad floor's linear-in-T scale ($T\theta$), not θ's value. The selectivity of next-token prediction ($K_{\text{eff}} \ll T$) is established for the natural-text distribution studied here, not claimed universal (Appendix K).

4 Benchmarks and setup

Position Recovery (broad). Input $[x_1, \ldots, x_T, \text{QUERY}, x_q]$; predict the position k with $x_k = x_q$. Tokens random and unique (no semantics). The query is revealed only after the window and any of the T positions is equally likely to be queried, so the pre-query state must keep *every* position addressable; we call this *broad* (pre-query coverage), in contrast to selective lookup where retaining one feature suffices (a full glossary: Table 5, Appendix A). $T = 64$.

MQAR (selective). Input $[k_1 v_1, \ldots, k_K v_K, q, \ldots]$ (Arora et al., 2024b); predict the value paired with query key q. $K = 8$.

Language modeling. Order-2 Markov chain (PPL floor 3.6) and char-level wikitext-103 (Merity et al., 2017) for the dissociation (§7). Default configuration (unless noted): $d_{\text{model}} = 128$, $n_{\text{layer}} = 2$, $d_{\text{state}} = 16$ (causal and K-sweep runs: 32), batch 32, AdamW lr 10^{-3}, 10k steps. Architectures: Transformer, Mamba/Mamba-2/Mamba-3, RWKV-6, RetNet (Sun et al., 2023), Linear Attention (Katharopoulos et al., 2020), Hybrid (SSM+attention), LinOSS-IM (Rusch & Rus, 2025), and `split_dt`, a Δt-decoupling variant used only as a diagnostic probe in Section 6.

5 Empirical validation

We validate in two tiers: consistency checks (P1–P4) that the theory recovers the motivating asymmetry, then the five theory-derived predictions (P5–P9) that carry the evidential weight: floors leave signatures in trained weights, and accessibility binds.

5.1 The capacity landscape (P1–P3): consistency with the asymmetry

Table 1 recovers the motivating asymmetry (P1–P3): Mamba is selective-capable (MQAR 0.43) yet broad-failed (PR 0.014, random); Transformer solves both; RetNet reaches broad in its high-PR regime. The d_{state} trajectory (Figure 7, Appendix H) is *horizontal* for Mamba (MQAR grows $0.17 \rightarrow 0.52$ as $d = 8 \rightarrow 64$, PR stays random) and *diagonal* for Hybrid (both grow); the deficit is not Mamba-specific (LinOSS-IM/IMEX (Rusch & Rus, 2025) fail PR too, despite oscillatory kernels' *stronger* covering capacity — an optimization-side failure, not capacity-side, Appendix B.6).

5.2 The MQAR K-sweep cliff (P6)

This is the closest available test of Theorem 1's $1/m$ structure: recall must overcome $m = K-1$ fillers, and a non-selective channel collapses when $1/(K-1)$ drops below θ. Sweeping K (Table 2; dense T-controlled grid below; Appendix C) shows the split Theorem 1 implies: RWKV collapses at $K = 8$, selective Mamba/split_dt degrade in order of their selectivity, and RetNet degrades slowest,

Table 1: Architecture comparison ($d_{\text{state}} = 16$, $T = 64$). Fixed-state SSMs (Mamba/Mamba-3, Linear Attn, RWKV, LSTM/GRU) fail broad PR; Transformer (attention state $\sim T$) and multi-scale RetNet escape. LinOSS-IM/IMEX ($T = 256$) also fail. †RetNet PR is bimodal ($1/5$ seeds reach broad); we list the maximum.

Architecture	PR (broad)	MQAR (selective)	Markov PPL
Transformer	**0.990**	0.135	7.73
Hybrid (SSM+attn)	0.456	0.416	8.11
RetNet†	0.978	0.138	44.9
RWKV-6	0.189	0.060	9.88
LSTM / GRU	0.047 / 0.045	0.138 / 0.127	10.7 / 13.7
Linear Attn	0.017	0.10	127
Mamba	0.014	**0.427**	34.4 ± 19.2
Mamba-3	0.014	0.310	8.2

at the same K for three d_{state} values. Reading the cliff as $1/(K-1) = \theta$ brackets θ: a *fit* supporting the $1/m$ structure, not its value.

Table 2: MQAR vs. K ($d_{\text{state}} = 32$, 4-seed mean). Random ≈ 0.008.

model	$K=4$	$K=8$	$K=16$	$K=32$
mamba (selective)	0.972	0.442	0.012	0.011
split_dt (selective)	0.989	0.670	0.013	0.013
rwkv (non-selective)	0.995	**0.057**	0.013	0.013
retnet (multi-scale)	0.266	0.138	0.069	0.017

The control variable is m, not T. A denser, *T-controlled* sweep ($K \in \{4, \ldots, 16\}$ at fixed context $T \in \{64, 128, 256\}$, independent of K; 3 seeds; Figure 4, Table 8, Appendix C) sharpens and generalizes the cliff. The curves are largely controlled by $1/(K-1)$ rather than T: RWKV's cliff is point-for-point identical at $T = 64$ and 128 (at $T = 256$ it shifts one K earlier, consistent with the bootstrap limit of Appendix I), and RetNet is identical to two decimals across all three T: the cliff's control variable is chiefly $m = K-1$, not context length. Two further readings (details in Appendix C). First, the cliff is a *per-run stochastic phase transition*: with 8 seeds at the edge, RWKV runs sit at ≥ 0.9 or ≤ 0.13 with nothing between (per-seed values in Appendix C), so the mean curve is a solve probability and the effective θ brackets to $(0.20, 0.25)$. Second, multi-scale RetNet's behavioral accuracy numerically tracks the theoretical $1/m$ SNR ceiling: its accuracy ratio to $1/(K-1)$ is $0.80 \to 1.03$ across $K = 4$–16 at all three T, a striking correspondence (a behavioral match to the Theorem-1 line, not a verification of it; Appendix C), and a level the selective models must exceed.

5.3 The specialization mechanism (P5)

Theorem 1 says failure to specialize is fatal; we test it on the trained checkpoints (per-channel $|A|$, dt, $\alpha_0 = e^{-dt|A|}$; state-dict only). Successful backbones have large *spread* across channels, failed ones stay uniform (Table 9, Appendix D). The mechanism differs by axis: broad success correlates with multi-timescale channels ($|A|$ and dt both differentiate), selective success with input-dependent gating (dt), consistent with Theorem 1. The signature replicates at scale (rank correlation of $|A|$-spread with PR: Spearman $\rho = +0.80$ over 8 24M checkpoints, $+0.41$ over 42 toy-scale ones); a curriculum-trained but *failed* backbone differentiates dt yet lacks timescale spread, landing with the failures. The broad signature is multi-timescale, not gating-only.

5.4 The readout-rank floor (P9)

Retrieval operates through $M = W^\top J$, so under the retrieval criterion a readout pinned below $T\theta$ cannot support broad. Ablating $\text{rank}(W)$ on Mamba-3 (Appendix E): PR fails at ranks $\{2, 8\}$ and

becomes reachable at $16 \approx T\theta$ ($4/5$ seeds), the sharp feasibility test on the readout rank, contrasting with the soft d_{state} curve (Appendix H). The selective floor is tested causally in §6.

6 ACCESSIBILITY: A SYMMETRIC CAUSAL TEST (P7, P9)

The theory says retrieval is representable above its floor (Proposition 1) but that SGD under-reaches it (the objective-and-landscape account of §3) — the spine's second clause, *accessibility binds*, now under causal test. Two hypotheses compete: ($\mathcal{H}_{\text{cap}}$) the function class cannot represent the task, so no training procedure reaches it; or ($\mathcal{H}_{\text{acc}}$) it is representable but SGD-inaccessible; under $\mathcal{H}_{\text{cap}}$ a curriculum *cannot* unlock the task. We test both axes and obtain a symmetric answer.

Broad: a T-curriculum unlocks broad retrieval (P7). Corollary 1 locates the only admissible mechanism class within permutation-invariant training: amplifying init-time channel asymmetry, with rising spread observed (Figure 5). We train Position Recovery under a 3-stage T-curriculum ($T = 16 \to 32 \to 64$) on our `split_dt` decoupling probe and on standard Mamba-3 (Lahoti et al., 2026), then fine-tune on MQAR; the no-curriculum control receives the same total training steps at $T = 64$ (so the comparison isolates staging, not optimization time) (Table 3). (`split_dt` widens the reachable broad set so the intervention is not confounded by its narrowness; the $1/m$ ceiling (Theorem 2) holds identically for coupled and decoupled recurrence — accessibility, not representability. Mamba-3 is the substantive replication.) Curriculum backbones reach broad PR (0.72–0.74) and

Table 3: Broad causal intervention, 6 seeds each, $d_{\text{state}} = 32$. Curriculum backbones reach broad PR and transfer to selective MQAR; no-curriculum (failed broad) stays near random; fresh $>$ no-curriculum (active trap). Replicates on standard Mamba-3.

architecture	backbone	PR (pre)	MQAR (post, mean±SD)	paired p
`split_dt` (probe)	curriculum (broad)	0.74	$\mathbf{0.92}\pm0.08$	—
	no-curriculum (failed)	0.015	0.11 ± 0.18	1.5×10^{-4}
	fresh (no pretrain)	—	0.76 ± 0.15	3.5×10^{-3} vs no-curr
Mamba-3 (standard)	curriculum (broad)	0.72	$\mathbf{0.80}\pm0.12$	—
	no-curriculum (failed)	0.015	0.08 ± 0.05	1.0×10^{-5}
	fresh (no pretrain)	—	0.69 ± 0.14	2.0×10^{-4} vs no-curr

transfer to selective MQAR (0.80–0.92), while no-curriculum backbones stay near random (paired $p \leq 10^{-4}$; Table 3). A failed backbone is an active *trap*: fresh inits beat no-curriculum (consistent with a worse basin), and pre-fine-tune PR predicts post-fine-tune MQAR (Pearson $r = +0.93$ over 36 backbones, Figure 5(B); descriptive). *Verdict:* a curriculum reaches broad on a standard architecture, impossible under $\mathcal{H}_{\text{cap}}$; the broad deficit is one of accessibility.

Selective: a K-curriculum unlocks the $K\theta$ floor (P9). The selective side gives the symmetric result. Under standard SGD, selective MQAR becomes learnable only near $d_{\text{state}} \sim 16$, far above the $K\theta \approx 1.9$ necessary floor (Proposition 1). That gap is accessibility slack, not necessity: a K-curriculum ($K = 2 \to 4 \to 8$, the direct mirror of the broad T-curriculum, matched total steps) unlocks selective MQAR down to $d_{\text{state}} = 2 \approx K\theta$ (Table 4). At $d_{\text{state}} = 2$ the curriculum lifts

Table 4: Selective MQAR vs d_{state} (toy, $d_{\text{model}} = 128$): standard SGD vs a K-curriculum ($K = 2 \to 4 \to 8$). Standard fails below $d_{\text{state}} \sim 16$; the K-curriculum solves MQAR down to $d_{\text{state}} = 2 \approx K\theta$ (12/12 seeds). 3 seeds.

d_{state}	$2 (\approx K\theta)$	4	8	16
standard $K = 8$	0.06	0.01	0.11	0.30
K-curriculum	$\mathbf{0.73}$	$\mathbf{0.84}$	$\mathbf{0.93}$	$\mathbf{0.95}$

MQAR from 0.06 to 0.73; the apparent ~ 16 threshold was accessibility slack, and the curriculum reaches down to the estimated $K\theta \approx 1.9$ necessary floor. The effect is *scale-invariant*: at 24M ($60\times$ toy scale) standard still fails and the K-curriculum rescues it (0.71, 0.83, $3/3$ seeds); the $K\theta$ floor is not relaxed by width because it is a rank floor already met by the readout, ample even at $d_{\text{state}} = 2$.

Both axes are accessibility-limited; the asymmetry is the *floor magnitude* $K\theta \ll T\theta$, not a works/fails binary: selective appears easy only because its low floor is reachable under standard SGD. The mechanism is specialization: $|A|$-spread rises under curriculum and stays flat without it (Figure 5, Appendix F).

6.1 A Pareto frontier: the objective controls the allocation

We make the objective-lever claim quantitative as a Pareto frontier on the same char-level wikitext corpus under $L = L_{\text{LM}} + \lambda L_{\text{broad}}$, sweeping $\lambda \in \{0, 0.03, 0.1, 0.3, 1, 3\}$, where L_{broad} is a real-text position-recovery auxiliary defined on the LM's own training distribution: query a 3-gram from the window (sampled to be unique within it) and predict its start position. As in PR the query is revealed after the window, so the pre-query state must keep all positions addressable (broad by the §4 criterion). Both objectives share the backbone through a dual head; a T-curriculum is used so broad is accessible; 3 seeds per point (6 in the transition region); Figure 3 plots LM PPL against broad accuracy.

