THE VALUE OF MEMORY IN LANGUAGE MODELS: JOINT LIMITS ON LANGUAGE MODEL EXPRESSIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

The value of memory in a language model depends on the model that must use it. Prompt information matters only if it is both retained and usable by the model's fixed parameters. We study this interaction through prompt accessibility, asking which outputs a frozen model can be prompted to generate. We place a causal cut between prompt processing and generation, so all prompt-derived information available afterward must pass through the retained state. This gives a joint bound in which prompt budget, retained memory, and compatible model-side constraints all limit the same accessible outputs, with the tightest applicable constraint governing. We then sharpen the memory-side bound under additional assumptions on the update dynamics and reachable-state geometry, and show how mechanism-specific structure can yield tighter bounds, using Gated Linear Attention (GLA) as a concrete example. Controlled experiments on two frozen GLA checkpoints show that increasing retained memory moves the discovered output frontier. On GLA-1.3B, increasing the retained payload from 12 to 24 to 36 MiB raises the evaluated 50% frontier from 8192 to 10240 to 12288 tokens. Controls further show that memory usefulness depends on both prompt budget and what information is retained. A same-weight sliding-window comparison reproduces the qualitative memory-frontier response, while Mamba provides complementary evidence from fixed-size recurrent memory. Together, these results support treating prompt budget, memory, and the learned model as coupled resources, motivating memory and architecture to be matched rather than scaled independently.

1 INTRODUCTION

To use a long prompt, a model must retain information from it during generation. Full-attention transformers do this with a growing key–value (KV) cache, which consumes HBM and memory bandwidth and can directly constrain context length, batch size, and throughput. These costs have motivated a wide range of alternative memory designs. Sliding-window attention retains only a recent subset of KV pairs, while linear attention, state-space models, and test-time-learning mechanisms compress or update past information through recurrent states or adapted parameters (Katharopoulos et al., 2020; Gu & Dao, 2024; Sun et al., 2025; Behrouz et al., 2025). Despite their differences, these mechanisms address the same basic design problem: what information from the prompt should remain available for later generation?

Accessibility provides a natural way to study what a fixed model can express. An output is accessible if there exists a prompt that makes the fixed model generate it. Meyer et al. (2026) use this notion to study how many different outputs a frozen transformer can produce as the prompt varies, showing that accessibility is limited by the prompt budget and, under stronger assumptions, by a model-side ceiling that does not grow with prompt length. They also probe accessibility empirically by optimizing continuous prompts for specific targets. We extend this analysis by treating retained memory as an additional bottleneck and scaling axis. After prompt processing, all influence of the prompt on generation must pass through the prompt-dependent state that remains, which we make explicit by placing a *causal cut* between prompt processing and generation. The retained state must therefore be complete, containing all prompt-derived information available afterward.

Our central design principle is that *memory and architecture should be matched, not independently maximized*. The retained state determines what prompt-dependent information remains available, while the model's fixed architecture and learned parameters determine how that information is writ-

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.

ten and read. A useful analogy is a student preparing for a test from a library. Before the test, they record what they need in a notebook, and during the test they must rely on that notebook together with their prior skills. The library represents the available context, the notebook the retained state, and the student's fixed abilities the model. A richer library helps only if useful information reaches the notebook, while a larger notebook helps only if the student can use it effectively. Because retained memory also carries a direct inference cost, the relevant design question is how much memory is useful for a given fixed model. This may be especially important for smaller models, where additional memory may offer little benefit once model-side constraints become limiting.

Theorem 4.1 formalizes this principle by adding retained memory to existing prompt- and model-side accessibility constraints. Each bounds the same set of robustly accessible outputs, so the smallest applicable term governs the result: relaxing one resource cannot improve the joint bound while another remains smaller. We then sharpen the memory term under assumptions of stable updates or geometric structure in the reachable states, and specialize the analysis to Gated Linear Attention (GLA), where rank-one writes and, under contractive retention, an effective memory horizon provide additional structure.

The experiments examine the same relationship in pretrained models. With model weights fixed, we optimize a prompt for each target and require the full target to be generated from the retained state after the prompt is removed. Increasing retained memory moves the discovered frontier on both GLA-340M and GLA-1.3B. Prompt-budget controls show that the usefulness of a fixed memory channel depends on the prompt budget, while equal-size channels can support very different outputs depending on what they retain. A same-weight Qwen sliding-window comparison shows a similar frontier response as the attention window grows. Mamba provides complementary evidence from fixed-size recurrent memory, where larger prompt budgets yield diminishing frontier gains. Together, these results show that prompt budget, memory, and the learned model jointly shape which outputs can be reached.

Our contributions are:

- We introduce a causal-cut extension of prompt accessibility that adds retained memory as a distinct bottleneck and combines it with existing prompt- and model-side constraints.
- We derive memory bounds under assumptions on stable update dynamics and reachable-state geometry, including a mechanism-specific specialization to Gated Linear Attention.
- We establish controlled memory frontiers on two GLA checkpoints, with prompt-budget and representation controls and a supporting same-weight sliding-window comparison.

2 Related work

Prompt accessibility and memory capacity. Prior work studies what frozen transformers can produce through prompting, and how this depends on the model architecture and prompt length (Wang et al., 2023; Petrov et al., 2024a;b; Hu et al., 2025). Meyer et al. (2026) formalize this idea as prompt accessibility, while related empirical work tests how much control optimized prompts or activations can achieve over exact outputs (Kuratov et al., 2025; Fort, 2025). A separate body of work asks how much information different memory mechanisms can retain. BASED studies the tradeoff between state size and associative recall (Arora et al., 2024), other analyses show limits of bounded recurrent memory on copying and retrieval (Jelassi et al., 2024; Wen et al., 2025; Bhattamishra et al., 2024), and Effective State-Size measures how much of a model's memory is actually used (Parnichkun et al., 2025). We connect these two lines of work by asking a simple question: once the prompt has been processed, how does the information that remains in the model's retained state limit the outputs that can still be reached?

Fixed-state and associative memory. A growing family of sequence models replaces or supplements the transformer's expanding KV cache with recurrent or associative memory, including Mamba, GLA, TTT, and Titans (Schlag et al., 2021; Sun et al., 2023; Gu & Dao, 2024; Yang et al., 2024; Sun et al., 2025; Yang et al., 2025b; Behrouz et al., 2025). Their strong benchmark performance has made bounded or compressed memory a practical alternative to storing the full context, and MIRAS provides a common view of how these memories are written and updated (Behrouz et al., 2026). Strong benchmark results, however, are often taken as evidence that fixed memory is sufficient in practice, even though such benchmarks do not directly reveal when the retained state

itself becomes the limiting resource. We make this limitation precise: at a fixed robustness level, a bounded finite-dimensional retained state can support only a bounded number of distinguishable continuations, no matter how long the prompt becomes. The growing use of hybrid designs such as Jamba and Kimi Linear further suggests that early benchmark success alone did not settle whether fully fixed memory remains sufficient as context demands grow (Lieber et al., 2024; Kimi Team et al., 2025).

State geometry and stability. The capacity of a fixed memory depends not only on its nominal size, but also on which states its dynamics can actually reach and how long information persists. Prior work on linear attention finds low effective rank in recurrent states and uses this structure for state reduction and head pruning (Nazari & Rusch, 2026; Sun et al., 2026). Meyer et al. (2026) similarly use embedding-space geometry to sharpen accessibility bounds. We use the same geometric perspective for the complete reachable-state set, asking how many post-prompt states can remain distinguishable under an enclosure that covers every admissible state. We also draw on classical stability ideas (Boyd & Chua, 1985; Sontag, 1995) to analyze contractive memory dynamics and the resulting effective memory horizon. Combined with standard packing arguments (Vershynin, 2018), these properties give sharper bounds on how much a fixed memory can support.

3 Causal-cut memory and accessible outputs

To study retained memory and the fixed model together, we separate prompt-dependent information carried into generation from the fixed mechanisms that operate on it. We place a *causal cut* between prompt processing and generation, so all prompt-dependent information available afterward must pass through the retained state.

Controlled setting. We fix the model, prompt budget, and generation protocol. A prompt contains m vectors in $\mathbb{R}^d$, with admissible prompts $\mathcal{P}_m \subseteq \mathcal{U}_m \subseteq \mathbb{R}^{dm}$, where $\mathcal{U}_m$ is the set of continuous prompts on which the model is defined. Only the target continuation and the prompt optimized for it vary. Under deterministic decoding, each retained state produces a single output sequence.

Definition 3.1 (Complete causal-cut memory)**.** The prompt-processing map $E_{\theta,m}$ sends an admissible prompt p to the complete post-prompt state $M = E_{\theta,m}(p)$. This state contains every prompt-dependent quantity that subsequent computation can inspect, and no additional prompt-dependent input is available after the cut. The reachable-state set is $\mathcal{M}_m = E_{\theta,m}(\mathcal{P}_m)$.

Accessible continuations. Fix a continuation length n and target alphabet $\mathcal{V}_0 \subseteq \mathcal{V}$ of size $K \geq 2$. The fixed readout $G_{\theta,n}$ produces an n-token output from the retained state using only its own preceding outputs, so

$$p \xrightarrow{E_{\theta,m}} M \xrightarrow{G_{\theta,n}} y, \qquad \mathcal{M}_m = E_{\theta,m}(\mathcal{P}_m), \qquad \mathcal{A}^0_{m,n} = G_{\theta,n}(\mathcal{M}_m) \cap \mathcal{V}_0^n. \tag{1}$$

Thus $\mathcal{A}^0_{m,n}$ is exactly the set of targets generated by some admissible prompt. The retained state may evolve during generation provided no new prompt information enters after the cut. Any prompt-dependent parameters or auxiliary quantities created during prompt processing are part of that state.

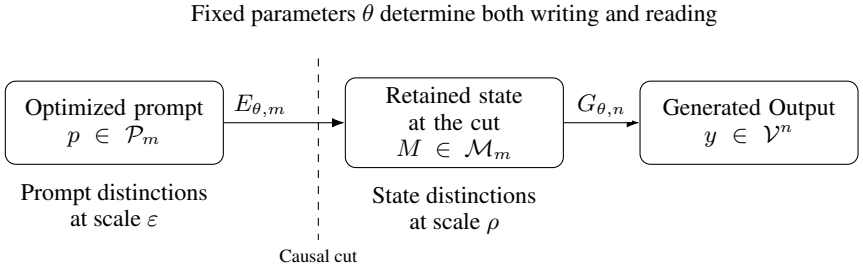


Figure 1: **Prompt information must be retained and usable.** All prompt-dependent information available after the causal cut is carried by the complete retained state, while the fixed model determines how that state is written and read.

Continuous prompts provide a controlled probe of what the frozen model can express, rather than what a user can necessarily elicit with ordinary text. Since both prompts and retained states are continuous, we introduce robustness at resolutions ε and ρ.

Robust accessibility. Equip $\mathbb{R}^{dm}$ with the Euclidean norm and write $\overline{B}(p, r)$ for the corresponding closed ball. Let $\mathcal{D}_M$ be the complete-state domain with metric $d_{\mathcal{M}}$. A target y is *jointly robustly accessible* if there exists a single witness $p_y \in \mathcal{P}_m$ such that

$$G_{\theta,n}(E_{\theta,m}(p')) = y \qquad \text{for every } p' \in \overline{B}(p_y, \varepsilon/2) \subseteq \mathcal{U}_m, \tag{2}$$

$$G_{\theta,n}(M') = y \qquad \text{for every } M' \in \mathcal{D}_M \text{ with } d_{\mathcal{M}}(M', E_{\theta,m}(p_y)) \le \rho. \tag{3}$$

The state condition includes perturbations that need not be reachable from another prompt, so it is distinct from prompt robustness. In both cases, the entire discrete output sequence must remain unchanged. We denote the resulting target set by $\mathcal{A}^{\varepsilon,\rho}_{m,n}$.

4 Joint constraints on memory and model expressivity

4.1 A joint bound on accessible continuations

The causal cut makes the joint constraint explicit. To robustly realize different targets, the model must distinguish them both before the cut, through their prompts, and after the cut, through the retained states. Hence the number of robustly accessible targets is bounded by the number of distinguishable prompts, the number of distinguishable retained states, and the size of the target space.

Let $\mathsf{Pack}_{>\delta}(X)$ be the largest size of a finite subset of X whose points are pairwise more than δ apart, or the supremum if no largest finite subset exists. It measures how many points in X can be distinguished at resolution δ.

Theorem 4.1 (Joint accessibility bound)**.** *For the fixed protocol above,*

$$|\mathcal{A}^{\varepsilon,\rho}_{m,n}| \le \min\left\{\mathsf{Pack}_{>\varepsilon}(\mathcal{P}_m), \mathsf{Pack}_{>\rho}(\mathcal{M}_m), K^n\right\}. \tag{4}$$

Any independently established upper bound on this same set may also enter the minimum.

Proof. Choose one joint witness for each accessible target. If two prompts are at distance at most ε, their midpoint lies in both required prompt balls and must generate two different targets, a contradiction. If their states are at distance at most ρ, either state lies in the other's required state neighborhood, giving the same contradiction. The target-space bound is immediate. □

For a nonempty accessible set, we take base-two logarithms. Let $U_{\text{prompt}}(m)$, $U_{\text{mem}}(m)$, and $U_{\text{model}}(m, n)$ denote valid prompt-, memory-, and model-side bounds. Then

$$\log_2 |\mathcal{A}^{\varepsilon,\rho}_{m,n}| \le \min\{U_{\text{prompt}}(m), U_{\text{mem}}(m), U_{\text{model}}(m, n), n \log_2 K\}. \tag{5}$$

Each term bounds the same accessible set, so the minimum gives the strongest applicable constraint. Bounds on broader sets may also be used when valid. Appendix A.3 gives the required conditions.

Existing prompt- and model-side bounds. For transformers satisfying the assumptions of Meyer et al. (2026), their Theorem 4.6 provides a finite-prompt bound and Theorem 4.10 provides a model-side ceiling independent of prompt length. These give concrete choices for the prompt and model terms in Equation (5). Appendix A.3 gives the corresponding expressions and assumptions.

What matching means. Equation (5) formalizes the matching principle. If the model-side bound is the tighter constraint, relaxing the memory bound alone cannot improve the joint bound. Conversely, if the memory bound is tighter, relaxing the model-side bound alone cannot improve it. Matching does not require the two bounds to be numerically equal. Rather, memory and the fixed model must be considered together because either can become the limiting constraint.

The remaining results sharpen the memory term using its update dynamics and reachable-state geometry. They bound how many reachable memory states remain distinguishable at a given resolution, rather than the raw size of the stored state.

4.2 Write dynamics and a memory ceiling

Under stable updates, the memory-side packing bound grows at most linearly with prompt length and is capped by a prompt-independent ceiling.

Consider a complete memory state updated as $M_t = F_t(M_{t-1}, w_t)$, where $w_t \in \mathcal{W}$ is a possible write. We use Euclidean distance for vector states and Frobenius distance for matrix states. Assume a common domain $\mathcal{D} \subseteq \mathbb{R}^{d_M}$, containing the initial state and 0, that is forward invariant under every admissible update. We also assume uniform stability on this domain:

$$\|F_t(M, w) - F_t(M', w')\| \le \gamma \|M - M'\| + L\, d_{\mathcal{W}}(w, w'), \qquad \|F_t(0, w)\| \le \mu, \tag{6}$$

for all $M, M' \in \mathcal{D}$ and $w, w' \in \mathcal{W}$, with $0 \le \gamma < 1$ and $L > 0$. The first inequality makes the update contractive in the existing state and Lipschitz in the incoming write. Thus $\gamma < 1$ causes state differences to decay, while together with the bound μ, it keeps reachable-state norms bounded. Finally, we assume that $\mathcal{W}$ admits a finite cover at every positive resolution, as any bounded finite-dimensional write set does.

For example, $M_t = \gamma M_{t-1} + w_t$ with $\|w_t\| \le \mu$ satisfies these assumptions with Euclidean writes and $L = 1$.

Theorem 4.2 (Stable-memory envelope)**.** *Under these assumptions, for any fixed $\rho > 0$, there exist finite constants a_ρ and U_ρ, independent of m, such that*

$$\log_2 \mathsf{Pack}_{>\rho}(\mathcal{M}_m) \le \min\{m a_\rho, U_\rho\}. \tag{7}$$

The constants depend on the possible writes, stability parameters, initial state, and resolution. The proof and explicit choices of a_ρ and U_ρ are given in Appendix B.

The term $m a_\rho$ scales with the number of writes, while U_ρ provides a ceiling independent of prompt length. Thus the memory packing bound remains capped regardless of prompt length. This memory term enters Equation (5) alongside the prompt and model bounds, with the fixed model determining how the retained state is used for generation. The ceiling relies on uniform contraction, whereas the general joint bound does not. Appendix B also gives a finite-length result without contraction.

4.3 A SHARPER BOUND WHEN STATE GEOMETRY IS KNOWN

We next use the geometry of the reachable states to tighten the memory bound. Trained linear-attention states have been observed to exhibit low effective rank, suggesting that most variation may lie in far fewer directions than the full state dimension (Nazari & Rusch, 2026). A dimension-and-radius bound ignores this structure by treating every point inside a containing ball as possible. When a tighter geometric description of the reachable states is available, it can therefore give a substantially smaller memory bound.

Suppose the complete reachable-state set $\mathcal{M}_m$ is contained in an ellipsoid $c + \mathcal{E}(Q)$, where $\mathcal{E}(Q) = \{Q^{1/2}u : \|u\|_2 \le 1\}$ and $Q \succeq 0$. The positive eigenvalues $\lambda_1, \dots, \lambda_r$ are the squared semiaxis lengths of this ellipsoid. The ellipsoid must contain every state reachable by any admissible prompt. A covariance fitted to sampled states is not sufficient, since it describes only the observed states.

Proposition 4.3 (Geometry-sensitive memory bound)**.** *Under this geometric assumption, for any $\rho > 0$,*

$$\log_2 \mathsf{Pack}_{>\rho}(\mathcal{M}_m) \le \inf_{\beta > 0} \frac{1}{2} \sum_{i=1}^{r} \log_2 \left(1 + \beta^{-1} + (1+\beta)\frac{4\lambda_i}{\rho^2}\right). \tag{8}$$

If $r = 0$, the bound is zero. When some directions are sufficiently narrow, they can be removed by projection, with a corresponding reduction in the separation scale as detailed in Appendix C.

The proof adapts the standard packing-volume argument to the ellipsoid. Because the bound accounts for width in each direction, narrow directions support fewer distinguishable states and can substantially tighten the result. When all directions have equal width, the bound reduces exactly to the standard dimension-and-radius ball bound.

Appendix C.3 provides a complete worked example of this bound.

5 GATED LINEAR ATTENTION: A MECHANISM-SPECIFIC BOUND

The previous results use only general properties of memory. When the memory mechanism is known, its structure can yield tighter bounds. We illustrate this with Gated Linear Attention (GLA), whose recurrent update combines rank-one key–value writes with a multiplicative retention gate

(Yang et al., 2024). The rank-one writes restrict reachable states at finite prompt lengths, while an additional contraction assumption yields an effective horizon independent of prompt length.

For one recurrent block with key dimension d_k and value dimension d_v, the state evolves as

$$S_t = D_t S_{t-1} + k_t v_t^\top, \qquad S_0 = 0, \qquad S_t \in \mathbb{R}^{d_k \times d_v}. \tag{9}$$

Here $D_t = \mathrm{Diag}(\alpha_t)$ is the operator corresponding to GLA's vector-valued retention gate, and k_t, v_t are the key and value written at step t. We assume $\|D_t\|_{\mathrm{op}} \le \gamma$, $\|k_t\|_2 \le \kappa$, and $\|v_t\|_2 \le \nu$, so κ and ν bound each write while γ controls how strongly previous state is retained.