Three results. (i) At $\lambda = 0$ broad sits at the random floor for *both* architectures, and a fresh probe on the frozen backbone also cannot decode the PR position (Appendix J); the same probe is at chance on 2.7B pretrained checkpoints of both families (Mamba-2, Pythia), so the absence is not an artifact of toy scale: standard language modeling does not teach broad retrieval. (ii) A broad auxiliary lifts broad at a measurable PPL cost: Transformer reaches 0.89, while Mamba-3 plateaus near 0.33 at toy width (rising with width, Appendix H); the emergence edge ($\lambda \approx 0.03$) is seed-dependent (0.40 ± 0.21), the same stochastic-transition signature as §5b. (iii) Transformer dominates Mamba-3 at every matched PPL (0.82 vs. 0.28 at PPL ≈ 5): the cost of co-existing broad retrieval with language modeling is architecture-dependent.

This is the controlled, same-architecture-same-corpus version of the Table 1 dissociation: matched PPL, different broad retrieval. It confirms the §3 thesis — the objective is the allocation lever — and locates the SSM-specific deficit in the *cost* of broad/PPL coexistence (natural-language objectives are predominantly selective, Appendix K, so the broad axis is rarely exercised).

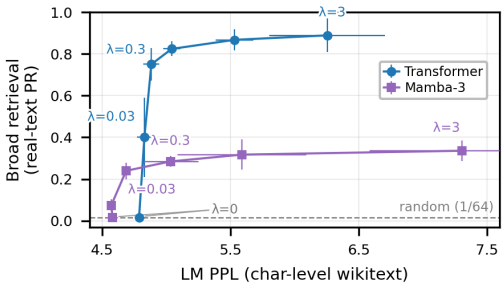

Figure 3: Pareto frontier of language modeling (char-level wikitext PPL, x, lower is better) versus broad retrieval (real-text position recovery, y), sweeping λ in $L_{\text{LM}} + \lambda L_{\text{broad}}$ on the same corpus and backbone (dual head, T-curriculum), 3 seeds per point. At $\lambda = 0$ both architectures sit at the random floor: PPL alone teaches no broad. Transformer dominates Mamba-3 at every matched PPL; Mamba-3 plateaus near broad 0.33. The broad/LM coexistence cost is architecture-dependent. Error bars: SD over seeds (6 seeds in the transition region near $\lambda = 0.03$).

7 Implications

Retrieval dissociates from language modeling. The spine's last clause — *the objective is the lever* — shows already in Table 1: broad-retrieval capacity and LM quality are decoupled — RetNet is the *worst* LM (PPL 44.9) yet the only recurrent architecture reaching broad PR ($1/5$ seeds, 0.978); Mamba-3 is a good LM (PPL 8.2) yet broad-fails (PR 0.014). The fixed-state cluster (Mamba-3, LSTM, GRU, Linear Attn, RWKV-6) shares the broad failure across PPL 8–127: the deficit is not a side effect of LM quality; PPL is blind to it, while a fresh linear probe tracks PR ($\rho = 0.83$; Appendix L).

Scale-invariance and the SSM–Transformer gap (P8). The pattern holds across a $60\times$ range: Mamba-3 solves selective at 24M/122M (MQAR 0.98/0.94) yet broad stays random under standard SGD (122M, 5 seeds); a T-curriculum reaches broad at 24M (PR 0.70) and $1/3$ seeds at 122M ($0.72/0.19/0.09$, the §5b stochastic-transition signature). Broad is where fixed-state compression bites (Transformer 0.99; Mamba 0.014); since broad transfers to selective, training toward the harder superset is principled.

Practical implication and the deficit's boundary. The broad deficit is substantially an *objective* artifact (with trap and bootstrap contributions, §3): on the natural-text distribution we study, PPL rewards selective/local recall rather than broad retrieval (Appendices B.9, K); adding a PR auxiliary (with curriculum, §6) forces broad specialization, and a representation meeting the broad floor meets the selective floor too (Proposition 1). Three downstream tasks localize the deficit precisely: SSMs fail PR (0.014 vs 0.99) but are on par with Transformer on verbatim copy (0.88 vs 0.70, sequential replay) and multi-needle (~ 0.99 both, selective recall) — specifically *random-access position discrimination*, which is why standard benchmarks (selective-dominated) do not see it. For Transformer, broad is architecturally free; for SSM, the objective is the lever.

The accessibility barrier on a real LM. On a from-scratch 30M Mamba LM, direct PR fine-tuning fails but curriculum activates it (0.61, $18\times$) with modest $|A|$-spread change; the rank budget dwarfs $T\theta$ there, so the barrier is optimization-dominant; Transformer learns PR from scratch (0.99), Mamba needs curriculum (0.64; details in Appendix B.10).

8 Discussion and limitations

Synthesis. The theory-derived predictions hold ($1/m$ cliff P6, specialization P5, scale-invariance P8, readout-rank floor P9), and both axes are accessibility-limited, a curriculum unlocking each near its necessary floor: the asymmetry is the magnitude gap $K\theta \ll T\theta$. We keep the theory–experiment tiers explicit: *proven* (identity, ceilings, rank laws, symmetry, Corollary 1); *empirical* (specialization, curriculum accessibility, Pareto); *exploratory* (d_{state} covering, W, bootstrap); *retired* (the low-rank-bias candidate, Appendix B.9).

Real tasks sit on a mixture spectrum. No natural task is purely broad: language dependencies are sparse ($K_{\text{eff}} \ll T$, Appendix K), so real inference is always partly selective. But the broad component is not optional: when *which positions matter* is fixed only at generation time (PR's query-after-window structure), the model must keep every candidate position addressable, and next-token prediction does not train this (§6.1). Whether grounding-type errors (position confusions in otherwise-grounded generations) trace to the broad deficit is testable on existing checkpoints; Appendix M works one example (subject-verb number agreement) end-to-end.

Limitations. PR/MQAR are indirect behavioral probes of SNR_k; per-position leverage is unstable on trained nonlinear SSMs (the identity is verified on synthetic linear models); the curriculum mechanism is not a theorem for nonlinear Mamba; $T \leq 256$; and the broad deficit is shown on controlled probes, not real-task downstream, partly because long-context benchmarks (Hsieh et al., 2024) are dominated by selective retrieval.

9 Conclusion

Is the broad-retrieval failure of SSMs a capacity limit or an optimization artifact? Both, at different scales: capacity sets the floors ($K\theta$ selective, $T\theta$ broad), but at every scale tested the binding constraint was *accessibility*: a curriculum reaches broad where standard SGD stays at chance, and transfers back to selective. The intervention point is timescale placement: broad needs timescales *placed* to match ages, and no permutation-invariant training creates the required asymmetry from an exactly symmetric init. For practice: not more state, not width alone, but training that amplifies specialization — a curriculum does so at every scale tested (at 122M, $1/3$ seeds, §5b's stochastic-transition regime). Whether the deficit surfaces on real tasks is the open question.

AI USE STATEMENT

In this work we used a generative AI tool (an LLM) for tasks in both the *required* and *recommended* disclosure categories. Required: assistance in developing the theoretical model and its conceptual framing (§3), formulating mathematical claims (Theorems 1–2, Propositions 1–5), drafting and checking proofs (Appendix B), refining the bias hypothesis, designing and iterating on experimental methodology, implementing methods (model and training code), and interpreting experimental results. Recommended: writing and editing software code, drafting and editing parts of the manuscript, improving readability, formatting references, and suggesting structure. We did not use generative AI to generate synthetic datasets. No empirical result, figure, or number in this paper was produced by an LLM: every reported value was computed by code that the authors wrote, ran, and inspected, and every proof was checked by the authors. Where the AI initially overstated a result (for example, treating a threshold fit from data as an a priori prediction, or an allocation-simplex volume statement as a loss-landscape claim), we corrected the scope. The authors directed the research, designed the theory and experiments, audited all AI-assisted text and claims, and take responsibility for the final content, including any text or claims produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the retrieval capacity of sequence-model architectures on synthetic tasks and public text corpora (wikitext-103). It involves no human subjects, no personal or sensitive data, and we are not aware of negative societal impacts beyond those generally associated with research on sequence-model analysis.

REPRODUCIBILITY STATEMENT

All experiments follow the architectures, tasks, and hyperparameters in §4 and Appendix G. Training/analysis scripts, per-seed result JSONs, and the commands to regenerate every table and figure are listed in Appendix N; we will release the code and configurations upon publication.

REFERENCES

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, 2019.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations*, 2024a.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, Dylan Zinsley, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2011. doi: 10.1007/s00211-010-0331-6.

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

Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

David C. Hoaglin and Roy E. Welsch. The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17–22, 1978. doi: 10.1080/00031305.1978.10479237.

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? In *First Conference on Language Modeling (COLM)*, 2024.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles, 2026. ICLR 2026.

Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. On the curse of memory in recurrent neural networks: Approximation and optimization analysis. In *International Conference on Learning Representations*, 2021.

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, 2024.

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

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Naoki Nishikawa and Taiji Suzuki. State space models are provably comparable to transformers in dynamic token selection. In *Proceedings of Thirty Seventh Conference on Learning Theory*, 2024.

Destiny Okpekpe and Antonio Orvieto. Revisiting associative recall in modern recurrent models, 2025. SCOPE Workshop at ICLR 2025.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, et al. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. doi: 10.18653/v1/2023.findings-emnlp.936.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

T. Konstantin Rusch and Daniela Rus. Oscillatory state-space models. In *International Conference on Learning Representations*, 2025. Oral.

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, 2023.

Shida Wang and Qianxiao Li. StableSSM: Alleviating the curse of memory in state-space models through stable reparameterization. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Kaiyue Wen, Xingyu Dang, and Kaifeng Lyu. RNNs are not transformers (yet): The key bottleneck on in-context retrieval. In *International Conference on Learning Representations*, 2025.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025.

A NOTATION AND TERMINOLOGY

The paper uses a small set of recurring terms. *broad* and *selective* name the two retrieval *properties* under study; Position Recovery (PR) and MQAR are the *tasks* that operationalize them. Table 5 collects the full vocabulary with the section where each term is defined.

Table 5: Terminology. "Property" entries name a retrieval property; "task" entries are its experimental operationalization; the rest are quantities of the theory.

Term	Meaning	Defined
Retrieval properties (and their tasks)		
broad	every input position stays discriminable in the pre-query state, so any position can be queried after the window	§1, §4
PR (Position Recovery)	task: given a post-window query token, recover its position	§4
selective	recall one associated value by its key; other content need not stay addressable	§1, §4
MQAR	task: multi-query associative recall, K key–value pairs	§4
Capacity quantities (§3)		
leverage budget	$\sum_k \mathrm{SNR}_k = \mathrm{rank}(J) \le d$: fixed budget a recurrent state splits across positions	§3
SNR_k	per-position leverage-SNR proxy of position k	§3
θ	per-position detection threshold; $\theta \in (0.20, 0.25)$, fit from the K-sweep cliff (floors quoted at the cliff-edge value ≈ 0.24)	§3, §5b
$T\theta$ / $K\theta$	necessary rank floors (under the retrieval criterion) for broad / selective retrieval	§3
$\mathrm{rank}(J)$, r_{mem}, r_{acc}	total, memory-side, and accessible-side ranks of the retrieval map M	§3
coverage window W	realized width of the addressable region (ages beyond W at chance); an optimization product, not a static rank	App. I
K_{eff}	effective dependency width of next-token prediction	App. K
Training and analysis		
T-curriculum / K-curriculum	staged training $T = 16 \to 32 \to 64$ (broad) / $K = 2 \to 4 \to 8$ (selective)	§6
specialization	channels taking distinct timescales (broad) or input-dependent gates (selective)	§3, §5c
$\lvert A\rvert$-spread	dispersion of per-channel decay rates; the observed specialization signature	§5c
accessibility	a reachable-in-principle solution that standard training does not find; contrasted with capacity (representability)	§1, §6

B THEORY EXTENSIONS

This appendix states the supporting results (in reading order: the exact leverage identity; the covering lemmas on timescale placement; the input-dependent extension of Theorem 1) and then proves the body results; numerics verify each piece (`verify_theorem_synthetic.py`, `verify_specialization_complexity.py`, `verify_minrank_bias.py`). We state only what holds without extra assumptions.

Why the linear theory guides the nonlinear model. The budget identity (1) is exact for any differentiable state map (Proposition 3), so the fixed-d budget is exact for nonlinear models; input-dependent decay provably cannot exceed $1/m$ (Theorem 2); stacking layers binds the same final d-dimensional state. We claim structure (the cliff's $1/m$ form, the specialization split's sign and ordering), leaving exact thresholds empirical.

B.1 EXACT LEVERAGE IDENTITY AND THE SNR INTERPRETATION

Equation (1) needs no linearization assumption; it is exact.

Proposition 3 (Exact identity). *For any differentiable $s_T = F(x)$ with Jacobian $J = [r_1, \dots, r_T]$ and $\Sigma = JJ^\top$, $\sum_{k=1}^{T} r_k^\top \Sigma^+ r_k = \operatorname{rank}(J) \le d$. Full row rank recovers $\sum_k \mathrm{SNR}_k = d$.*

Proof. Compact SVD $J = U_r S V_r^\top$ with $r = \operatorname{rank}(J)$; $r_k = Je_k = U_r S w_k$ for $w_k = V_r^\top e_k$. Since $\Sigma^+ = U_r S^{-2} U_r^\top$ and $U_r^\top U_r = I_r$, $r_k^\top \Sigma^+ r_k = w_k^\top w_k = \|w_k\|^2$. Summing, $\sum_k \|w_k\|^2 = \|V_r\|_F^2 = \operatorname{tr}(V_r^\top V_r) = r$. □

The reading of $r_k^\top \Sigma^+ r_k$ as a per-position SNR *is* a linearization: for the true finite-difference sensitivity $\eta_k(\delta)$, Taylor expansion gives $|\eta_k(\delta) - r_k^\top \Sigma^+ r_k| = O(\delta)$ (measured $3{\times}10^{-3}$, $3{\times}10^{-4}$, $3{\times}10^{-5}$ at $\delta = 10^{-1}, 10^{-2}, 10^{-3}$), a *linearly*-vanishing residual.