Proposition 5.1 (Native GLA rank and retention)**.** *Assume the bounds above hold with some $\gamma \ge 0$. Because retention cannot increase rank and each update adds a rank-one matrix, after m writes every reachable state satisfies $\mathrm{rank}(S_m) \le r_m := \min\{m, d_k, d_v\}$ and $\|S_m\|_F \le R_m := \kappa\nu \sum_{j=0}^{m-1} \gamma^j$. These rank and norm constraints imply*

$$\log_2 \mathsf{Pack}_{>\rho}(\mathcal{S}_m) \le \min\left\{ d_k d_v \log_2(1 + 2R_m/\rho), r_m(d_k + d_v + 1)\log_2(1 + 12R_m/\rho) \right\}, \tag{10}$$

where $\mathcal{S}_m$ is the reachable set of the block.

If, in addition, the uniform bound above satisfies $\gamma < 1$, the contribution of older writes decays. Keeping only the most recent $H \ge 1$ write contributions approximates every reachable state, uniformly over m, within $\tau_H = \kappa\nu\gamma^H/(1-\gamma)$. Therefore, whenever $\rho > 2\tau_H$,

$$\log_2 \mathsf{Pack}_{>\rho}\Big(\bigcup_{m \ge 0} \mathcal{S}_m \Big) \le r_H(d_k + d_v + 1)\log_2\left(1 + \frac{12R_H}{\rho - 2\tau_H}\right). \tag{11}$$

Appendix D gives the proof and extends the block-level result to the model's complete retained state. The finite-prompt restriction comes from GLA's rank-one writes, which constrain the reachable state before the matrix reaches full rank. Under the stronger condition $\gamma < 1$, sufficiently old writes become negligible at resolution ρ, giving the prompt-independent effective horizon in Equation (11).

Corollary 5.2 (Gated DeltaNet)**.** *The bounds of Proposition 5.1 also apply to the recurrent matrix state of Gated DeltaNet (Yang et al., 2025b). After transposing its standard state convention,*

$$\widetilde{S}_t = A_t \widetilde{S}_{t-1} + \beta_t k_t v_t^\top, \qquad A_t = \alpha_t(I - \beta_t k_t k_t^\top).$$

If $\|k_t\|_2 = 1$, $0 \le \beta_t \le 1$, $\|v_t\|_2 \le \nu$, and $\alpha_t \le \gamma$ over all admissible trajectories, then Proposition 5.1 holds with $\kappa\nu$ replaced by ν. Thus $R_m^\Delta = \nu \sum_{j=0}^{m-1} \gamma^j$, and, when $\gamma < 1$, the effective-horizon bound holds with $R_H^\Delta = \nu \sum_{j=0}^{H-1} \gamma^j$ and $\tau_H^\Delta = \nu\gamma^H/(1-\gamma)$.

Proof sketch. Since $\|I - \beta_t k_t k_t^\top\|_{\mathrm{op}} \le 1$, we have $\|A_t\|_{\mathrm{op}} \le \gamma$, while $\beta_t k_t v_t^\top$ has rank at most one and Frobenius norm at most ν. The rank, norm, and recent-write arguments of Proposition 5.1 therefore apply unchanged. As with GLA, a complete-model bound must additionally account for any other prompt-dependent state, such as the short-convolution buffers.

6 Experiments

The theory bounds robust accessibility under specified prompt and state perturbations. The experiments study the same interaction in pretrained checkpoints using exact target recovery. With model weights fixed, we vary retained memory, its representation, and the prompt budget available to use it. This lets us test when additional memory expands what the model can be prompted to generate and when that gain depends on prompt budget or how the memory is represented.

6.1 What we measure

Following prior cramming work (Kuratov et al., 2025; Meyer et al., 2026), we freeze the model and optimize a continuous prompt for each target using teacher-forced cross-entropy. At evaluation, the prompt is removed after producing the retained state, and the model must generate the full target using only that state and its own previous outputs. A target is recovered if any declared run reproduces it exactly, and the fraction recovered is our *discovered exact coverage*. We summarize coverage using L_{50}, the longest tested length with at least 50% recovery. For the primary GLA projection

experiments, only the saved state is passed to a separate generation process, which receives neither the optimized prompt nor the target. Appendix E gives the full protocol and result tables.

We call the prompt-dependent state crossing the causal cut the *saved-state payload*. In the GLA experiments, we vary this payload by projecting each recurrent matrix and keeping only the projection coefficients. The projection basis is fixed and therefore belongs to the readout rather than the prompt-dependent state. Because the dense state is still formed before projection, saved-state payload is not total working memory, and these experiments do not measure end-to-end memory savings.

6.2 Memory frontiers and their dependence on prompt budget

Our main experiments vary prompt budget and retained memory while keeping the model fixed. We first increase retained memory at a fixed prompt length, then vary prompt length for fixed memory channels. This lets us see not only whether each resource helps, but how their effects depend on one another.

More retained memory moves the frontier. We begin with the frozen GLA-340M checkpoint at $m = 512$. Five projected channels of rank $32, 48, 64, 80, 96$ retain $3, 4.5, 6, 7.5, 9$ MiB, compared with the 12-MiB dense state. Here projection rank refers to the rank of the experimental projection, not the native GLA write rank. Twelve targets are paired across channels at each length, with up to 20K optimization updates.

As the targets get longer, a larger retained channel is needed to maintain majority coverage. At length 3072, rank 32 recovers $8/12$ targets. At 4096, rank 64 is the smallest channel reaching a majority, with $10/12$, and at 5120 this moves to rank 96, again with $10/12$. The evaluated L_{50} therefore moves from 3072 to 4096 to 5120 as the retained payload increases from 3 to 6 to 9 MiB.

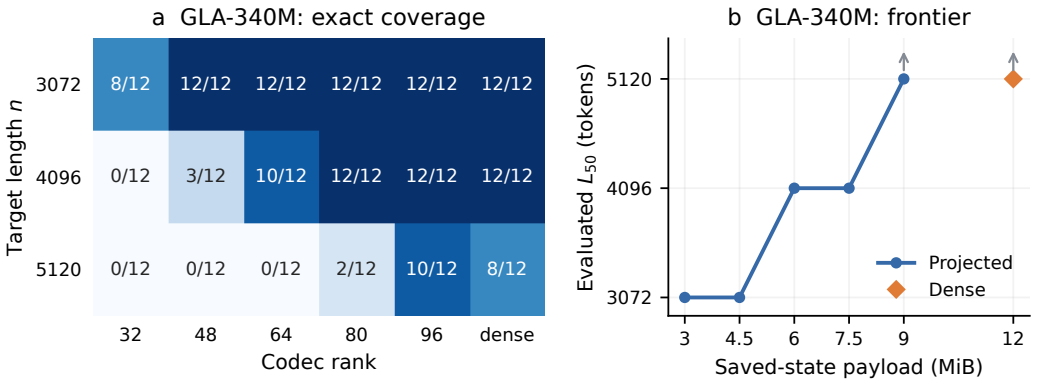


Figure 2: **Retained memory moves the GLA-340M frontier.** At fixed prompt budget $m = 512$, larger projected channels maintain exact coverage to longer continuations. **Left:** discovered exact coverage. **Right:** L_{50}, the longest tested length with at least 50% coverage. Rank 96 and dense are right-censored at 5120. The GLA-1.3B replication appears in Figure 6.

The same pattern appears on a larger checkpoint. We repeat the experiment on GLA-1.3B at the same prompt budget $m = 512$, using 24 fresh paired targets across seven continuation lengths. Projected ranks $64, 128, 192$ retain 12, 24, and 36 MiB, while the dense state retains 48 MiB, with up to 10K optimization updates.

The evaluated L_{50} rises from 8192 tokens at 12 MiB to 10240 at 24 MiB and 12288 at 36 MiB. The dense 48-MiB state also reaches 12288. Every condition falls below 50% coverage at a longer tested length, and the three projected-channel L_{50} values are identical across the two independently executed groups of 12 targets. Taken together, the two checkpoints show that increasing retained memory can push the accessibility frontier outward. For GLA-1.3B, the 36- and 48-MiB conditions reach the same frontier on the tested grid. Figure 6 and Table 4 in Appendix E.2.3 give the full results.

Prompt budget changes what a fixed memory channel can support. We next fix the target length at 4096 and compare the same 24 targets across three prompt budgets, $m = 256, 512, 1024$, and four memory conditions. Each condition uses two initializations and the same 20K search schedule.

Prompt budget	rank 32	rank 64	rank 96	dense
$m = 256$	$0/24$	$0/24$	$0/24$	$0/24$
$m = 512$	$0/24$	$20/24$	$24/24$	$24/24$
$m = 1024$	$16/24$	$24/24$	$24/24$	$24/24$

Increasing the prompt budget improves or saturates coverage in every memory condition. At $m = 256$, all four channels recover $0/24$ targets. At $m = 512$, larger memory becomes useful, with coverage rising from $0/24$ at rank 32 to $20/24$ at rank 64 and $24/24$ at rank 96. At $m = 1024$, even rank 32 reaches $16/24$, while all larger channels reach $24/24$. Together, these results show that prompt budget and retained memory shape accessibility jointly. A larger prompt can make a fixed memory channel more effective, while additional memory may provide little benefit when the prompt budget is small.

More prompt does not always move the frontier. Finally, we test whether increasing prompt budget continues to help the dense 12-MiB state. Using 20 paired targets per length and 6K-, 10K-, and 20K-update searches from matched initializations, we obtain:

Prompt budget	length 4096	length 8192
$m = 256$	$0/20$	not tested
$m = 512$	$20/20$	$0/20$
$m = 1024$	$20/20$	$0/20$

Increasing the prompt from 256 to 512 changes coverage at length 4096 from $0/20$ to $20/20$, while coverage at length 8192 remains at $0/20$ even at $m = 1024$. Together with the memory interventions above, these results show that accessibility depends on the combination of prompt budget and retained memory rather than on either quantity alone.

6.3 A NATIVE-MEMORY COMPARISON WITH SLIDING-WINDOW ATTENTION

We use sliding-window attention as a native-memory counterpart to the projected GLA experiments above.

We keep one Qwen3-0.6B weight set fixed, use a prompt budget of $m = 1000$, and compare attention memory windows of $16, 32, 64, 128, 256, 512$ with full attention. Each evaluated length uses 12 targets, paired with its same-weight full-attention control, and up to 5K optimization updates per prompt. As the attention memory window grows, the discovered continuation frontier moves outward. Adaptive refinement places the observed 50% crossings within 64–72, 126–136, 252–262, 476–484, 862–872, and 1844–1856 tokens, respectively. The lower endpoint is the evaluated L_{50}; the upper endpoint is the shortest tested length below 50%, not a bound on all longer continuations. Full attention remains above 50% at every crossing endpoint and recovers $11/12$ targets at the longest tested length of 2048, so its frontier is not reached.

This comparison shows the SWA memory-frontier response under our causal-cut protocol across the evaluated window sizes. Increasing the attention memory window allows the same fixed model to recover longer continuations. The response is approximately proportional over this range, but the refined crossings do not follow an exact linear law. Window size changes both retained memory and attention access, so this is a native-memory comparison rather than a storage-only intervention. Appendix E.3 gives the complete counts and search procedure.

Additional breadth from fixed-state Mamba. Mamba-130M, 370M, and 790M provide complementary evidence from fixed-size recurrent memory. Teacher-forced prompt-length sweeps show substantial growth followed by smaller gains at larger prompt budgets, while the recurrent state dimensions remain fixed within each checkpoint. These descriptive results do not identify a strict memory ceiling. Separate Mamba-130M saved-state controls verify that the fixed-size saved state can serve as the complete interface across the causal cut. Appendix E.4 reports the curves, their protocol differences, and the strict controls.

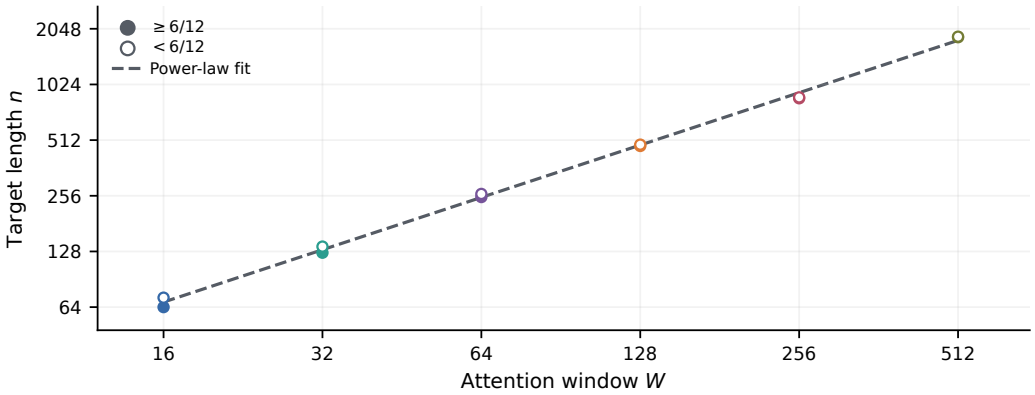


Figure 3: **Sliding-window memory moves the Qwen3-0.6B frontier.** Measured 50% crossings under fixed weights and $m = 1000$. Filled and open markers denote qualifying and nonqualifying endpoints; the short connecting segments indicate search resolution, not confidence intervals. The dashed power-law fit to bracket midpoints is a descriptive guide. Paired full-attention controls remain above 50% at each endpoint and reach $11/12$ at 2048; their counts appear in Appendix E.3.

6.4 Equal-size channels can support different outputs

The previous experiments show that memory size matters. We next ask if two equally sized memory channels are equally useful. On GLA-340M, we compare two rank-64 representations. One preserves directions that commonly appear in earlier states, while the other uses a random orthonormal basis. Both retain the same 6-MiB payload and are fixed before evaluation on 24 fresh targets.

At $m = 512, n = 4096$, the informed channel recovers $23/24$ targets, compared with $1/24$ for the random channel and $24/24$ for dense state. The difference is highly significant even after correcting for multiple comparisons (Holm-adjusted $p = 1.43 \times 10^{-6}$). Thus memory size alone does not determine accessibility. Which parts of the state are retained also matter. Appendix E.2.6 gives the full comparison and additional controls with longer optimization and independently sampled random bases.

7 Discussion and Conclusion

Model-specific memory sizing. Memory budgets should be model-specific. Memory and model capacity constrain the same accessible outputs, so enlarging memory alone cannot improve the joint bound once a model-side constraint becomes tighter. This may matter especially for smaller models, where compact memory could avoid spending HBM and bandwidth on retained state that provides little additional benefit.

Optimization as an empirical probe. As in prior cramming work (Kuratov et al., 2025; Meyer et al., 2026), continuous-prompt optimization may fail to find outputs that are nevertheless accessible. We therefore treat changes in exact recovery as evidence about discovered accessibility, not as proof that unrecovered targets are impossible for the model.

Future work. A key next step is to locate where memory ceases to be the limiting resource in trained models by tightening memory- and model-side bounds for the same checkpoint and controlling its memory mechanism. Extending the analysis to other architectures, approximate or stochastic generation, and end-to-end HBM, throughput, and latency measurements could turn the matching principle into practical guidance for memory sizing.

Taken together, this work provides a new understanding of how prompt budget, retained memory, and the fixed model behave as coupled resources. The benefit of a larger prompt or memory depends on the model that must use it, while the model can only act on the information made available through the prompt and retained state. The design problem is therefore to balance these resources rather than scale each independently.

Ethics Statement

This work studies frozen language models using synthetic target sequences and does not involve human participants or the collection of personal data. We do not identify ethical risks specific to the methods studied here.

Reproducibility Statement

The appendix gives all theorem proofs, a fully specified analytical memory example, target construction, optimization schedules, complete-state accounting, and the primary result tables. Experimental comparisons use frozen checkpoints, paired targets, declared search schedules, and saved-state validation. The supplementary code provides a minimal reference implementation of the causal-cut evaluation framework and representative memory interventions, together with example configurations for running new experiments.

Statement on use of AI tools

AI assistants were used for literature search and related-work discovery, research ideation and mathematical exploration, development and checking of mathematical claims and proofs, feedback on experimental design, analysis, and interpretation of results, assistance with experimental and analysis code, and manuscript writing and revision. All AI-assisted work was reviewed by the authors, who take full responsibility for the paper's content, claims, and conclusions.

References

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, 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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1763–1840, 2024. URL https://proceedings.mlr.press/v235/arora24a.html.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/a4ca07aa108036f80cbb5b82285fd4b1-Abstract-Conference.html.

Ali Behrouz, Meisam Razaviyayn, Peilin Zhong, and Vahab Mirrokni. It's all connected: A journey through test-time memorization, attentional bias, retention, and online optimization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gZyEJ2kMow.

Satwik Bhattamishra, Michael Hahn, Phil Blunsom, and Varun Kanade. Separations in the representational capabilities of transformers and recurrent architectures. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3f630b20b7b3ac76d3a0016fe29b6dc0-Abstract-Conference.html.

Stephen Boyd and Leon O. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161, 1985. doi: 10.1109/TCS.1985.1085649. URL https://stanford.edu/~boyd/papers/pdf/fading_volterra.pdf.

Stanislav Fort. Scaling laws for adversarial attacks on language model activations and tokens. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5434a6b40f8f65488e722bc33d796c8b-Abstract-Conference.html.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Jerry Yao-Chieh Hu, Wei-Po Wang, Ammar Gilani, Chenyang Li, Zhao Song, and Han Liu. Fundamental limits of prompt tuning transformers: Universality, capacity and efficiency. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/48dfc849640344e2d58df0b5bb78c33b-Abstract-Conference.html.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21502–21521, 2024. URL https://proceedings.mlr.press/v235/jelassi24a.html.

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*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165, 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Kimi Team et al. Kimi Linear: An expressive, efficient attention architecture, 2025. URL https://arxiv.org/abs/2510.26692. arXiv:2510.26692.

Yuri Kuratov, Mikhail Arkhipov, Aydar Bulatov, and Mikhail Burtsev. Cramming 1568 tokens into a single vector and back again: Exploring the limits of embedding space capacity. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19323–19339, 2025. doi: 10.18653/v1/2025.acl-long.948. URL https://aclanthology.org/2025.acl-long.948/.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A hybrid Transformer-Mamba language model, 2024. URL https://arxiv.org/abs/2403.19887. arXiv:2403.19887.

Maxime Meyer, Mario Michelessa, Caroline Chaux, and Vincent Y. F. Tan. How many different outputs can a transformer generate? In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.22223v2.

Philipp Nazari and T. Konstantin Rusch. The key to state reduction in linear attention: A rank-based perspective, 2026. URL https://arxiv.org/abs/2602.04852. arXiv:2602.04852.

Robert G. Newcombe. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, 17(22):2635–2650, 1998. doi: 10.1002/(SICI)1097-0258(19981130)17:22<2635::AID-SIM954>3.0.CO;2-C. URL https://doi.org/10.1002/(SICI)1097-0258(19981130)17:22<2635::AID-SIM954>3.0.CO;2-C.

Rom Parnichkun, Neehal Tumma, Armin W. Thomas, Alessandro Moro, Qi An, Taiji Suzuki, Atsushi Yamashita, Michael Poli, and Stefano Massaroli. Quantifying memory utilization with effective state-size. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48276–48334, 2025. URL https://proceedings.mlr.press/v267/parnichkun25a.html.

Aleksandar Petrov, Philip H. S. Torr, and Adel Bibi. When do prompting and prefix-tuning work? a theory of capabilities and limitations. In *International Conference on Learning Representations*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/18a367688479c1b08b001584218a4443-Abstract-Conference.html.

Aleksandar Petrov, Philip H. S. Torr, and Adel Bibi. Prompting a pretrained transformer can be a universal approximator. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40523–40550, 2024b. URL https://proceedings.mlr.press/v235/petrov24a.html.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366, 2021. URL https://proceedings.mlr.press/v139/schlag21a.html.