B.2 WHY CHANNELS MUST SPECIALIZE: A COVERING LEMMA

The §5c finding (successful broad backbones spread per-channel timescales) has a rigorous counterpart. In the linearized single-channel model a channel with decay α contributes $r_k = \alpha^k$ at position k; a bank of N channels gives $J_{c,k} = \alpha_c^k$.

Proposition 4 (Single-channel ceiling and identical-channel collapse).

(a) *For one channel, $\max_\alpha \min_k \mathrm{SNR}_k = 1/T$ (attained as $\alpha \to 1$); a single channel cannot solve broad retrieval at any threshold $\theta > 1/T$.*

(b) *If all channels share one decay ($\alpha_c \equiv \alpha$), J has rank 1 and Proposition 3 gives $\sum_k \mathrm{SNR}_k = 1$ (not N): adding channels without timescale diversity adds no budget.*

Proof. (a) $\mathrm{SNR}_k = \alpha^{2k} / \sum_{j=0}^{T-1} \alpha^{2j}$, minimized at $k = T{-}1$. For $\alpha < 1$, every summand with $j < T{-}1$ has $\alpha^{2j} > \alpha^{2(T-1)}$, so $\sum_j \alpha^{2j} > T\alpha^{2(T-1)}$ and $\mathrm{SNR}_{T-1} < 1/T$; as $\alpha \to 1$ the sum $\to T$ and $\mathrm{SNR}_{T-1} \to 1/T$. (b) Identical rows $\Rightarrow \operatorname{rank}(J) = 1$; apply Proposition 3. □

This is the structural reason broad retrieval needs diverse multi-timescale channels (the form of specialization §5c measures). Distinct timescales are necessary but not sufficient; the two-pole case admits an exact statement, which isolates the geometric mechanism — angular separation between adjacent ages — that any covering argument must control.

Proposition 5 (Specialization complexity). *$\operatorname{rank}(J) = \sum_k \mathrm{SNR}_k \le d$ is the retrieval rank; achieving rank r requires at least r linearly-independent channel responses (Proposition 4), and the capacity ratio broad/selective is T/K.*

Lemma 1 (Adjacent-age resolution in a two-pole bank). *Let $v_k = (\rho_1^k, \rho_2^k)$ with $\rho_1 \le \rho_2$, $\kappa = \rho_1/\rho_2 \in (0,1)$ the pole* ratio*, and $\psi_k = \angle(v_k, v_{k+1})$. Then:*

(i) *ψ_k depends only on κ:*

$$\sin\psi_k = \frac{\kappa^k(1-\kappa)}{\sqrt{(1+\kappa^{2k})(1+\kappa^{2k+2})}}, \qquad \tan\psi_k = \frac{\kappa^k(1-\kappa)}{1+\kappa^{2k+1}}. \tag{2}$$

(ii) *The optimal resolution decays linearly with age: for $k \geq 2$,*

$$\frac{1}{4k} \;\leq\; \sup_{\kappa\in(0,1)} \sin\psi_k \;\leq\; \frac{1}{3k}, \tag{3}$$

with sharp asymptotic constant $\lim_{k\to\infty} k \sup_\kappa \sin\psi_k = \frac{c^*+1}{2} e^{-c^*} \approx 0.3314$, where $c^* \approx 1.1997$ is the unique root of $e^{-2c} = (c-1)/(c+1)$. The squared separation correspondingly obeys $\sup_\kappa(1-\cos\psi_k) = \Theta(k^{-2})$.

(iii) *The supremum in* (3) *is attained, up to $o(1)$, at pole ratio $\kappa^* = e^{-c^*/k}$, i.e. at inverse-timescale separation $\frac{1}{\tau_1} - \frac{1}{\tau_2} = \frac{c^*}{k}(1+o(1))$, where $\tau_i = -1/\ln\rho_i$.*

Proof. (i) Write $v_k = \rho_2^k(\kappa^k, 1)^\top$; the common factor cancels in the angle, and $\varphi_k = \arctan(\kappa^k)$, $\varphi_{k+1} = \arctan(\kappa^{k+1})$ give $\psi_k = \varphi_k - \varphi_{k+1}$, hence $\tan\psi_k = \frac{\kappa^k - \kappa^{k+1}}{1+\kappa^k\kappa^{k+1}}$. The sine form follows from $(1+\kappa^{2k})(1+\kappa^{2k+2}) - (1+\kappa^{2k+1})^2 = \kappa^{2k}(1-\kappa)^2 \geq 0$. (ii)–(iii) Substitute $u = -\ln\kappa > 0$ and bound the denominator by $1 + e^{-(2k+1)u}$ (the difference of the squares is $e^{-2ku}(1-e^{-u})^2 \geq 0$): $\sin\psi_k \leq \frac{e^{-ku}(1-e^{-u})}{1+e^{-(2k+1)u}}$. At $u = c/k$ the right side tends to $\frac{c}{k} \cdot \frac{e^{-c}}{1+e^{-2c}}$, maximized at c^* solving $e^{-2c} = (c-1)/(c+1)$, giving the sharp constant and (iii). The two-sided bound (3) is elementary calculus on the same expression (verified numerically for all $k \leq 10^4$; `theory_upgrade/verify_two_pole.py`); the $1-\cos$ claim follows from $1-\cos\psi = \sin^2\psi/(1+\cos\psi)$ with $\cos\psi_k \to 1$ at the optimizer. □

Timescale matching. The complementary half of the covering story says *which* timescale serves age k:

Lemma 2 (Timescale matching)**.** *For a single pole $\rho = e^{-1/\tau}$ in a length-T window, $\mathrm{SNR}_k(\tau) = \frac{e^{-2k/\tau}(1-e^{-2/\tau})}{1-e^{-2T/\tau}}$, and for $\tau \leq T$: $\mathrm{SNR}_k(\tau) \leq \frac{2e^{-2k/\tau}}{\tau(1-e^{-2})} \leq \frac{2.32\, e^{-2k/\tau}}{\tau}$. Consequently (i)* fast-pole starvation*: if $\tau \leq k/2$ then $\mathrm{SNR}_k \leq 0.085/k$; (ii)* matching*: the matched pole $\tau = k$ attains $\mathrm{SNR}_k \approx 2e^{-2}/k \approx 0.271/k$, more than $3\times$ the starved value; (iii) ages beyond every pole are starved: $\mathrm{SNR}_k \leq 2.32\, e^{-2k/\tau_{\max}}/\tau_{\max}$ whenever all $\tau_c \leq \tau_{\max} \leq T$.*

Proof. $\mathrm{SNR}_k = \rho^{2k}\frac{1-\rho^2}{1-\rho^{2T}}$ (Proposition 4); substitute $\rho = e^{-1/\tau}$ and use $1 - e^{-2T/\tau} \geq 1 - e^{-2}$, $1 - e^{-2/\tau} \leq 2/\tau$ for $\tau \leq T$. (i) follows from $e^{-2k/\tau} \leq e^{-4}$ and $1/\tau \geq 2/k$; (ii) evaluates at $\tau = k$ with $1 - e^{-2/k} \approx 2/k$; (iii) is immediate. □

Together, Lemmas 1–2 give the local mechanism behind the covering effect of §5c–§6: to resolve ages near k, a bank needs a pole pair *tuned to k* — timescale $\approx k$ (Lemma 2) and relative separation $\approx c^*/k$ (Lemma 1) — and even then the achievable angular resolution of adjacent ages decays as $\Theta(1/k)$ (squared separation $\Theta(1/k^2)$). This is the first result here that quantifies *where* distinct timescales must be placed, not merely that they must be distinct. We emphasize the limits: both lemmas are local two-pole statements and do not by themselves yield a global law $W(d_{\mathrm{state}})$ for the coverage window observed empirically (Appendix I); the global question is addressed numerically next.

B.3 Input-dependent per-channel impossibility

Theorem 1 uses a fixed decay α. Real SSMs use input-dependent gates, so a natural worry is that input-dependence evades the $1/m$ ceiling. It does not, on the retention side. Consider one channel with per-step retention $A_t \in (0,1)$ and write gate B_t, $s_t = A_t s_{t-1} + B_t x_t$, processing $[\textsc{key}, f_1, \ldots, f_m, \textsc{query}]$ read out at the query. Let $P_k = \prod_{t=k+1}^m A_t$ (so $P_0 = \prod_{t=1}^m A_t$ is the key residue, up to B_0); the filler-k residue is P_k.

Theorem 2 (Input-dependent retention bound). (a) ***Two-class gate.*** *If $A_t \equiv A$ and the write gate is B_K on the key, B_F on fillers, then* $\mathrm{SNR} \le (B_K/B_F)^2/m$*; a non-selective channel ($B_K = B_F$) gives* $\mathrm{SNR} \le 1/m$*, recovering Theorem 1.*

(b) ***Non-selective retention.*** *If the write gate is uniform ($B_t \equiv B$) but $\{A_t\}$ is arbitrary (any input-dependence), then* $\max_{\{A_t \in (0,1)\}} \mathrm{SNR} = 1/m$*, attained iff $A_t \to 1$ for all t.* Input-dependent decay alone cannot exceed $1/m$.

Proof. (a) $\mathrm{SNR} = (B_K/B_F)^2 A^{2m} / \sum_{j=0}^{m-1} A^{2j}$. Since $A \in (0,1)$ and $j \le m-1 < m$, each $A^{2j} \ge A^{2m}$, so $\sum_j A^{2j} \ge m A^{2m}$ and $\mathrm{SNR} \le (B_K/B_F)^2/m$; equality needs all A^{2j} equal, i.e. $A = 1$. (b) With $B_t \equiv B$ the prefactors cancel, $\mathrm{SNR} = P_0^2 / \sum_{k=1}^{m} P_k^2$. As $A_t \le 1$, the sequence $P_m = 1 \ge P_{m-1} \ge \cdots \ge P_0$ is monotone, so every $P_k \ge P_0$, hence $\sum_k P_k^2 \ge m P_0^2$ and $\mathrm{SNR} \le 1/m$, with equality iff every $A_t = 1$. □

Part (b) is the substantive claim: it makes the Scope statement ("input-dependent α_t does not remove the specialization pressure") a theorem. We do *not* claim the converse, that exceeding $1/m$ requires a *content*-dependent gate; a position-dependent gate can also exceed it, so the content/position distinction is empirical, not implied here.

B.4 Proofs of the main results

Theorem 1. $\mathrm{SNR} \le 1/m \iff g(\alpha) := \alpha^{2m}(m+1-m\alpha^2) \le 1$. $g(1) = 1$ and $g'(\alpha) = 2m(m+1)\alpha^{2m-1}(1-\alpha^2) \ge 0$ on $(0,1)$, so g is increasing and $g(\alpha) \le g(1) = 1$.

Proposition 5. Selective needs $\mathrm{SNR}_k \ge \theta$ on K queried positions, broad on all T; sum and apply Proposition 3 ($\sum_k \mathrm{SNR}_k = \mathrm{rank}(J)$). Distinct-timescale channels raise rank (Proposition 4(b)); coincident decays contribute rank 1 regardless of count.

Proposition 2. (a) $\mathrm{SNR}_k = r_k^\top \Sigma^+ r_k$ with $r_k = J e_k$, $\Sigma = J J^\top$, is invariant under the simultaneous row-permutation $J \mapsto \Pi J$ (channels are rows of J). (b) Permutation invariance makes ∇L symmetric at any symmetric point (all entries equal), so gradient flow maps the symmetric subspace to itself; rank-1 along it is Proposition 4(b). (c) follows from (b) by induction on time. Escaping needs symmetry breaking; linearized, the spread on the $(N-1)$-dim breaking subspace grows as $e^{\lambda_b t}$ at a rate λ_b set by the Hessian, so the time to reach the broad-required spread $\sigma_{\mathrm{target}} \propto T$ scales as $\lambda_b^{-1} \log(\sigma_{\mathrm{target}}/\sigma_{\mathrm{init}}) \propto \log T$.