Eduardo D. Sontag. On the input-to-state stability property. *European Journal of Control*, 1(1):24–36, 1995. doi: 10.1016/S0947-3580(95)70005-X. URL https://www.sciencedirect.com/science/article/pii/S094735809570005X.

Ao Sun, Hongtao Zhang, Heng Zhou, Yixuan Ma, Yiran Qin, Tongrui Su, Yan Liu, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, and Renqing He. State rank dynamics in linear attention LLMs, 2026. URL https://arxiv.org/abs/2602.02195. arXiv:2602.02195.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (learn at test time): RNNs with expressive hidden states. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57503–57522, 2025. URL https://proceedings.mlr.press/v267/sun25h.html.

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. URL https://arxiv.org/abs/2307.08621. arXiv:2307.08621.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596. URL https://doi.org/10.1017/9781108231596.

Yihan Wang, Jatin Chauhan, Wei Wang, and Cho-Jui Hsieh. Universality and limitations of prompt tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/eef6aecfe050b556c6a48d9c16b15558-Abstract-Conference.html.

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. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/79dc391a2c1067e9ac2b764e31a60377-Abstract-Conference.html.

An Yang et al. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56501–56523, 2024. URL https://proceedings.mlr.press/v235/yang24ab.html.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4904fad153f6434a7bcf04465d4be2cc-Abstract-Conference.html.

A Protocol details and consequences of the causal cut

A.1 Complete state and fixed conventions

A state is complete relative to a declared continuation protocol. If two ingestions leave the same state and every downstream public convention is fixed, they must produce the same continuation. The learned parameters belong to the fixed write and read maps. Prompt-dependent fast weights, normalization accumulators, cache positions, or other fields do not: they must be included in M unless a proof shows they cannot affect the readout. A length counter may be public at a fixed m but require accounting when comparing the union over prompt lengths.

The complete-state domain $\mathcal{D}_M$ contains every state on which the perturbed readout is defined. Robustness in (3) is relative to this declared domain. No interpolation of a discrete field is implicit. The metric can be Euclidean on concatenated tensors, a product metric with discrete auxiliary fields, or another specified metric. Bounds derived for one choice cannot silently be used at another resolution or under a different perturbation domain.

For exact generation, EOS handling, tie-breaking, delimiter placement, and the first-token convention are part of $G_{\theta,n}$. Truncating at n tokens and treating an earlier EOS as failure is one valid fixed protocol. Providing a target token before evaluation instead defines a different readout, unless that token was independently verified as the model's own first prediction.

A.2 Same-witness counting and exact finite precision

Theorem 4.1 holds for every finite subset of the accessible targets, so it also applies when packing is defined as a supremum. The prompt midpoint argument uses the full closed ball contained in $\mathcal{U}_m$, not merely its intersection with an arbitrary nonconvex admissible prompt set. The state argument only needs one center to lie in the other center's ρ-neighborhood, which gives separation $> \rho$. It does not require midpoint closure of $\mathcal{D}_M$. Under a convex Euclidean perturbation domain, midpoint reasoning can sometimes strengthen this to $> 2\rho$; the paper uses the more general convention consistently.

Exact and robust bounds are distinct. At a fixed protocol,

$$|\mathcal{A}^0_{m,n}| \le \min\{|\mathcal{P}_m|,\ |\mathcal{M}_m|,\ |G_{\theta,n}(\mathcal{M}_m)|,\ K^n\}.$$

Here the image term can be much smaller than the state cardinality. If a sufficient code $q(M)$ takes at most 2^B values and $G_{\theta,n}(M) = \widetilde{G}_{\theta,n}(q(M))$ on every reachable state, then $|\mathcal{A}^0_{m,n}| \le 2^B$. Sufficiency, not the dimension of a tensor by itself, is the requirement. A real-valued scalar state can encode infinitely many exact messages if precision and robustness are unrestricted.

If at least a fraction η of the K^n targets is jointly robustly accessible, every valid logarithmic bound U on that set must satisfy

$$\log_2\lceil \eta K^n \rceil \le U.$$

In particular, the weaker but convenient inequality $n \log_2 K + \log_2 \eta \le U$ follows for $\eta > 0$. This is a necessary-resource statement, not an achieved coverage guarantee.

A.3 Combining Meyer's prompt and model bounds with retained memory

We use the numbering in arXiv v2 of Meyer et al. (2026). For embedding dimension d, embedding radius r, and prompt precision ε_p under their conventions, Theorem 4.6 gives the logarithmic bound

$$U^{\text{Meyer}}_{\text{prompt}}(m) = dm \log_2\left(1 + \frac{2r}{\varepsilon_p}\right). \tag{12}$$

Under their additional Assumption 4.9, with Wasserstein order $q \ge 1$ and precision $\varepsilon_W > 0$, Theorem 4.10 gives

$$U^{\text{Meyer}}_{\text{model}} = \left(1 + \frac{2r}{\varepsilon_W}\right)^d \log_2\left(e + e\left(\frac{2r}{\varepsilon_W}\right)^q\right). \tag{13}$$

The latter is independent of prompt length when its precision constants hold uniformly across that range. The assumption concerns the model's invariance within the reference classes of prompt empirical measures. It is stronger than ordinary vector precision, as their Appendix D explains.

To import either result, let $\mathcal{B}_{m,n}$ be a set of continuations covered by that result for the same generation rule. It is enough to establish

$$\mathcal{A}^{\varepsilon,\rho}_{m,n} \subseteq \mathcal{B}_{m,n}, \qquad |\mathcal{B}_{m,n}| \leq 2^U.$$

For example, if the theorem covers every output of that generator on a prompt domain containing $\mathcal{P}_m$, the inclusion follows directly. The additional state-robustness requirement only reduces our accessible set; the two results do not require identical robustness definitions. When both Meyer bounds apply, they can be used in Equation (5) alongside any valid retained-memory bound.

The source precisions $\varepsilon_p, \varepsilon_W$ are not automatically our witness scales ε, ρ. Their domains and assumptions must hold for the generator being bounded. In particular, fixing learned weights does not transfer a precision property to an altered memory/readout rule. We do not assume that the experimental GLA channels or Mamba satisfy Meyer's transformer assumptions. The import gives a concrete model term where applicable, rather than a universal parameter-count formula.

A.4 Exact budgets, at-most budgets, and eventual saturation

The theoretical set at an at-most budget is $\mathcal{A}_{\leq m,n} = \bigcup_{j=0}^{m} \mathcal{A}_{j,n}$. It grows monotonically with m by definition. At an exact budget, increasing m can remove previously accessible targets. A valid padding operation prevents this by preserving shorter-prompt solutions. Bounds proved separately at each exact length do not automatically bound this union. One may always sum the exact-length bounds. A common complete-state enclosure or a common sufficient finite code instead bounds the union directly.

For a fixed finite target family, this increasing sequence of sets eventually stabilizes: each accessible target has a finite witness length, and the maximum of finitely many such lengths suffices. A finite number of complete behavioral state classes gives the analogous statement for the full continuation behavior. This elementary existence result does not locate the stabilization point. It also does not show that a fixed-length state bound is the active restriction in a trained model.

Experimentally, let $I_i(j, n)$ indicate whether a declared search finds target i at prompt budget j. Exact-budget coverage is $N^{-1} \sum_i I_i(m, n)$. Where identities are shared across evaluated budgets, the cumulative statistic is

$$\widehat{C}_{\leq m,n} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\exists j \leq m \text{ evaluated} : I_i(j, n) = 1\}.$$

This is only discovered coverage over the tested budgets, not the theorem's full at-most-budget accessible set. Counts from different target populations cannot be pooled to form it.

A.5 When an exact witness can be certified robust

For a target y, let $Z^y_t(M)$ be the vector of logits obtained from initial state M when the preceding target tokens $y_{<t}$ are supplied. Suppose the correct logit has a strictly positive margin Δ_t at M_y, and every coordinate of Z^y_t is L_t-Lipschitz on a state ball of radius r. If $\rho \leq r$ and $2L_t\rho < \Delta_t$ for all $t \leq n$, the correct token remains the unique maximum at every step. Induction on t therefore proves exact autonomous generation throughout that ball. The Lipschitz constants must cover the complete continuation dynamics, not only the final linear head.

If ingestion is also Λ-Lipschitz on the required prompt ball and $\Lambda\varepsilon/2 \leq \rho$, that same state margin implies prompt robustness. This supplies a route from successful exact witnesses to robust lower bounds. No such worst-case margin certification is claimed for the experiments in this paper.

B Proof of the stable-memory envelope

B.1 Definitions and the write-cover construction

Writes may be features computed from tokens rather than the raw tokens themselves. A particular model may restrict which write sequences can occur. Allowing all sequences in $\mathcal{W}$ only enlarges the set being bounded.

An internal δ-cover of $\mathcal{W}$ is a finite set of centers in $\mathcal{W}$ such that every write is within δ of one center. Its minimal cardinality is $\mathsf{N}_\delta(\mathcal{W})$. The explicit constants in Theorem 4.2 are

$$a_\rho = \log_2 \mathsf{N}_{\rho(1-\gamma)/(2L)}(\mathcal{W}),$$
$$U_\rho = \min\left\{H_\rho \log_2 \mathsf{N}_{\rho(1-\gamma)/(4L)}(\mathcal{W}),\ d_M \log_2(1 + 2R_\infty/\rho)\right\}.$$

where $R_\infty = \max\{\|M_0\|, \mu/(1-\gamma)\}$ and $H_\rho \geq 1$ satisfies $\gamma^{H_\rho} R_\infty \leq \rho/4$. Centers remain admissible writes, so the approximate trajectories stay in the invariant domain required by (6). For a reachable M_m, replace each write by a cover center and call the result $\widetilde{M}_m$. Starting from the same M_0, the error satisfies

$$e_t := \|M_t - \widetilde{M}_t\| \leq \gamma e_{t-1} + L\delta, \qquad e_m \leq L\delta \sum_{j=0}^{m-1} \gamma^j \leq \frac{L\delta}{1-\gamma}.$$

Take $\delta = \rho(1-\gamma)/(2L)$. The $\mathsf{N}_\delta(\mathcal{W})^m$ approximate endpoints cover all reachable states with balls of radius $\rho/2$. Such a ball contains at most one point of any $> \rho$-separated set, giving the linear term of (7). The empty-write case is a singleton and has zero logarithmic packing.

B.2 Containing ball and recent-write cap

Setting $M' = 0$ and $w' = w$ gives $\|F_t(M, w)\| \leq \gamma\|M\| + \mu$. Therefore

$$\|M_t\| \leq \gamma^t \|M_0\| + \mu \sum_{j=0}^{t-1} \gamma^j \leq R_\infty, \qquad R_\infty = \max\{\|M_0\|, \mu/(1-\gamma)\}.$$

Disjoint open balls of radius $\rho/2$ around a finite $> \rho$-separated subset of this ball lie in the ball of radius $R_\infty + \rho/2$. Comparing volumes gives

$$\mathsf{Pack}_{>\rho}(\mathcal{M}_m) \leq (1 + 2R_\infty/\rho)^{d_M}.$$

This is the usual Euclidean volume argument (Vershynin, 2018).

Choose $H_\rho \geq 1$ with $\gamma^{H_\rho} R_\infty \leq \rho/4$. For $m > H_\rho$, replace the state at time $m - H_\rho$ by 0 and retain the last H_ρ writes. The resulting endpoint error is at most $\gamma^{H_\rho} R_\infty$. Quantize those writes at $\bar{\delta} = \rho(1-\gamma)/(4L)$ to add at most $\rho/4$ further error. Thus $\mathsf{N}_{\bar{\delta}}(\mathcal{W})^{H_\rho}$ balls of radius $\rho/2$ cover the exact-length reachable set. For $m \leq H_\rho$, quantizing all writes from the true initial state uses no more centers and error at most $\rho/4$. Together these prove Theorem 4.2.

For $0 < \gamma < 1$, an explicit choice is

$$H_\rho = \max\left\{1,\ \left\lceil \frac{\log \max\{1, 4R_\infty/\rho\}}{-\log \gamma} \right\rceil \right\}.$$

For $\gamma = 0$, take $H_\rho = 1$. This is a bound on the *count* at each m; when F_t depends on time, the recent-write cover centers may depend on m. The ambient containing ball, in contrast, bounds the union over m.

B.3 A memory satisfying the stability assumptions

Consider $F_t(M, w) = \gamma M + w$ on $\mathcal{D} = \mathbb{R}^{d_M}$ with $0 \leq \gamma < 1$ and $\mathcal{W} = \{w : \|w\|_2 \leq \mu\}$. The triangle inequality gives $\|F_t(M, w) - F_t(M', w')\|_2 \leq \gamma\|M - M'\|_2 + \|w - w'\|_2$, so $L = 1$, and $\|F_t(0, w)\|_2 \leq \mu$. The write ball has finite internal covers at every positive resolution, and the full Euclidean state domain is invariant. Thus all assumptions of Theorem 4.2 hold for any fixed initial state. The containing-ball argument above gives $\|M_t\|_2 \leq \max\{\|M_0\|_2, \mu/(1-\gamma)\}$ uniformly in t.

B.4 Write rate and the memory ceiling

More generally, the number $T(m)$ of prompt-dependent ingestion updates replaces m in Equation (7). These are events during prompt processing, not steps of the prompt optimizer. As a deliberately naive example, consider an interface that admits one unaggregated write every s positions, starting at the first position, and performs no update at the others. Then $T(m) = \lceil m/s \rceil$. Holding

the state domain and per-write dynamics fixed divides the asymptotic slope of the linear upper envelope by s, while leaving its cap unchanged. This compares upper bounds, not the prompt length at which observed performance saturates.

The example isolates only how often writes enter memory. Practical recurrent mechanisms are more nuanced: they may update at every position while gating, mixing, compressing, or forgetting information. Their memory bound depends on these dynamics and on the geometry of the states they produce, not simply on an update count.

B.5 Without contraction

A finite-horizon version does not need $\gamma < 1$. Under (6) with $\gamma \geq 0$, define $g_m = \sum_{j=0}^{m-1} \gamma^j$ and $R_m = \gamma^m \|M_0\| + \mu g_m$. For $m \geq 1$,

$$\log_2 \mathsf{Pack}_{>\rho}(\mathcal{M}_m) \leq \min\left\{ m \log_2 \mathsf{N}_{\rho/(2Lg_m)}(\mathcal{W}),\ d_M \log_2(1 + 2R_m/\rho) \right\}.$$

The same error recursion and volume proof apply. When $\gamma \geq 1$, the radius and required write resolution may deteriorate with m, so this argument alone gives no prompt-independent cap. Other invariance or boundedness assumptions can still supply one.

The contraction condition is related to fading memory and input-to-state stability (Boyd & Chua, 1985; Sontag, 1995). Here its particular role is to turn finite write covers into an explicit packing bound at a declared resolution. The result requires contraction of the complete state or a separately valid sufficient representation, not merely decay in one submodule.

C Geometry-sensitive bounds and worked application

C.1 Ellipsoid enlargement and packing proof

Let $r = \operatorname{rank}(Q) > 0$, and work in the r-dimensional affine span $c + \operatorname{range}(Q)$. Distances within this span are the original Euclidean distances. Translation is irrelevant, so set $c = 0$. For $t > 0$ and $\beta > 0$,

$$\mathcal{E}(Q) + tB_2^r \subseteq \mathcal{E}\big((1+\beta)Q + (1+\beta^{-1})t^2 I_r\big). \tag{14}$$

To prove this inclusion, for any direction v the support function of the left side is $\sqrt{v^\top Q v} + t\|v\|_2$. The scalar inequality

$$(\sqrt{a} + \sqrt{b})^2 \leq (1+\beta)a + (1+\beta^{-1})b$$

shows it is at most the support function of the right side. The support-function order implies inclusion for compact convex sets.

For any finite $>$ ρ-separated subset of $\mathcal{M}_m$, open balls of radius $t = \rho/2$ are disjoint and lie in $\mathcal{E}(Q) + tB_2^r$. The volume ratio in (14) is

$$\prod_{i=1}^{r} \left((1+\beta)\lambda_i/t^2 + 1 + \beta^{-1}\right)^{1/2}.$$

Taking logarithms and then the infimum over $\beta > 0$ proves (8). No covariance estimation or distributional assumption enters this argument. Rank zero gives a singleton, with packing number one. The ingredients are classical volume and convexity inequalities; the refinement applies them to the complete-state term in the causal-cut bound.

C.2 Removing narrow directions with the correct penalty

Order the positive eigenvalues so that $\lambda_1 \geq \cdots \geq \lambda_r > 0$. For $1 \leq j < r$, let Π_j project onto the first j eigenvectors. Every $x \in \mathcal{E}(Q)$ satisfies $\|(I - \Pi_j)x\| \leq \sqrt{\lambda_{j+1}}$. Thus, for x, x' in the ellipsoid,

$$\|\Pi_j(x - x')\|^2 = \|x - x'\|^2 - \|(I - \Pi_j)(x - x')\|^2 > \rho^2 - 4\lambda_{j+1}$$

whenever $\|x - x'\| > \rho$. If $4\lambda_{j+1} < \rho^2$, define $\rho_j = (\rho^2 - 4\lambda_{j+1})^{1/2} > 0$. Projection is injective on the separated set, giving

$$\log_2 \mathsf{Pack}_{>\rho}(\mathcal{M}_m) \leq \inf_{\beta > 0} \frac{1}{2} \sum_{i=1}^{j} \log_2 \left(1 + \beta^{-1} + (1+\beta)\frac{4\lambda_i}{\rho_j^2}\right). \tag{15}$$

One may minimize this over admissible j and the untruncated result. If $2\sqrt{\lambda_1} \le \rho$, the ellipsoid's diameter is at most ρ and packing is one directly. Discarding eigenvalues without reducing ρ would not be a valid bound.

C.3 Fully specified mechanism and strict improvement

A complete worked application. To isolate geometry while holding dimension and radius fixed, consider a memory with eight directions that can vary widely and 120 whose variation is restricted. Let one admissible write be any $u \in \mathbb{R}^{128}$ with $\|u\|_2 \le 1$, and retain $M = A_\alpha u$, with $A_\alpha = \mathrm{diag}(10I_8, 10\alpha I_{120})$ and $0 < \alpha \le 1$. The reachable set is exactly $\mathcal{E}(A_\alpha A_\alpha^\top)$. Its dimension and radius do not change with α, so the ball bound on the logarithm of the state count is always $128\log_2(21) = 562.22$ at $\rho = 1$. At $\alpha = 10^{-3}$, Equation (8) lowers this bound to 56.66. Accounting separately for the small error caused by discarding the narrow directions tightens it to 35.14 (Figure 4). The derivation below proves strict improvement over the ball bound for every $0 < \alpha < 1$ without numerical optimization. The improvement therefore follows from a specified memory mechanism: its full state still has 128 dimensions, but most directions cannot vary enough to support many distinguishable states at this resolution.

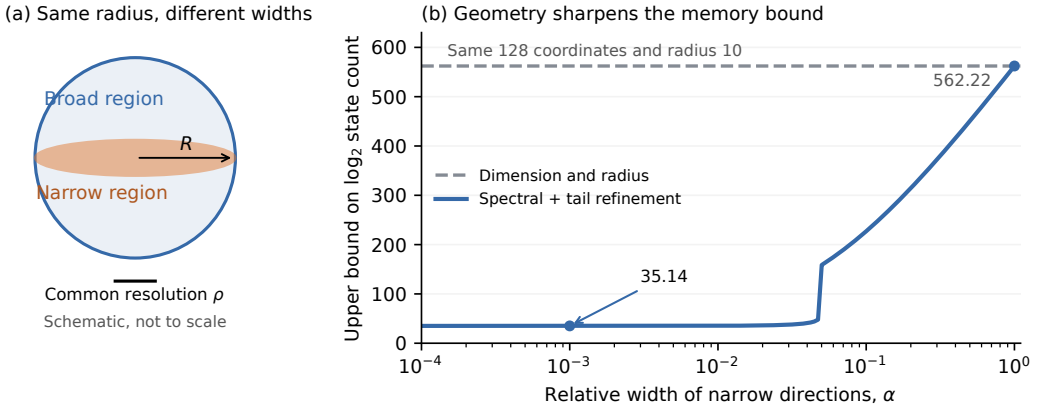


Figure 4: **The same dimension and radius can yield different memory bounds.** A specified 128-dimensional memory has eight wide directions and 120 directions of relative width α. All positive α give full rank and radius 10. At separation $\rho = 1$, accounting for narrow directions substantially reduces the upper bound on robust state distinctions. These are theoretical upper bounds for the defined mechanism, not measured checkpoint capacities.