B.5 Curriculum invariance: proof and scope

Corollary 1 (Curriculum invariance). *Let $\mathcal{L}_1, \mathcal{L}_2, \ldots$ be* any *finite or infinite sequence of channel-permutation-invariant losses (any curriculum over permutation-invariant objectives, including staged T- or K-curricula), and let gradient descent with any positive step sizes be applied in any interleaving. If the initialization is exactly symmetric ($\alpha_1 = \cdots = \alpha_N$), then every iterate remains symmetric, hence* $\mathrm{rank}(J_t) = 1$ *and* $\min_k \mathrm{SNR}_k < 1/T$ *at every step t. In particular, no curriculum over permutation-invariant objectives can reach broad retrieval from an exactly symmetric initialization: a curriculum can help only by amplifying* pre-existing *asymmetry (symmetry breaking), not by creating it.*

Proof. By induction. If $\alpha^{(t)}$ is symmetric, then for every permutation P of the channel indices, $\mathcal{L}_{t+1}(\alpha^{(t)}) = \mathcal{L}_{t+1}(P\alpha^{(t)})$, so the gradients satisfy $\nabla\mathcal{L}_{t+1}(\alpha^{(t)}) = P^{-1}\nabla\mathcal{L}_{t+1}(\alpha^{(t)})$ for all P. The only vectors fixed by every channel permutation are multiples of $\mathbf{1}$, so the per-channel gradient components are all equal, and $\alpha^{(t+1)} = \alpha^{(t)} - \eta_{t+1}\nabla\mathcal{L}_{t+1}(\alpha^{(t)})$ is again symmetric. The rank and SNR claims are Proposition 4. □

Scope. The corollary holds for the linearized bank and for exactly symmetric initializations. The corollary covers full-batch gradient descent with any positive step sizes; Adam with zero-initialized

moments performs the same symmetric update (elementwise operations commute with permutations), and minibatch training requires the per-example loss to be equivariant under the joint channel-and-readout permutation, which the retrieval loss satisfies. Real Mamba's per-channel parameters are only approximately symmetric at random init, so the corollary does not constrain what a curriculum can do from such an init (§6: it reaches broad); what it rules out is the specific route of *creating* channel asymmetry from exact symmetry. In particular, a curriculum that succeeds must be amplifying init-time asymmetry into specialization — the observed rising per-channel spread under curriculum (Figure 5, Appendix F) is this route, not an alternative to it.

B.6 The covering law of the linearized model: a numerical observation

Lemmas 1–2 are local; the global counterpart is a covering question. Define $\mathrm{OPT}(N,T) = \max_{\alpha\in(0,1)^N} \min_k \mathrm{SNR}_k$: the best uniform coverage of T positions by N decay channels. We report numerical facts (`theory_upgrade/covering_rate2.py`, `theory_upgrade/geom_clock.py`; differential evolution plus multi-start Nelder-Mead over half-lives), stated as observations, not theorems:

- *Geometric clocks are near-optimal.* Restricting the half-lives to geometric progression $\tau_c = \tau_{\min}(\tau_{\max}/\tau_{\min})^{c/(N-1)}$ and optimizing the two endpoints reproduces the free N-parameter optimum exactly at $N \in \{4, 6, 8\}$ and within 4–8% at $N = 16$ ($T \le 256$): the optimally placed bank *is* a log-spaced clock, which is precisely the placement Lemma 2 asks for ($\tau \approx k$ for each covered age k).
- *Covering efficiency is bounded below* 1*, and decreases.* $\mathrm{OPT}(N,T) \approx 0.6\, N/T$ for $N \le 8$; the efficiency $\mathrm{OPT}/(N/T)$ then *decreases* with N (0.50 at $N=16$, 0.29 at $N=32$, 0.19 at $N=48$; $T=64$).
- $\mathrm{OPT}(N,T)$*T saturates.* At $T=64$ it grows $1.3, 2.4, 4.9, 6.9, 8.1$ for $N=2, 4, 8, 12, 16$ and plateaus at ≈ 9.0–9.1 for $N \ge 24$ (within optimizer noise across $T \in \{32, \ldots, 256\}$). Within the pure-decay family, additional channels stop buying uniform coverage: $\min_k \mathrm{SNR}_k \lesssim 9/T$ no matter how many decay channels are available.

Two connections, both stated with care. First, the saturation is the global counterpart of Lemma 1's local decay: at small N reaching broad ($\min_k \mathrm{SNR}_k \ge \theta$) needs $N \gtrsim 1.7\, T\theta$ — consistent with the empirical onset of broad at $d_{\mathrm{state}} \approx 32 \approx 2T\theta$ (Appendix H) — while at large N the ceiling $\approx 9/T$ makes broad unreachable in the pure-decay family at any $\theta = O(1)$, so the trained model must use structure beyond pure decay (readout-time binding, Appendix L). Second, this explains why the d_{state} curve is *soft* (Appendix H): d_{state} buys covering efficiency at rate ≈ 0.6, not a rank threshold. We do not claim a closed-form $W(N,T)$ law; the saturation value and efficiency curve are empirical regularities of the decay family. (Optimization caveat: a local L-BFGS run from generic inits *underestimates* OPT relative to the geometric clock; all reported values use the global-search agreement above.)

Kernel-family envelopes. The decay family is not the only recurrent kernel; the covering analysis extends to other families, and the comparison isolates *where the crowding lives* (`theory_upgrade/directionB_envelopes.py`, `directionB_quick.py`; same OPT protocol, state budget counted fairly — a conjugate pole pair costs 2 state dimensions):

family	adjacent-age resolution	$\mathrm{OPT}\cdot T$ at state budget		
	(scaling in k)	4	8	16
exponential (decay)	$\Theta(1/k)$ (Lemma 1)	2.4	4.9	$\sim$8–9 (saturates)
oscillatory (conj. pairs)	$\Theta(1)$	3.3	6.5	13.4 (no saturation)
power-law (fractional)	$\sim k^{-1.5}$	1.9	2.6	2.8 (saturates)

Three readings. (i) *Oscillatory kernels dominate both axes*: their adjacent-age resolution is $\Theta(1)$ (the angle between conjugate-quadrature responses at ages $k, k{+}1$ is set by the frequency ω, independent of k), and their covering efficiency stays ≈ 0.82 at every tested budget while the decay family's drops — the frequency basis is uncrowded where the decay basis crowds. (ii) *Power-law kernels are worse on both axes*: resolution $\sim k^{-1.5}$ and a covering ceiling near 2.8, below even the exponential family.

(iii) *The LinOSS puzzle*: oscillatory SSMs such as LinOSS (Rusch & Rus, 2025) sit in the *strong* family — their covering capacity should be ample — yet they fail PR entirely (mean PR 0.004; Table 1). The envelope comparison therefore predicts their failure is *accessibility*, not capacity: the optimization barrier of §6, transplanted to a family whose capacity is not the binding constraint. This closes the loop on our earlier forward-validation miss ("oscillation should help broad"): oscillation helps *capacity*, and capacity was never the binding side. A testable prediction follows — a T-curriculum (or symmetry-breaking init) should unlock PR on LinOSS; we leave the test to future work.

B.7 What the linearized model can and cannot show

The covering law predicts that within the pure-decay class at our working size ($N=16$, $T=64$), no training procedure can reach broad: the covering ceiling ($\mathrm{OPT}\approx 0.127$) is below $\theta\approx 0.225$. We verify this directly on linearized training dynamics (`theory_upgrade/escape_dynamics.py`; Adam on a soft $\min_k$SNR broad surrogate, 10 seeds):

- *Exact symmetry is a hard trap* (rigorous, Corollary 1): from a symmetric init the spread stays at machine zero throughout training.
- *Near-symmetric inits specialize but misplace.* From init spread $\sigma \in \{0.005, \ldots, 0.1\}$, training reaches channel rank $\approx$ 7–9 but $\min_k \mathrm{SNR}_k \leq 0.06$ in $0/10$ seeds at every σ — the flow finds *some* timescale diversity but not the log-spaced placement of Lemmas 1–2.
- *A T-curriculum does not rescue the linearized model.* Staged training ($T=16\to32\to64$, matched steps) raises the channel spread but leaves $\min_k \mathrm{SNR}_k \leq 0.09$ ($0/10$ seeds): the covering ceiling binds before any curriculum benefit can materialize.

The division of labor is thus deliberate. The linearized theory supplies the *necessity* side — symmetry barrier (Proposition 2, Corollary 1), timescale matching and resolution decay (Lemmas 1–2), and the covering ceiling (above) — while the *sufficiency* side, that a curriculum and not standard SGD reaches broad in the real (nonlinear, gated) model, remains an empirical accessibility finding (§6, §8), as the scope statement in §3 promises.

B.8 Specialization complexity and the symmetric obstruction: numerical verification

We verify Propositions 2 and 5 on the linearized bank-of-channels model $J_{c,k} = \alpha_c^k$ (N channels, $T = 64$ positions), with $\mathrm{SNR}_k = r_k^\top \Sigma^+ r_k$, $\Sigma = JJ^\top$ and rank via the SVD (`scripts/verify_specialization_complexity.py`).

Proposition 5 (specialization complexity). $\mathrm{rank}(J)$ tracks the number of *distinct* decays, not the channel count: 4 distinct decays give rank 4 ($\sum_k \mathrm{SNR}_k = 4.00$), 4 coincident give rank 1 ($\sum = 1.00$), and 8 decays in 4 coincident pairs give rank 4 ($\sum = 4.00$)—coincident channels add no budget, exactly Proposition 4(b). The capacity floors are sharp: broad needs $\mathrm{rank} \geq T\theta = 15.4$, selective only $\mathrm{rank} \geq K\theta = 1.9$ (ratio $T/K = 8$; at the cliff edge $\theta = 0.24$). The necessity is exacting: over 2000 random allocations, the best achievable $\min_k \mathrm{SNR}_k$ stays below $\theta = 0.24$ even at $N = 48$ channels (best ≈ 0.10), and direct optimization of the decays saturates at $\min_k \mathrm{SNR}_k \approx 0.14$ ($T = 64$, N large; Appendix B.6), reflecting the rank-vs-retention tension of the pure decay model (distinct decays raise rank; slow decays retain late positions; both are needed). *The floor above is a rank floor, not a hyperparameter floor:* Proposition 1 constrains $\mathrm{rank}(J)$, *not* the Mamba hyperparameter d_{state}, and the sub-floor sweep of Appendix H empirically rules out identifying the two (broad reaches 0.19 at $d_{\mathrm{state}} = 2 \ll T\theta$ via readout-time binding). This is consistent with broad being reached at $d_{\mathrm{state}} = 32$ ($\sim 2\times$ the floor) in §6 and via readout-time binding (Appendix L), not by the pure state-content model alone; the linear theory delivers the scale-correct *necessity*, and the non-uniform leverage reflects the rank-vs-retention tension of the pure decay model.

Proposition 2 (symmetric fixed point). (a) Permutation invariance holds to numerical precision: permuting the channel indices changes every SNR_k by $< 5\times10^{-9}$. (b) The symmetric configuration $\alpha = (\mu, \ldots, \mu)$ collapses $\mathrm{rank}(J)$ to 1 for every $\mu \in \{0.3, \ldots, 0.999\}$, giving $\sum_k \mathrm{SNR}_k = 1 \ll T\theta$; as $\mu \to 1$, $\min_k \mathrm{SNR}_k \to 1/T = 0.0156$, recovering Proposition 4(a). (c) Gradient flow on a

permutation-invariant broad-retrieval surrogate $\sum_k \mathrm{softplus}(\theta - \mathrm{SNR}_k)$ preserves symmetry: from a symmetric init the channel spread stays 0.000 (loss $51.95 \to 51.94$), whereas from an asymmetric init ($\sigma \approx 0.02$) the spread grows to 0.34 and $\mathrm{rank}(J)$ reaches 16. Symmetry, once absent at init, is never created by the gradient; asymmetry is the only route to specialization, and the linearized escape rate gives the $\log T$ specialization-time scaling (`theory_upgrade/escape_dynamics.py`).

The bias hypothesis on trained backbones. We also test the hypothesis's directional prediction on the trained PR backbones themselves (`scripts/analyze_linearized_rank.py`): building the linearized $J_{c,k} = \alpha_{0,c}^k$ from each checkpoint's per-channel discrete decay $\alpha_0 = e^{-dt \cdot |A|}$ (42 toy-scale backbones), $\mathrm{rank}(J)$ is higher for broad-success than broad-fail (6.3 vs. 5.3, Mann–Whitney $p = 10^{-4}$), as is the spread (0.10 vs. 0.06, $p = 6 \times 10^{-4}$). Both groups sit well below the $T\theta = 15.4$ floor—the linearized rank is conservative in absolute terms because the nonlinear model solves broad via readout-time binding (Appendix L), not by the pure state-content rank—so the test of the hypothesis is the *ordering* (successful broad specializes more channels), not the absolute rank, and the ordering holds.

B.9 The low-rank-bias candidate: motivation, tests, and why it is retired

Why does SGD under-reach broad retrieval from approximately symmetric initializations? The causal results (§6) establish *that* it does and *where the lever is*, but not the microscopic mechanism. This appendix documents the natural candidate — an implicit low-rank bias of the optimizer — and why we retired it: its motivation does not transfer to the SSM factorization, a gauge redundancy decouples parameterization knobs from the pole placement the mechanism would need to act on, and three interventions at three budget margins returned no support. We keep the account because the negative result constrains future explanations: the mechanism is not a generic implicit bias of training on this architecture. The matrix-sensing structure itself remains load-bearing for other parts of the paper (readout-time binding, Appendix L).

Retrieval as matrix sensing. With a linear readout W and the linearized state map $s_T = Jx$, the input→output map is the effective retrieval operator

$$M = W^\top J, \qquad M[\ell, k] = w_\ell^\top r_k, \qquad M = W^\top \mathrm{diag}(\bar{B})\, V(\alpha), \quad V(\alpha)[c, k] = \alpha_c^{T-k}, \quad (4)$$

a depth-≥ 2 linear factorization in the free factors (W, B) over the structured pole factor $V(\alpha)$ ($\bar{B}_c$ is the effective gain $((e^{\Delta t A_c} - 1)/A_c)B_c$, the gauge-invariant object; for PR, M binds the query to the matching position; for MQAR, to the queried key's value; the readout-time binding of Appendix L). Retrieval training senses M through its loss.