General mechanism and proof. Take dimension d, wide-direction count $1 \le k < d$, radius $R > 0$, and $0 < \alpha \le 1$. The admissible write is the entire unit Euclidean ball, the initial state is fixed, and a single update sets

$$M = A_\alpha u, \quad \|u\|_2 \le 1, \qquad A_\alpha = \mathrm{diag}(RI_k, \alpha R I_{d-k}).$$

There are no hidden retained fields. This defined mechanism has exactly the reachable ellipsoid with squared axes R^2 and $\alpha^2 R^2$. For every positive α, its rank is d and its radius is exactly R.

Write $q = 2R/\rho$. The ball bound is $U_{\mathrm{ball}} = d\log_2(1+q)$. Evaluating the spectral expression at the explicit choice $\beta = 1/q$ gives

$$U_{\mathrm{fixed}} = k\log_2(1+q) + \frac{d-k}{2}\log_2\big((1+q)(1+q\alpha^2)\big). \tag{16}$$

Consequently,

$$U_{\mathrm{ball}} - U_{\mathrm{fixed}} = \frac{d-k}{2}\log_2\frac{1+q}{1+q\alpha^2} > 0 \quad \text{for } 0 < \alpha < 1. \tag{17}$$

The infimum in (8) is no larger than this explicit value. Thus strict improvement is analytical, not a consequence of the accuracy of a numerical optimizer. At $\alpha = 1$, the optimal spectral expression

equals the ball bound. This follows either by differentiation or by minimizing $\beta q^2 + \beta^{-1}$, whose minimum is $2q$.

When $2\alpha R < \rho$, projection onto the first k coordinates yields

$$U_{\text{tail}} \le k \log_2 \left(1 + \frac{2R}{\sqrt{\rho^2 - 4\alpha^2 R^2}} \right).$$

For $d = 128$, $k = 8$, $R = 10$, $\rho = 1$, and $\alpha = 10^{-3}$, the bounds on the base-two logarithm of the distinguishable state count are

$$U_{\text{ball}} = 562.216630, \qquad U_{\text{spec}} = 56.658809, \qquad U_{\text{tail}} = 35.140738.$$

The explicit nonoptimized choice (16) gives 298.679316. As α tends to zero the projected bound tends to $8 \log_2(21) = 35.138539$.

Figure 5 shows the intermediate spectral expression. The numerical calculation sweeps 167 positive widths and optimizes the one-dimensional objective in $\log \beta$; the explicit formulas above independently check the isotropic endpoint, strict-improvement sign, and tail regime. These calculations compare upper bounds for an exactly specified memory. A decoder is not constructed to attain them, and the informed GLA projection basis is not used to define this ellipsoid.

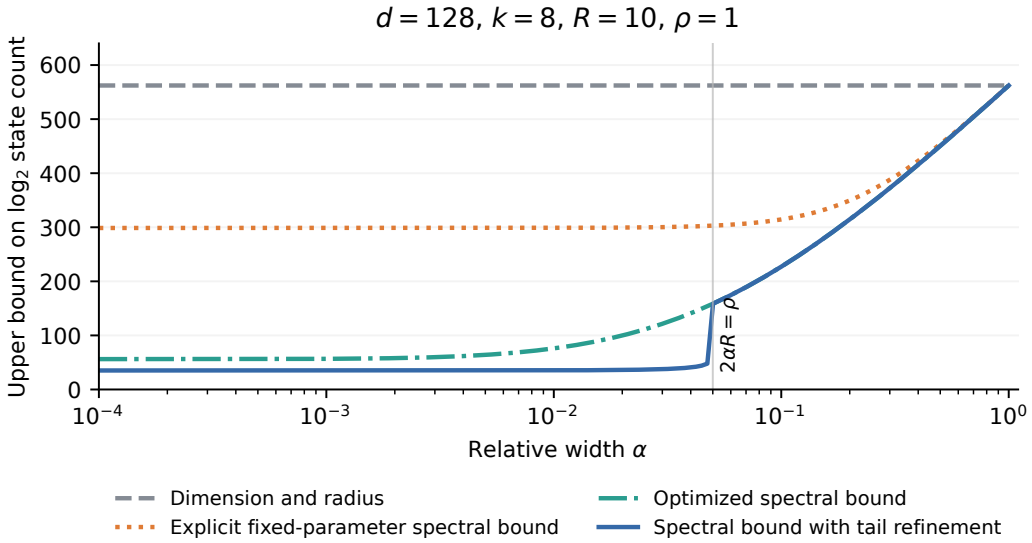


Figure 5: **Analytical bound comparison, including the untruncated refinement.** The containing ellipsoid follows exactly from the admissible write mechanism. The same $\rho = 1$ separation convention is used for every bound; tail removal includes the reduced projected separation.

D Proof and scope of the native GLA specialization

D.1 Rank, norm, and low-rank covering

The rank statements concern the recurrence in real arithmetic; numerical snapshots are not treated as exact realizations of those rank identities. Unrolling (9) gives

$$S_m = \sum_{t=1}^{m} (D_m D_{m-1} \cdots D_{t+1}) k_t v_t^\top.$$

The empty gate product is the identity. Each summand has rank at most one and Frobenius norm at most $\kappa\nu\gamma^{m-t}$. Rank subadditivity and the triangle inequality give r_m and R_m. This argument permits trajectory-dependent keys, values, and gates.

For completeness, define the rank-constrained ball

$$\mathcal{L}_r(R) = \{S \in \mathbb{R}^{d_k \times d_v} : \operatorname{rank}(S) \le r,\ \|S\|_F \le R\}.$$

For $r \ge 1$ and $R > 0$, it has a cover by balls of Frobenius radius η with at most

$$\mathsf{N}_\eta(\mathcal{L}_r(R)) \le (1 + 6R/\eta)^{r(d_k+d_v+1)}. \tag{18}$$

Here cover centers may lie in the ambient matrix space. To verify the count, factor $S = U \operatorname{diag}(\sigma) V^\top$ with orthonormal padded columns, $\|\sigma\|_2 \le R$, and nonnegative σ. Cover the operator-norm unit balls for U and V at radius $\eta/(3R)$ and the Euclidean r-ball for σ at radius $\eta/3$. The volumetric bounds are $(1 + 6R/\eta)^{d_k r}$, $(1 + 6R/\eta)^{d_v r}$, and $(1 + 6R/\eta)^r$. Centers can be chosen inside their respective balls. Replacing the three factors one at a time bounds the total Frobenius error by $\eta/3 + \eta/3 + \eta/3$. Setting $\eta = \rho/2$ bounds $> \rho$ packing and gives the second term of (10). The first is the ambient ball bound. If $R = 0$ or $r = 0$, the set is a singleton.

D.2 Uniform recent-write approximation

For $\gamma < 1$, retain only the final $\min\{m, H\}$ summands of the unrolled recurrence, calling the result $S_m^{[H]}$. It belongs to $\mathcal{L}_{r_H}(R_H)$ and

$$\|S_m - S_m^{[H]}\|_F \le \kappa\nu \sum_{j=H}^{m-1} \gamma^j \le \tau_H = \frac{\kappa\nu\gamma^H}{1-\gamma}.$$

An empty tail contributes zero. Images of a $> \rho$-separated state set are therefore $> (\rho - 2\tau_H)$-separated under this approximation. When $\rho > 2\tau_H$, apply (18) at half that separation to prove (11). The approximating set is common to all m, so here the bound is valid for the union over lengths. The ambient bound $d_k d_v \log_2(1 + 2\kappa\nu/((1-\gamma)\rho))$ also applies to that union.

Uniform $\gamma < 1$ is additional to the finite-length rank result. A sigmoid gate can be below one at every sampled position while having supremum one on the admissible domain. Moreover, contraction in a recurrent matrix with keys and values held fixed is not contraction of the full multi-layer nonlinear state update. Either a domain-wide GLA component bound with complete-state assembly, or a valid full-state stability bound, is needed for a checkpoint-level theorem.

D.3 Assembling the complete state

Suppose block b is within distance τ_b of a set $\mathcal{L}_{r_b}(R_b)$, and any auxiliary field is within τ_{aux} of a declared set $\mathcal{K}$. Use the concatenated Euclidean metric and let $\tau_{\text{tot}}^2 = \sum_b \tau_b^2 + \tau_{\text{aux}}^2$. For positive radii $\eta_b, \eta_{\text{aux}}$ satisfying

$$\sum_b \eta_b^2 + \eta_{\text{aux}}^2 \le \left(\frac{\rho - 2\tau_{\text{tot}}}{2}\right)^2, \qquad \rho > 2\tau_{\text{tot}},$$

a product cover proves

$$\mathsf{Pack}_{>\rho}(\mathcal{M}) \le \left[\prod_b (1 + 6R_b/\eta_b)^{r_b(d_{k,b}+d_{v,b}+1)}\right] \mathsf{N}_{\eta_{\text{aux}}}(\mathcal{K}). \tag{19}$$

Indeed, the product center is at distance at most $\tau_{\text{tot}} + (\sum_b \eta_b^2 + \eta_{\text{aux}}^2)^{1/2} \le \rho/2$ from each represented state. Omit an auxiliary factor if no such field exists, and treat singleton blocks as factors of one. Discrete auxiliary states can instead be covered exactly, with their cardinality multiplying the result. No independence of layers or heads is assumed.