Motivation (not a theorem), and why it does not transfer. Implicit-bias results for *unrestricted* depth-2 matrix factorizations associate small-initialization GD with minimum-nuclear-norm solutions (a low-rank surrogate) (Gunasekar et al., 2017; Arora et al., 2019). Three gaps: our factorization is structured ($V(\alpha)$ is optimized jointly with (W, B), outside these theorems); minimum nuclear norm is not minimum rank; and the results concern GD on the free factors, not our optimizer. A numerical illustration of the tendency (`scripts/verify_minrank_bias.py`): a free depth-2 factorization ($T = N = 32$) converges to effective rank $4.4 \to 2.5$ (target rank 2) as init scale shrinks $\alpha = 10^{-2} \to 10^{-5}$; the structured SSM factorization contracts the same way. This illustrates the *tendency*, not a transferable theorem. Independently of any bias, the task statistics alone put PPL optima at low rank: next-token prediction depends effectively on $K_{\text{eff}} \ll T$ positions (Appendix K), so positions beyond that context can have their SNR_k driven to ≈ 0 at no loss — a statistical floor $\mathrm{rank}(J(\theta^\star)) \sim K_{\text{eff}}\theta$ that requires no optimizer bias at all.

Tests: three interventions, no support. We tested the candidate channel (weight decay pulling $A_{\log}$, hence each head's decay rate, toward the coincident-decay configuration of Proposition 4(b)) by exempting $A_{\log}$ from the penalty at matched decay, curriculum training, mamba3, across three budget margins; and the bias itself via two theory-designated knobs (init scale; optimizer family). Table 6 collects the results; scripts: `sweep_wd_hypothesis.py`, `sweep_wd_24m.py`, `sweep_wd_margin.py`, `sweep_bias_intervention.py`.

Neither harm nor rescue is established at any margin. Two design facts qualify the 24M cell (weaker cumulative dose at its learning rate; an order-of-magnitude larger budget margin, where a floor-margin mechanism would not bind), which is why the floor-margin row exists: it removes those

Table 6: Weight-decay/A-exemption tests (broad PR under curriculum, mean over seeds; "solved" = seeds with PR ≥ 0.5). Toy: $d_s = 32$, margin $2.2\times$ floor. 24M: causal recipe. Margin: $d_s = 16$, margin $1.1\times$ floor, toy-matched cumulative dose.

scale	config	PR (mean±SD)	solved	read
toy (3 seeds)	wd 0 / 0.01 / 0.05	0.25 / 0.46 / 0.25	—	non-monotone, all cells SD 0.17–0.32
	wd 0.1 full / exclA	0.27 / 0.45	1/3 vs 3/3	Fisher $p = 0.4$; collapse also at wd 0
24M (3+3+3)	0.01 (default) / 0.1 full / 0.1 exclA	0.70 / 0.58 / 0.42	3/3, 3/3, 2/3	no A-specific effect; $\lvert A\rvert$-spread flat across cells
margin (3+3)	0.1 full / 0 baseline	0.56 / 0.20	3/3 vs 1/3	at pinned margin the *decayed* runs are stable
bias knobs (5 seeds)	init $0.3/1/3\times$	0.45 / 0.68 / 0.02	—	$3\times$ = training-regime failure, not a bias signature
	GD / SGdm / AdamW	0.01 / 0.01 / 0.68	—	GD never leaves the initial plateau (under-optimized)

qualifiers and still shows no harm — the zero-decay baseline is the one that collapses, at the floor's stochastic-transition rate (§5b). We report the full account to caution against prescribing parameterization-level interventions from toy-scale signatures.

Why parameterization knobs do not couple. A structural fact explains *a posteriori* why the wd intervention could not move pole placement: the linearized recurrence carries an exact gauge redundancy $(A, \Delta t, B) \to (\lambda A, \Delta t/\lambda, \lambda B)$, preserving M and the loss exactly (verified to 10^{-11}; `theory_upgrade/directionA_gauge.py`); the training gradient is orthogonal to the gauge direction. Weight decay on $A_{\log}$ alone is not gauge-invariant, so the optimizer can neutralize it by drifting the gauge (the input-dependent Δt channel) while leaving every pole in place, consistent with the flat $|A|$-spread above.

Status. The candidate is retired: motivation does not transfer (structured factorization), its parameterization consequence is gauge-absorbed, and direct interventions were confounded or negative across three margins. The mechanism of the under-reach remains open; the causal core (objective sets the allocation, failure is an active trap, bootstrapping is barriered) does not depend on it.

B.10 Sequential entanglement: why SSM representations are zero-sum

The zero-sum has a structural root in the recurrence: positions' SNR allocations are *coupled* through the shared decay chain, unlike attention's independent per-position weights. By direct calculation of the recurrence, $s_T = (\prod_{t=k+1}^{T} A_t) B_k x_k + \ldots$, so the Jacobian column for position k is $r_k = \partial s_T / \partial x_k = \Phi_k B_k$ with $\Phi_k = \prod_{t=k+1}^{T} A_t$ the cumulative decay. For $j < k$, $r_j = \Phi_k \cdot \Psi_{j,k} \cdot B_j$ with $\Psi_{j,k} = \prod_{t=j+1}^{k} A_t$. The tail Φ_k is shared: any parameter change that adjusts r_k (via some A_t for $t > k$) simultaneously adjusts r_j for all $j < k$. Attention weights $\alpha_k = f(Q_T, K_k)$, in contrast, depend on (Q_T, K_k) independently per position (apart from the weak softmax-normalization coupling).

Why this creates the SSM-specific barrier (not the zero-sum). PPL's optimal allocation concentrates SNR on $\sim K_{\text{eff}}$ recent positions (the statistical argument above); PR's spreads it uniformly over all T (Proposition 5). To move from PPL's allocation to PR's, one must raise distant positions' SNR by modifying the decay chain, but the shared tail Φ_k means this simultaneously perturbs all earlier positions' SNR. This coupling makes the PR optimization landscape narrow and hard for SGD, requiring curriculum (§6). Attention avoids this difficulty: each position's weight is independently tunable ($Q_T \cdot K_k$), so redirecting attention from selective to uniform is a simple per-position adjustment. Note: this explains why SSMs learn PR *harder* (barrier), and is silent on why PR and PPL conflict (zero-sum). The zero-sum is general catastrophic forgetting (both architectures show it), arising because any backbone modification for PR disrupts PPL features. The SSM-specific contribution is the barrier (the zero-sum is general).

B.11 Theorem 1 scope

Theorem 1 concerns the abstract scalar recurrence with discrete decay $\alpha \in (0, 1)$ ($\alpha \to 1$ = retention), not any particular parameterization of a trained SSM's A or Δt. We validate it via the K-sweep cliff (§5b) and our own spread-based specialization (§5c), not via absolute decay markers, whose polarity relative to this α is parameterization-dependent.

C MQAR K-SWEEP FULL TABLE

Full 4-seed MQAR table (Table 7) (mean $\pm$ SD over seeds $\{0, 1, 2, 42\}$; per-seed values in `experiments/mqar_Ksweep/`). θ is read off as the K where non-selective accuracy first collapses ($K = 8$, $1/m = 0.143$), bracketed above by $K = 4$ ($1/m = 0.33$): $\theta \in (0.143, 0.33)$. The denser T-controlled sweep below (3 seeds) and its 8-seed extension at the cliff edge tighten this to $\theta \in (0.20, 0.25)$.

Table 7: MQAR accuracy (4-seed mean $\pm$ SD) vs. K ($d_{\text{state}} = 32$). The selective/non-selective split appears at $K = 8$; the K=8 per-seed values are Mamba $0.37/0.47/0.44/0.49$, split_dt $0.85/0.79/0.53/0.52$, RWKV $0.02/0.07/0.08/0.06$, RetNet $0.14/0.13/0.14/0.14$.

model	K=4	K=8	K=16	K=32
Mamba (selective)	0.972 ± 0.034	0.442 ± 0.055	0.012 ± 0.002	0.012 ± 0.001
split_dt (selective)	0.989 ± 0.003	0.670 ± 0.173	0.013 ± 0.001	0.011 ± 0.001
RWKV (non-selective)	0.995 ± 0.004	0.057 ± 0.025	0.014 ± 0.006	0.012 ± 0.000
RetNet (multi-scale)	0.266 ± 0.003	0.138 ± 0.003	0.069 ± 0.004	0.017 ± 0.005

T-controlled dense sweep. The legacy sweep above lets the padded context scale with K; to test that the cliff's control variable is $m = K-1$ rather than T, we re-swept with a dense grid at a *fixed* context $T \in \{64, 128, 256\}$, independent of K (`scripts/sweep_mqar_K_T.py`; $d_{\text{state}} = 32$, 3 seeds; `experiments/mqar_Ksweep_T.json`). Mean accuracies:

Table 8: T-controlled dense K-sweep (3-seed mean). Rows: model $\times$ T; columns K (the Theorem-1 variable $1/(K-1)$ in the header). RWKV's cliff and RetNet's slow decay are T-invariant; selective models collapse in order of selectivity. Random ≈ 0.008.

	K=4 .33	5 .25	6 .20	7 .17	8 .14	10 .11	12 .09	16 .07
rwkv T=64	1.00	0.41	0.39	**0.07**	0.06	0.03	0.02	0.01
rwkv T=128	1.00	0.42	0.39	**0.06**	0.04	0.03	0.02	0.01
rwkv T=256	0.99	0.12	0.09	0.07	0.05	0.03	0.02	0.02
mamba T=64	0.97	0.95	0.72	0.48	0.42	0.11	0.01	0.01
mamba T=128	0.95	0.81	0.59	0.51	0.45	0.21	0.08	0.01
split_dt T=64	0.99	0.99	0.99	0.97	0.64	0.03	0.01	0.03
split_dt T=128	0.99	0.99	0.99	0.99	0.96	0.08	0.01	0.02
split_dt T=256	0.99	0.99	0.99	0.99	0.98	0.08	0.03	0.01
retnet T=64	0.27	0.21	0.18	0.16	0.14	0.11	0.09	0.07
retnet T=128	0.27	0.21	0.18	0.16	0.14	0.11	0.09	0.07
retnet T=256	0.27	0.21	0.18	0.16	0.14	0.12	0.09	0.07

The cliff is a per-run stochastic phase transition. Per-seed accuracies at the cliff edge are bimodal: RWKV at $K = 5, 6$, $T = 64$, 8 seeds gives $K = 5$: $0.97, 0.13, 0.12, 0.99, 0.99, 1.00, 0.90, 0.13$ ($p_{\text{solve}} = 6/8$) and $K = 6$: $0.09, 0.09, 0.99, 0.07, 0.08, 0.09, 0.08, 0.07$ ($p = 1/8$) — every run at ≥ 0.9 or ≤ 0.13, no intermediate values (`experiments/cheap_validations.json`, P6). The mean curve is therefore a solve probability, and the cliff location behaviorally brackets the effective θ under the retrieval criterion at $(1/5, 1/4) = (0.20, 0.25)$ (midpoint 0.225, consistent with the 3-seed bracket). This is a behavioral threshold read off task accuracy, not a direct SNR measurement, and it is not perfectly T-invariant: at $T = 256$ the RWKV cliff shifts one K earlier ($K = 5$ drops to 0.12), consistent with the larger-T bootstrap being harder (Appendix I). The same bimodal signature recurs elsewhere: the broad T-curriculum click-in and the λ-frontier edge region (§6.1), suggesting a common stochastic-transition structure.

RetNet tracks the $1/m$ line. Across the full T-controlled grid, multi-scale RetNet's accuracy divided by the Theorem-1 bound $1/(K-1)$ is $0.80, 0.84, 0.91, 0.95, 0.96, 1.01, 1.03, 1.03$ for $K =$

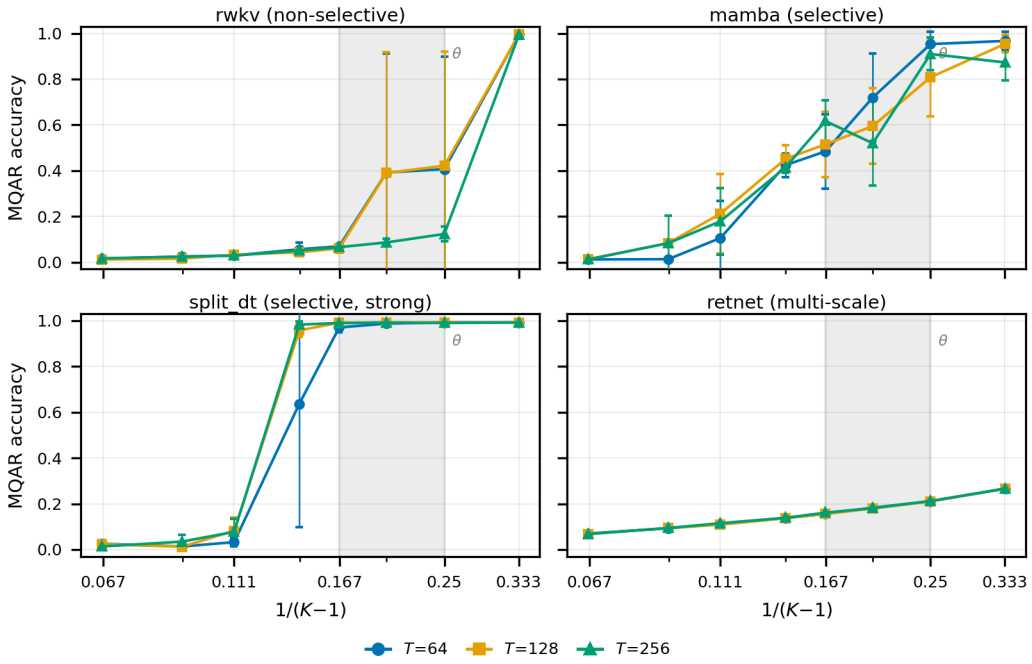

Figure 4: MQAR accuracy vs the Theorem-1 control variable $1/(K-1)$ (the per-position SNR bound $1/m$), three fixed context lengths overlaid (3 seeds, mean±SD; Table 8). The curves are largely controlled by $1/(K-1)$ rather than T: RetNet collapses nearly exactly across T (identical to two decimals), while RWKV and the selective models show residual T dependence near their transition (at $T = 256$ the RWKV cliff shifts one K earlier, consistent with the bootstrap limit of Appendix I); the shaded band marks the θ region $1/(K-1) \in (0.167, 0.25)$ spanned by the dense RWKV cliff; the 8-seed edge extension tightens the effective bracket to $(0.20, 0.25)$ (Appendix C).

$4, 5, 6, 7, 8, 10, 12, 16$ — within $\sim 20\%$ of the line over a $5\times$ range of the bound, at all three context lengths (identical to two decimals across T). We note the quantity distinction: Theorem 1 bounds per-channel SNR, while this ratio compares task accuracies, so the match is a behavioral correspondence to the theoretical ceiling rather than a direct verification of it. RetNet behaves as an ideal *non-selective* working point in this behavioral sense: it does not beat the filler-noise scaling and decays as the bound predicts; selective architectures exceed the line at small K and cross it at their own threshold.

D PER-CHECKPOINT SPECIALIZATION

Per-backbone $|A|/\alpha_0/dt$ values and the success/fail grouping (`experiments/alpha_specialization_{pr,mqar}.json`); M1 (state-dict-only) metrics summarized in Table 9.

Table 9: Channel specialization, success (PR/MQAR≥ 0.5) vs. fail (<0.1): mean spread (a uniform backbone has near-zero spread). Spread is consistently higher in successful backbones on every metric.

task	metric	success	fail
broad (PR)	$\lvert A\rvert$ spread	0.447	0.250
broad (PR)	α_0 spread	0.287	0.159
broad (PR)	dt std	0.191	0.021
selective (MQAR)	dt std	0.243	0.029

E READOUT-RANK ABLATION (P9)

Retrieval operates through $M = W^\top J$ (Appendix B.9), so $\mathrm{rank}(M) \le \mathrm{rank}(W)$ (Table 10): a readout constrained to rank r cannot support a task needing $\mathrm{rank} \ge T\theta$. We ablate the readout rank (low-rank head, standard Mamba-3, $d_{\mathrm{state}} = 32$, $T = 64$, 3-stage curriculum, 5 seeds): PR fails at

Table 10: P9 (broad): PR mean over 5 seeds (success rate in parentheses). PR fails at $\mathrm{rank}(W) \in \{2, 8\}$ and becomes reachable at 16 $(4/5)$, consistent with the estimated $T\theta \approx 15.4$ readout-rank floor.

$\mathrm{rank}(W)$	2	8	16	full
PR (broad)	$0.02_{(0/5)}$	$0.11_{(0/5)}$	$\mathbf{0.45}_{(4/5)}$	$0.44_{(3/5)}$

$\mathrm{rank} = 8$ $(0/5)$ and becomes reachable at $\mathrm{rank} = 16$ $(4/5)$, consistent with the $T\theta \approx 15.4$ floor: with the readout pinned below $T\theta$, broad is unrepresentable ($\mathrm{rank}(M) \le \mathrm{rank}(W) < T\theta$), even though the state and curriculum are unchanged. Even at full rank, $2/5$ seeds fail—the accessibility limit documented in Section 6. (Selective is content-addressed; its readout must span the vocab output, so the readout-rank knob isolates the broad floor; the selective floor is isolated causally in Section 6 instead.)

F CURRICULUM DYNAMICS: SPECIALIZATION, NOT SPEEDUP

The curriculum's effect is not a generic training speedup. Tracking per-channel $|A|$-spread across the $T = 64$ phase (Figure 5), the curriculum condition's spread rises monotonically ($0.30 \to 0.42$) as PR climbs to ~ 0.7; the no-curriculum condition (identical LR and initialization) keeps spread flat (≈ 0.30) and PR at random. The two diverge *during* training: curriculum drives the multi-timescale specialization broad retrieval requires (Proposition 4), which standard training does not produce on its own (the objective-and-landscape account of §3). The effect is robust to the schedule ($[16, 32, 64]$ and $[8, 16, 32, 64]$ both reach broad), so the operative mechanism is the specialization dynamics rather than schedule granularity.

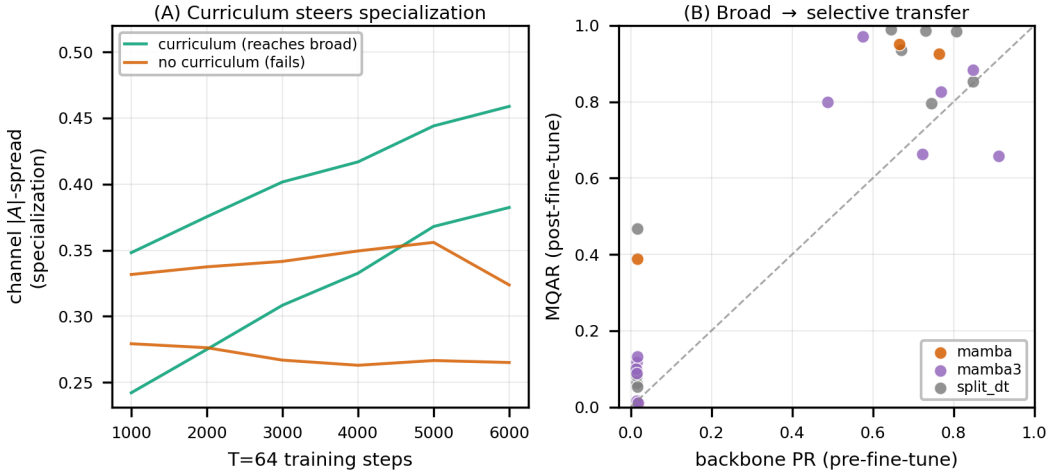

Figure 5: The two mechanism views of the causal intervention. (A) Channel $|A|$-spread (specialization) vs. $T = 64$ training step: curriculum traces drive spread up and reach broad (PR ~ 0.7); no-curriculum traces (same LR/init) stay flat and fail (PR ~ 0.01); two seeds. (B) Backbone PR vs. post-fine-tune MQAR, 36 backbones (three architectures; curriculum/no-curriculum). Pearson $r = +0.93$.

G SCALE INVESTIGATION: DIAGNOSIS AND THE WIDTH-CORRECTED RECIPE

This appendix documents why our minimal SSM implementations fail under the *naive* toy protocol at width, and the two ingredients that fix it (stated at the end; scale results in Appendix H). The main causal result (Section 6) uses Mamba-3 at toy scale ($d_{\text{model}} = 128$), where it trains well; the question here is whether the toy training protocol scales. Under the naive protocol it does not, for any of the three variants. We first took the decoupled `split_dt` block to $d_{\text{model}} = 512$ (24M): the $d_{\text{model}} = 512$ width barrier was diagnosed by bisection (only d_{model}/n_{layer} break training, not d_{state}) and by a forward pass showing layer-output std ≈ 5.6 vs. ≈ 0.67 at $d_{\text{model}} = 128$. Eight fixes were tried with a correctly-sized data loader: LR $\{10^{-3}, 3 \times 10^{-3}, 10^{-2}\}$ (10^{-2} diverges, others plateau), optimizer ($\beta_2 = 0.95, \text{wd} = 0.1$), init std-scaling, mimetic init ($dt \to \Delta \approx 1$, $W_C^\top W_B \approx I$), output norm, internal SSM-input norm, and the full norm chain; all leave eval accuracy at random (0.008). The failure is not specific to the decoupling: the coupled Mamba and the separate Mamba-3 implementation plateau identically at $d_{\text{model}} \geq 256$ and explode at width (per-layer output std ≈ 25 and ≈ 19 at $d_{\text{model}} = 512$, versus ≈ 3 and ≈ 2.5 at $d_{\text{model}} = 128$), so decoupling is not the cause; a width-scaled initialization is the missing ingredient.

The fix. The two ingredients that recover training at width are (i) a *width-scaled residual initialization*: out_proj and the MLP down-projection $\sim \mathcal{N}(0, 0.02/\sqrt{L \cdot d/128})$ plus a less-retentive $A_{\log} \sim U(-1, 0)$, which holds the layer-output std near 3 even at $d = 1024$ (versus ≈ 10 for depth-only scaling); and (ii) a *width-scaled learning rate* $\propto 1/d$ ($3{\times}10^{-4}$ at $d = 512$, 10^{-4} at $d = 1024$; the working window narrows with width, and $2{\times}10^{-4}$ at $d = 1024$ already plateaus). With both, Mamba-3 converges at 24M and 122M (Appendix H); the earlier failures were this protocol mismatch rather than an architectural or decoupling limit.

H SCALE: THE BROAD CEILING RISES WITH WIDTH, NOT STATE DIMENSION

The Pareto frontier of Section 6.1 is at toy scale ($d_{\text{model}} = 128$). We replicate it at 24M ($d_{\text{model}} = 512$, $n_{\text{layer}} = 6$, $d_{\text{state}} = 64$, 2 seeds) and ask which dimension of scale lifts the broad ceiling.

Scale-invariance of the objective claim. At $\lambda = 0$, broad is at the random floor for *both* architectures at 24M as well (Transformer 0.016, Mamba-3 0.014). "Language modeling does not teach broad" is a property of the objective at these scales, surviving a $4\times$ width and $3\times$ depth increase.

The Mamba-3 ceiling rises with scale. At 24M Mamba-3 reaches broad 0.71 ($\lambda = 0.3$, PPL 4.5), more than double its toy ceiling (0.33). The architecture gap narrows (toy 0.89 vs 0.33; 24M 0.96 vs 0.71 at matched $\lambda = 3$) but Transformer still leads. At 24M Mamba-3 is non-monotonic in λ: 0.71 at $\lambda = 0.3$ falls to 0.59 at $\lambda = 1, 3$ with high seed variance: more broad pressure destabilizes the larger model; we report this rather than cherry-pick the peak.

Width lifts the ceiling; d_{state} has a soft covering effect. To separate "capacity" (representation) from "accessibility" (optimization) we sweep each dimension independently (Figure 6). At fixed $d_{\text{model}} = 128$, we extend the d_{state} sweep *below* the $T\theta \approx 15.4$ floor to $d_{\text{state}} \in \{2, 4, 8, 16, 32, 64, 128\}$ (3 seeds, curriculum, $\lambda = 3$). Broad is not abolished below the floor: it rises on a soft, saturating curve from 0.19 ± 0.08 ($d_{\text{state}} = 2$, $\sim 12\times$ random) to a ~ 0.35–0.47 plateau by $d_{\text{state}} \approx 8$–$16$, with no sharp $T\theta$ threshold. So d_{state} is a per-channel *covering* factor (poles per channel, governing how uniformly channels address positions), not the rank budget—and the linearized model makes "covering" precise: optimally placed banks are log-spaced timescale clocks covering age k with pole $\tau \approx k$ at efficiency ≈ 0.6, with a saturating ceiling (Appendix B.6); consistent with the readout rank being ample even at $d_{\text{state}} = 2$ ($\text{rank}(W) \gg T\theta$, Appendix B.8). The sharp feasibility floor instead lies on the readout rank: the P9 ablation (§5d) collapses broad below $\text{rank}(W) \sim T\theta$ even at full d_{state}. At fixed $d_{\text{state}} = 64$ and $n_{\text{layer}} = 2$, broad rises with d_{model} up to 256 ($0.15 \to 0.38 \to 0.60$ across $\{64, 128, 256\}$, 5 seeds) and then *plateaus* at 512 (0.38 ± 0.11, 7 seeds): pure width lifts the ceiling over a $4\times$ range and saturates within fixed depth, while the 24M model (6 layers at $d = 512$) reaches 0.71 — width and depth contribute jointly, and neither alone accounts for the scale gain. Together with the curriculum intervention (Section 6, which reaches broad in the *same* function class standard SGD misses), the width dependence is consistent with

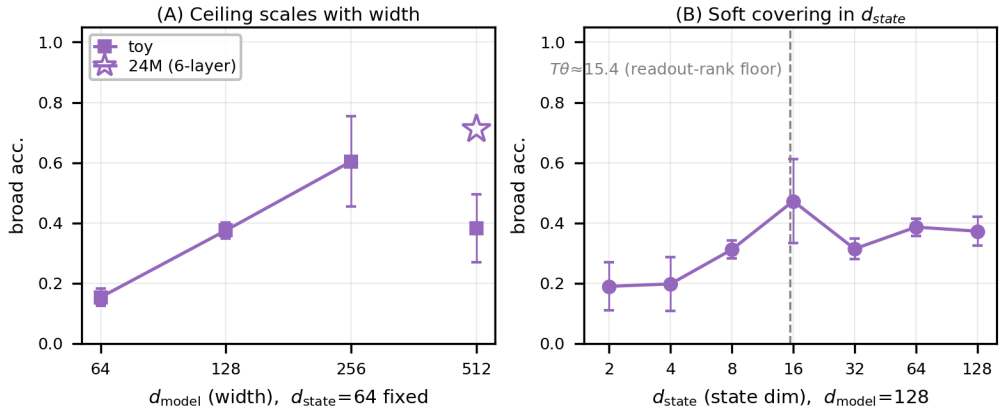

Figure 6: Capacity vs accessibility (mamba3, broad at the ceiling). (A) Broad accuracy vs d_{model} (width) at fixed 2 layers: $0.15 \to 0.38 \to 0.60 \to 0.38$ across $\{64, 128, 256, 512\}$ (512: 7 seeds, ± 0.11) — width lifts the ceiling and saturates within fixed depth; the 6-layer 24M model at $d = 512$ reaches 0.71 (width and depth jointly). The ceiling is consistent with an accessibility (optimization) limit. (B) Broad vs d_{state} (3 seeds, curriculum, $\lambda = 3$), extended below the $T\theta \approx 15.4$ reference to $d_{\text{state}} \in \{2, 4, 8, 16, 32, 64, 128\}$: broad rises on a soft, saturating curve from 0.19 ($d_{\text{state}} = 2$, $\sim 12\times$ random) to a ~ 0.35–0.47 plateau by $d_{\text{state}} \approx 8$–16, with no sharp $T\theta$ threshold. d_{state} is a per-channel covering factor, not the rank budget (the sharp floor lies on the readout rank, P9). The width dependence, together with the curriculum (Section 6), is consistent with broad being accessibility-limited (optimization).

a widening *accessible basin* under optimization rather than a d_{state} effect. The horizon factorial (Appendix I) extends this to the T axis: under bootstrap-safe curricula, scaling T at fixed (d_m, d_s) yields no consistent d_{state} effect and a positive width effect at every T — consistent with the same accessibility picture, with the realized coverage window W (not a static rank) setting which T is reachable.

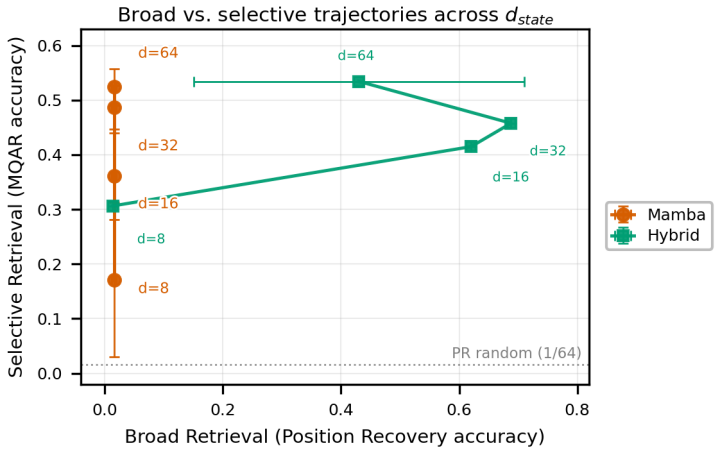

Figure 7: Broad vs. selective trajectories vs d_{state} (P2). Mamba: broad stays random while selective grows with d_{state} (near-vertical stack at the random floor). Hybrid (SSM+attn): both grow.

I THE BROAD HORIZON: WHAT SETS THE RETRIEVAL-RANK BUDGET?

The two-rank decomposition (§3) leaves open which side of $\min(r_{\text{mem}}, r_{\text{acc}})$ sets the broad horizon $T^\star$. We adjudicate with a factorial $T \in \{64, 128, 256\} \times (d_{\text{model}}, d_{\text{state}}) \in \{(128, 16), (128, 64), (256, 16), (256, 64)\}$ on broad PR (mamba3, T-curriculum, 2 seeds; a no-

curriculum leg on the two extreme columns), reading the collapse profile in T. Three competing accounts: H_{state} ($T^\star$ tracks $d_{\text{state}}/r_{\text{mem}}$), H_{width} ($T^\star$ tracks $d_{\text{model}}/r_{\text{acc}}$), H_{opt} (the observed $T^\star$ stays well below the representational bound; curriculum shifts it). *Staging audit.* Our first pass at this factorial used schedules starting at $T/4$ ($[32,\dots]/[64,\dots]$), which is beyond mamba3's from-scratch bootstrap limit: first-stage accuracies stayed at random (0.035 at $T=32$, $d_s=16$; 0.018 at $T=64$, $d_s=64$), so $12/36$ cells (all $T=128$ at $d_s=16$; all $T=256$) were *staging-confounded* and were re-run with bootstrap-safe curricula ($[16,32,64,128]/[16,32,64,128,256]$; `sweep_broad_horizon_fix.py`). Table 11 reports the corrected factorial. The audit overturned the first pass's verdict, and is itself a finding about the from-scratch bootstrap limit (below).

H_{opt} wins the horizon: neither rank axis alone sets it. Four readings (Figure 8). (i) *No d_{state}*

Table 11: Broad PR vs T (corrected factorial: curriculum with bootstrap-safe stages, mean over 2 seeds). $T=64$ cells use the original (valid) $[16,32,64]$ schedules; $T=128/256$ cells are the corrected re-runs. No consistent d_{state} effect at any T (within 2-seed noise, d_s = 16 matches or slightly exceeds $d_s=64$); width is positive at every T. $T=256$ is partial coverage at all configs (30–90× random $1/256$). The no-curriculum leg fails everywhere (≤ 0.02, all configs, all T).

config	$T=64$	$T=128$	$T=256$
$d_m=128$, $d_s=16$	0.76	0.63	0.12
$d_m=256$, $d_s=16$	0.88	0.79	0.19
$d_m=128$, $d_s=64$	0.85	0.45	0.16
$d_m=256$, $d_s=64$	0.81	0.68	0.27

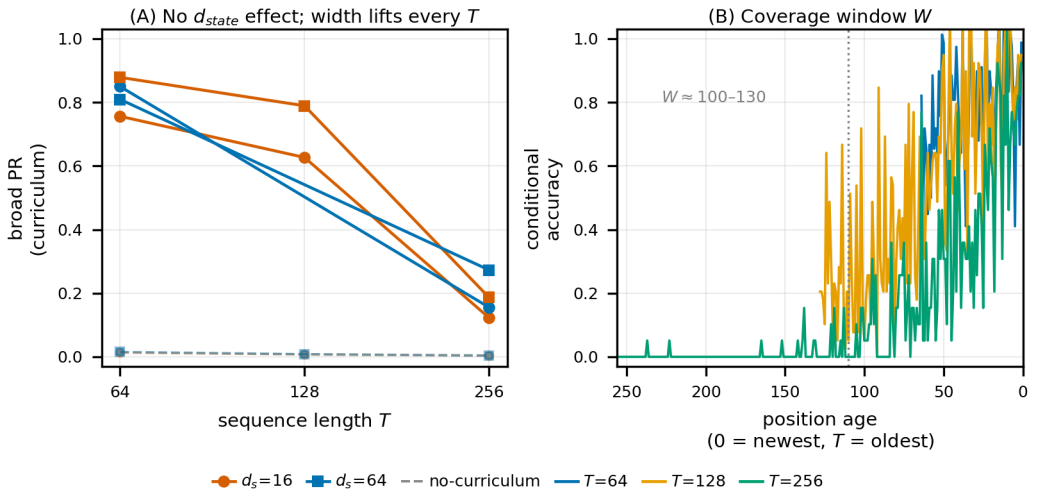

Figure 8: The broad horizon (mamba3, corrected factorial; Table 11). (A) Broad PR vs T for the factorial configs (color = d_{state}, marker = d_{model}; dashed = no-curriculum, at chance everywhere). No consistent d_{state} effect; width positive at every T. (B) Per-position conditional accuracy vs position age (seed 0, $d_m=128$, $d_s=16$) at $T\in\{64,128,256\}$: a finite coverage window $W\approx 100$–130 with a ~ 50-wide ramp. Ages beyond W are *exactly* zero (hard truncation), full T coverage at $T=64$, and $T=256$ shows the window with dilution.

effect under curriculum: at every T, $d_s=16$ matches $d_s=64$ within 2-seed noise ($T=128$ even inverts: 0.63 vs 0.45 at $d_m=128$) — the first-pass "memory rank sets the horizon" reading was an artifact of the staging confound, and H_{state} is rejected in the curriculum regime. (ii) *Width is positive at every T* (mean $+0.04/+0.20/+0.09$ at $T=64/128/256$): a ceiling lift, not an unlock — no-curriculum training stays at chance at every width. (iii) *The mechanism is a realized coverage window.* The per-position profiles (Fig. 8B) show a hard truncation: positions older than $W\approx 100$–130 are recovered at exactly chance while recent ages reach 0.7–0.9, with a ~ 50-wide ramp between. Broad PR for $T>W$ is then $\approx W/T$ times the in-window rate — the "soft horizon" is a hard window diluted by $1/T$, which reproduces the graded decline of Table 11 ($T=256$ values

0.12–0.27 are precisely the dilution regime). W is a learned/realized coverage property strongly controlled by optimization: the trained model's per-head decays do not broaden with T, and the window a run achieves — not d_{state} or d_{model} — is the operative budget. (We state no formal bound linking rank to positional span; memory-side resources plausibly cap W, but this is a conjecture.) (iv) *Curriculum is necessary everywhere*: the no-curriculum leg fails at every configuration (≤ 0.02, including $T = 64$). *The bootstrap limit that caused the confound is itself the one clean d_{state} effect.* From-scratch training bootstraps only up to T_{boot} ($d_s = 16$ bootstraps at $T = 16$ but not $T = 32$; $d_s = 64$ at $T = 32$ but not $T = 64$; roughly $(0.5{-}1)\, d_s$ over these two levels — an association, not a fitted scaling law), which curriculum extends $\geq 4\times$ ($d_s = 16$: $\sim 16 \nrightarrow \geq 64$). T_{boot} is strongly associated with d_{state} but not with width in our tested configurations: at $T = 32$ no-curriculum, doubling width at fixed $d_s = 16$ ($d_m = 128 \to 256$, doubling the pole count to 8 heads) leaves all seeds dead ($0.030/0.030/0.029$, indistinguishable from $d_m = 128$), while quadrupling d_s at fixed width ($d_m = 128$, $d_s = 16 \to 64$) solves ($0.50/0.61/0.18$; 3 seeds) — the from-scratch barrier reads out the per-head state dimension, not the architecture's total pole count. Whether d_{state} enters T_{boot} through the per-head spectral structure or the dt-modulation channel is open.

J FROZEN $\lambda = 0$ BACKBONE PROBE

The $\lambda = 0$ point of the Pareto frontier (Section 6.1) could in principle reflect an untrained broad head rather than absent broad information in the representation. We rule this out. We train a pure-LM ($\lambda = 0$) backbone (Transformer and Mamba-3, $d_{\text{model}} = 128$, $T = 64$, 6k steps on char-level wikitext), *freeze* it, and fit a fresh linear probe ($d \to T$) and a 2-layer MLP probe ($d \to 4d \to T$) to decode the Position-Recovery answer from the frozen query-position hidden state. Inputs are standard PR (random unique tokens, vocab 256, fed through the frozen char-trained embedding) so the probe tests position encoding free of semantic confound. Both probes are at chance: Transformer linear 0.012 / MLP 0.020; Mamba-3 linear 0.023 / MLP 0.027 (random $1/64 = 0.016$). These inputs (random unique tokens) are out-of-distribution for a char-wikitext LM, so a reader could worry the chance result reflects distribution shift rather than absent broad information. We rule this out with an *in-distribution* control: the broad query is the real-text position-recovery task itself—a unique 3-gram sampled from a real wikitext window (§6.1), the LM's own training distribution—fed through the same frozen $\lambda = 0$ backbone, decoding the 3-gram's start position with a fresh linear/MLP head. Both architectures remain at chance: Transformer linear 0.017 / MLP 0.016; Mamba-3 linear 0.016 / MLP 0.019 (random 0.016). On-distribution too, the $\lambda = 0$ representation carries no linearly- or shallow-nonlinearly-decodable broad position information (`scripts/probe_rtpr_frozen.py`). The frontier's $\lambda = 0$ point is thus a statement about the representation PPL induces, not about the untrained head or an OOD artifact (`scripts/probe_ppl_backbone.py` for the OOD control).

The same absence at 2.7B, both architecture families. We repeat the frozen probe on two open pretrained checkpoints—Mamba-2 2.7B (Dao & Gu, 2024) and Pythia-2.8B (Biderman et al., 2023), no finetuning—with the same two input distributions (random unique subwords; in-distribution real-text 3-gram queries with a uniqueness-verified target) and $n = 3000$ sequences per cell. Both models are at chance under the MLP probe on the in-distribution cell (each $0.015 \approx 1/64$) and at 1.4–1.8$\times$ random under the linear probe, with no per-position structure (no sustained age band above 3$\times$ random). Three controls rule out probe artifacts: window-position hiddens do decode their own token identity (2.6–2.9$\times$ random at position 0), pooled and per-position variants are also at chance, and a strongest-case probe—a linear head over the full flattened $65 \times d$ hidden block, training accuracy 1.0 on both models—tests at 1.3–1.6$\times$ random. The absence is not a readout limitation: pretrained representations of *both* families carry no static broad position binding, consistent with the $\lambda = 0$ frontier point and with broad readout being something training must build (Transformer solves PR only through a trained head, §5a) (`scripts/sota_pr_probe.py`; pre-registered decision rules in `experiments/sota_probe_preregistration.md`).

K EFFECTIVE DEPENDENCY WIDTH OF NEXT-TOKEN PREDICTION

The Pareto frontier (Section 6.1) shows pure language modeling ($\lambda = 0$) leaves broad retrieval at the random floor. We verify the premise directly: the predictive signal of next-token prediction is concentrated on few positions, so PPL is a selective objective that does not exercise the broad axis.

Two measurements on trained char-level wikitext LMs (Transformer and Mamba-3, $d_{\text{model}} = 128$, $T = 64$, 6k steps, 3 seeds each; `scripts/measure_keff.py`; Figure 9). **(A) Truncation curve.** We give the LM only the last K tokens before a target and measure the next-token NLL, sweeping K. NLL saturates at $K \approx 8$ for both architectures (Transformer: 2.54 at $K = 1$ to 1.46 at $K = 8$, flat thereafter; Mamba-3: 2.54 to 1.43; SD < 0.03 over seeds): predictions stop improving once $\sim$ 8 recent tokens are visible. **(B) Leave-one-out influence.** For a target y_t we mask each history position k in turn and measure $\Delta\text{NLL}_{t,k}$; the effective number of predictive positions is $K_{\text{eff}}(t) = \exp(H(q_t))$ over the normalized influence q_t. Median K_{eff} is 3.6 (Transformer, per-seed 3.4–3.8) and 4.8 (Mamba-3, per-seed 4.2–5.3) over 3 seeds.

Both put the effective dependency width of next-token prediction at $K_{\text{eff}} \approx 4$–$8 \ll T = 64$ for this char-level wikitext setup. On this distribution the retrieval demand of language modeling is predominantly selective ($\sim K_{\text{eff}}\theta$, the easy axis of Proposition 1) rather than broad ($\sim T\theta$): the typical prediction depends on few positions, which is why the broad axis is unreachable under PPL on the tasks we test (Section 6.1). We do not claim PPL is universally selective; one can construct autoregressive distributions whose next token depends on the full history, and we claim only that it holds for the natural text and context lengths studied here.

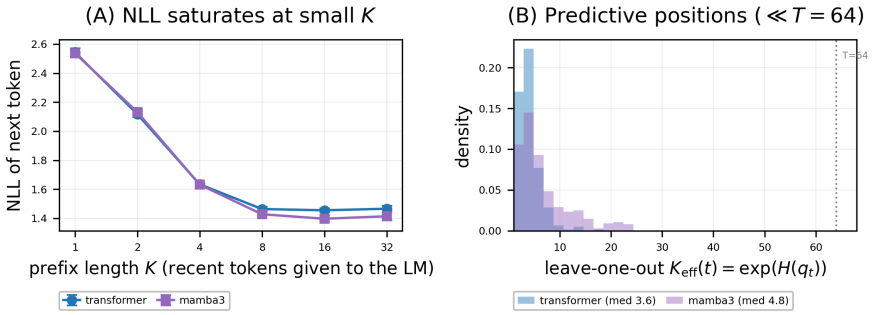

Figure 9: Effective dependency width of next-token prediction (char-level wikitext, Transformer and Mamba-3). (A) Truncation: NLL of the next token given only the last K tokens saturates at $K \approx 8$ for both architectures ($T = 64$). (B) Leave-one-out $K_{\text{eff}}(t) = \exp(H(q_t))$, the effective number of predictive positions, has medians 3.6 (Transformer) and 4.8 (Mamba-3), far below $T = 64$. Standard language modeling is a selective objective.

L POSITION-ANSWER DECODABILITY PROBE

We test the representation side of the §7 prediction (position-indexed information tracks PR) with a readout probe on the frozen PR-trained checkpoints. For each checkpoint we read the trained-readout hidden state $h_{\text{query}} = \text{norm}(\text{backbone}(\text{seq}))[:, \text{seq_len}+1, :]$ on 6000 fresh PR inputs (random unique tokens, query at seq_len+1), then fit a *fresh* linear probe and a 2-layer MLP probe ($d \to 4d \to T$) to decode the PR answer (which content position matches the query). Probes are trained on 66%, model-selected on 17%, tested on 17%; random baseline $1/T = 0.016$. (An earlier content-at-position variant, decoding the token at each position k, returned $\sim$random for all architectures including Transformer (PR $= 0.999$): a fixed-d state cannot hold $T \gg d$ distinct content tokens, which is the paper's own capacity argument; PR is solved by readout-time binding, not by pre-storing all content. The position-answer probe is the correct measurement.)

The probe tracks PR almost exactly (Figure 10): Transformer (0.999) and RetNet (0.978) decode to 0.97–1.0; Hybrid (0.62) and the partial RetNet seed (0.31) decode proportionally; every PR-failed

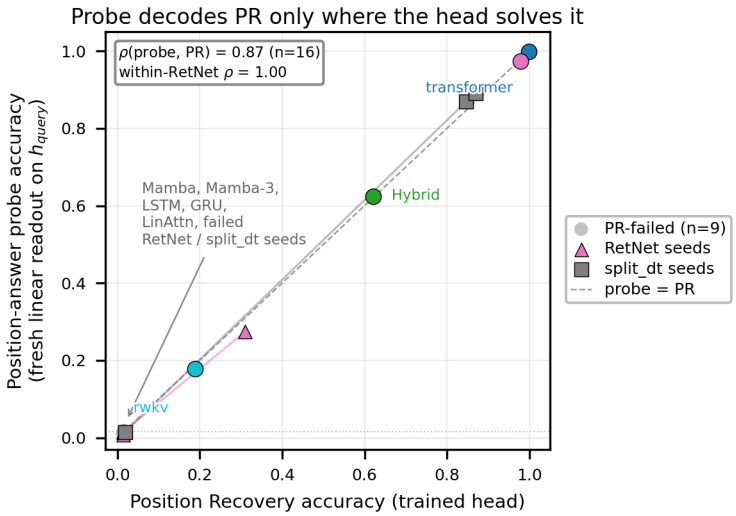

Figure 10: Fresh-linear-readout probe accuracy vs trained-head PR, 16 checkpoints. Cross-architecture points (circles) and within-architecture controls: RetNet bimodal seeds (triangles) and Mamba `split_dt` broad-vs-failed seeds (squares). Points lie on the diagonal (probe=PR); Spearman $\rho = 0.87$ overall, 1.0 within each family.

checkpoint (Mamba, Mamba-3, RWKV-low, LSTM, GRU, Linear Attention, and the failed RetNet and `split_dt` seeds) stays at chance. Spearman(probe, PR) is 0.83 cross-architecture (over the $n = 9$ architecture means, $p < 0.01$; 0.87 over all 16 checkpoints, Figure 10), 1.0 within the five RetNet seeds, and 1.0 within Mamba `split_dt`. The within-RetNet control rules out an absolute-position-embedding explanation: failed RetNet seeds *have* a positional embedding yet their answer is not decodable, so the deficit is the learned content–position binding, not the absence of positional input.

The MLP probe never exceeds the linear one (mean gap -0.01): a richer two-layer readout finds no decodable binding beyond the linear one. The broad-PR deficit is thus not recoverable by either a linear or our two-layer nonlinear probe—consistent with a representational (not readout) limitation, though we cannot rule out decodability by a richer probe family. Finally, decodability dissociates from language modeling: RetNet (probe 0.97, Markov PPL 44.9) vs Mamba-3 (probe 0.01, PPL 9.1; a separate training run from Table 1's 8.2).

M Subject-verb agreement: a behavioral position-indexed task

We test the *selective* side of the broad/selective split (Proposition 1) on a real linguistic task: subject-verb number agreement (Figure 11), where the verb must agree with a subject separated by intervening "attractor" nouns of the opposite number (e.g. "the key near the doors is" vs "are"). Agreement is position-indexed (the verb depends on the subject's number across distance) but *selective* (the state need only retain one feature, not all positions), so by Proposition 1 it is the easy axis ($d \gtrsim \theta$) all models should handle, like the selective retrieval (MQAR) Mamba already solves (§5). We train word-level wikitext-103 LMs (≈31M, $d_{\text{model}} = 512$, 6 layers, 10k-word vocab, full-sequence loss) for Transformer, RetNet, Mamba-3 to matched PPL (33–35), and evaluate templated agreement pairs (distance 1–10, 0–3 attractors), comparing the LM logit on the agreeing vs disagreeing verb.

Result. At distance 1 all three architectures agree almost perfectly (≥ 0.97); with distance all decay toward chance together, with overlapping error bars (0.35 ± 0.07 / 0.41 ± 0.08 / 0.47 ± 0.05 for Mamba-3 / Transformer / RetNet at $d \approx 10$). Mamba-3 (the best LM, PPL 33.2 ± 0.2) is statistically comparable to the broad-position models on agreement—it does *not* show the broad-retrieval deficit. This is the broad/selective prediction: agreement is selective (one feature across distance, like MQAR), the easy axis all models solve; the broad deficit is specific to tasks that require the state to carry *all* positions

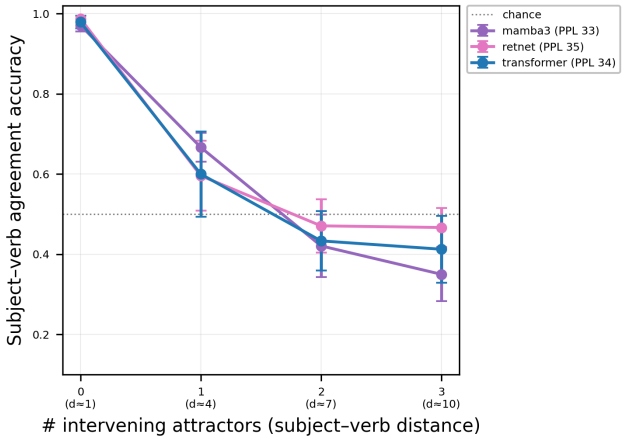

Figure 11: Subject-verb agreement accuracy vs subject-verb distance (intervening attractors), word-level wikitext LMs at matched PPL (33–35); 3 seeds, error bars ± 1 SD. All architectures agree almost perfectly at distance 1 and decay toward chance *together* at long distance with overlapping error bars (0.35–0.47 at $d \approx 10$); Mamba-3 (lowest PPL, 33.2) is comparable to the others. Agreement is selective (one feature across distance), so, as the broad/selective theory predicts, all models handle it; the broad deficit does not appear here.

(PR; the broad readout of Appendix L), which natural selective/local language tasks do not demand. The negative result is itself informative—it localizes the deficit to genuinely broad tasks.

N Reproducibility

Code, training scripts, and per-seed result JSONs are provided. Key entry points, grouped by paper section:

- **Tasks and backbones:** `train_mqar.py` (MQAR, `--num_kv` for the K-sweep); `train_curriculum.py` (PR + causal-intervention backbones); `causal_intervention_multiseed.py` (fine-tune + transfer).
- **Theory verification:** `verify_specialization_complexity.py` (Propositions 2 and 5); `analyze_linearized_rank.py` (linearized rank on checkpoints); `verify_minrank_bias.py` (matrix-sensing illustration). Theory-upgrade checks (Lemmas 1 –2, Appendices B.6–B.7): `theory_upgrade/verify_two_pole.py`, `verify_timescale_matching.py`, `covering_rate2.py`, `geom_clock.py`, `escape_dynamics.py`, `directionB_envelopes.py`.
- **Empirical sections:** `analyze_alpha_specialization.py` (specialization, §5c); `experiment_minrank_predictions.py` (P9, §5d); `sweep_mqar_K_T.py` + `sweep_cheap_validations.py` (T-controlled cliff + 8-seed edge, §5b); `sweep_broad_horizon.py` / `_fix.py` and `plot_broad_horizon.py` (App. I); `capacity_scaling.py` (Fig 5 sweeps); `probe_rtpr_frozen.py` and `sota_pr_probe.py` (App. J frozen probes, incl. the 2.7B controls).
- **Batch drivers:** `sweep_mqar_K.sh`, `sweep_dstate_curriculum.sh`. The 122M numbers are from a fully re-run batch (original per-seed artifacts lost; every paper number reproduced): `scale_122m_pr.sh` / `_curriculum.sh` / `_transfer.sh`, `mqar_122m_seed.py`, `scale_b3_width.sh`.