This construction explains why multiplying component packing numbers at the same radius is generally invalid under a Euclidean product metric. It also makes the scope of a GLA specialization explicit: every prompt-dependent component must be included. The empirical codec changes the complete state passed across the cut, but does not supply the domain-wide constants in this proposition.

E EXPERIMENTAL DESIGN AND SUPPORTING RESULTS

E.1 COMMON PROTOCOL AND CAUSAL-CUT VALIDATION

Checkpoints and numerical conventions. The primary checkpoints are `fla-hub/gla-340M-15B` and `fla-hub/gla-1.3B-100B`, with frozen revisions:

GLA-340M	`6e04029dc090a2c55df712f18814db80aa39894f`
GLA-1.3B	`46b15820a4df269e99aed9d709e017677c15d24b`

The same-weight attention comparison uses Qwen3-0.6B (Yang et al., 2025a). All model weights and tokenizers are frozen within their comparisons. The isolated GLA runtime uses Flash Linear Attention 0.5.1, PyTorch 2.8.0, Triton 3.4.0, and Transformers 4.48.2. Both GLA protocols use the fixed public newline token 13 after the cut, without a supplied target token. No inference about parameter scaling is drawn from different tokenizers, training histories, or search schedules across checkpoints.

Targets are random word sequences drawn reproducibly from the same ordered 100K-word list used in the original experimental pipeline. The deterministic SHA256-counter recipe is keyed by target-set identifier, seed, requested stream length, and target index. Words are joined with spaces, tokenized by the frozen tokenizer, and truncated to the requested token length; insufficient tokenizations are rejected. The supplementary code includes the target-construction procedure and example configurations. The two primary GLA length sweeps use common maximum-length streams ($N = 5120$ and 13312) and exact prefixes at shorter lengths. Initial prompt hashes match across channels within each target and length. The matched prompt-budget comparison, geometry control, longer-search control, and crossed surface each use a separate target set; populations are not pooled.

Prompt optimization. GLA prompt vectors are FP32, initialized from a seeded centered Gaussian scaled by the checkpoint's input-embedding standard deviation. The frozen model weights are BF16 and the retained recurrent state is FP32. AdamW uses learning rate 0.01, $(\beta_1, \beta_2) = (0.9, 0.9)$, and weight decay 0.01. Gradients pass through ingestion, the projection and reconstruction map, and teacher-forced continuation. Only the prompt is updated. Qwen uses the same AdamW hyperparameters, FP32 prompts, BF16 weights, and a 5K-update cap with patience 2000. The fixed primary schedules are summarized in Table 1.

Table 1: Primary search schedules. Counts are independent target identities per cell; restarts and checkpoints are not additional targets.

Comparison	Targets	Restarts	Evaluation updates
GLA-340M rank–length	12	1	6K, 10K, 20K
GLA-1.3B rank–length	24	1	4K, 7K, 10K
GLA-340M crossed prompt–memory control	8	1	6K, 10K, 20K
GLA-340M paired budget confirmation	24	2	6K, 10K, 20K
GLA-340M equal-size geometry	24	2	6K, 10K, 20K
GLA-340M longer random search	24	2	20K, 40K, 60K
Qwen3-0.6B window envelope	12	1	up to 5K

The earlier dense-state optimization control uses 20 targets per length and separate 6K-, 10K-, and 20K-update runs, rather than three autonomous evaluations along one trajectory (Appendix E.2.5).

Autonomous saved-state evaluation. At each declared endpoint, the runner ingests the optimized prompt up to the cut, which precedes the fixed public delimiter, and saves every prompt-dependent field needed downstream. It then deletes the prompt and all prompt-derived activations outside that state, restores the saved state, and supplies the public delimiter. Greedy generation must reproduce every target token, including the first, without receiving a target prefix. For the primary GLA projection experiments, a separate target-blind worker receives the saved channel, frozen model, and public protocol, but not the prompt, target, optimizer state, or dense precursor of a compressed channel. In-process and restored-worker outputs agree at accepted endpoints. Model-freezing, basis identity, target identity, payload, serialization, and output checks are recorded. The adapter is checked for deterministic resets, complete-state enumeration, and round-trip reconstruction.

All included primary GLA and Qwen results pass the stated acceptance checks; jobs, restarts, and checkpoints are not additional independent target identities.

Floating-point evaluation is protocol-dependent. The Qwen envelope fixes one-shot ingestion rather than requiring bitwise agreement between arbitrary alternative chunkings. Likewise, teacher-forced and recurrent implementations must not be assumed numerically identical merely because their real-arithmetic formulas agree. Final exactness is determined by the declared native saved-state computation.

When teacher forcing implies autonomous exactness. Teacher forcing is appropriate for optimization and is not itself evidence of leakage. If every teacher-forced next-token prediction, including the first, matches the target under exactly the same deterministic computation used in generation, induction shows that autonomous generation matches too. Supplying an unchecked first target token, retaining an undeclared prompt route, or using a numerically different parallel kernel does not satisfy this equivalence automatically. We therefore evaluate the declared native pathway directly and distinguish historical teacher-forced flags from whole-target saved-state generation.

E.2 GLA MEMORY INTERVENTIONS AND CONTROLS

E.2.1 PROJECTION CHANNEL AND COMPLETE-STATE PAYLOAD

The theoretical S in (9) uses key-by-value orientation. The runtime stores its transpose, which we call $H_{\ell h} \in \mathbb{R}^{d_v \times d_k}$ to avoid conflating conventions. For a fixed orthonormal key-side basis $U_{\ell h} \in \mathbb{R}^{d_k \times d_k}$, let $U_{\ell h}^{(r)}$ contain its first r columns. The experimental channel is

$$Z_{\ell h}^{(r)} = H_{\ell h} U_{\ell h}^{(r)}, \qquad \widetilde{H}_{\ell h}^{(r)} = Z_{\ell h}^{(r)} (U_{\ell h}^{(r)})^{\top}. \tag{20}$$

It is applied simultaneously to every layer and head. Only the coefficients $Z^{(r)}$ cross the cut in projected arms. The random control preserves prompt-dependent coefficients along random directions; it does not replace memory by independent random values.

Both GLA checkpoints have 24 layers and four heads per layer. The per-head runtime state shapes are 256×128 for GLA-340M and 512×256 for GLA-1.3B. At batch size one, each checkpoint retains 24 FP32 tensors containing these four head matrices. All remaining continuation metadata in this protocol is fixed or determined by the public prompt length and decoding position; no additional prompt-content-dependent tensor is passed to the worker. With four bytes per FP32 value,

$$B_{\text{coefficients}}(r) = 24 \cdot 4 \cdot d_v \cdot r \cdot 4 \quad \text{bytes.}$$

Table 2 distinguishes causal tensor payload from the fixed basis. Dense and projected ingestion can use substantially more working memory; the experiment does not claim an HBM or throughput saving.

Table 2: Saved-state payload and shared basis storage. MiB denotes 2^{20} bytes. The basis is fixed read-map metadata, not prompt-dependent state, but remains part of the operational memory ledger.

Checkpoint	Dense state	Full basis	Projected rank: payload
GLA-340M	12 MiB	6 MiB	32/48/64/80/96: 3/4.5/6/7.5/9 MiB
GLA-1.3B	48 MiB	24 MiB	64/128/192: 12/24/36 MiB

For GLA-340M, the distribution-informed basis diagonalizes, block by block,

$$C_{\ell h} = \sum_i w_i \frac{H_{i,\ell h}^{\top} H_{i,\ell h}}{\|H_{i,\ell h}\|_F^2}.$$

The 408 deduplicated source states span 48 earlier prompt-budget, length, and target-set cells, with equal total weight per cell and equal weight per state within a cell. Eigenvectors are ordered by decreasing eigenvalue. The GLA-340M calibration targets are disjoint from those used in the 24-target equal-size geometry comparison and the 12-target rank–length sweep, and the basis is held fixed throughout both evaluations. The GLA-1.3B basis is similarly frozen from earlier calibration states, whose targets do not overlap the 24 confirmation targets. This use of empirical state geometry selects a channel; it does not provide the global ellipsoid enclosure required by Proposition 4.3.

Table 3: GLA-340M strict coverage at $m = 512$, out of 12 targets. A dagger marks a right-censored evaluated frontier.

n	rank 32	rank 48	rank 64	rank 80	rank 96	dense
3072	8/12	12/12	12/12	12/12	12/12	12/12
4096	0/12	3/12	10/12	12/12	12/12	12/12
5120	0/12	0/12	0/12	2/12	10/12	8/12
L_{50}	3072	3072	4096	4096	$5120^{\dagger}$	$5120^{\dagger}$

Each random construction uses frozen-seed FP64 Gaussian QR independently for every layer and head, with deterministic sign fixing and FP32 storage. Orthogonality is checked. Rank 32 and rank 64 are nested prefixes of the same basis in each corresponding primary arm. A single global random construction is used in the primary equal-size comparison; four additional independently generated constructions are evaluated in the longer-search control.

E.2.2 GLA-340M RANK–LENGTH FRONTIER

Table 3 gives the complete primary surface at $m = 512$. A target counts when its one declared trajectory is strict-exact at any of 6K, 10K, or 20K updates. Projected-channel success sets are nested in the observed results: no target found at one projected rank is lost at the next larger projected rank. This is an empirical observation, not a general monotonicity theorem. At length 5120, rank 96 reaches two targets not found by dense identity. Projection changes both the retained interface and the optimization problem; this comparison does not establish an ordering of intrinsic accessibility between the two systems.

E.2.3 GLA-1.3B REPLICATION

Figure 6 shows the larger-checkpoint replication of the main GLA-340M frontier. The seven lengths in Table 4 are exact prefixes of 24 common 13312-token streams. Targets count if exact at any of 4K, 7K, or 10K. Every channel falls below majority coverage within the evaluated grid. Rank-192 coverage exceeds dense coverage at some lengths; this is consistent with differences in optimization difficulty, but does not diagnose their cause or identify an architectural cap.

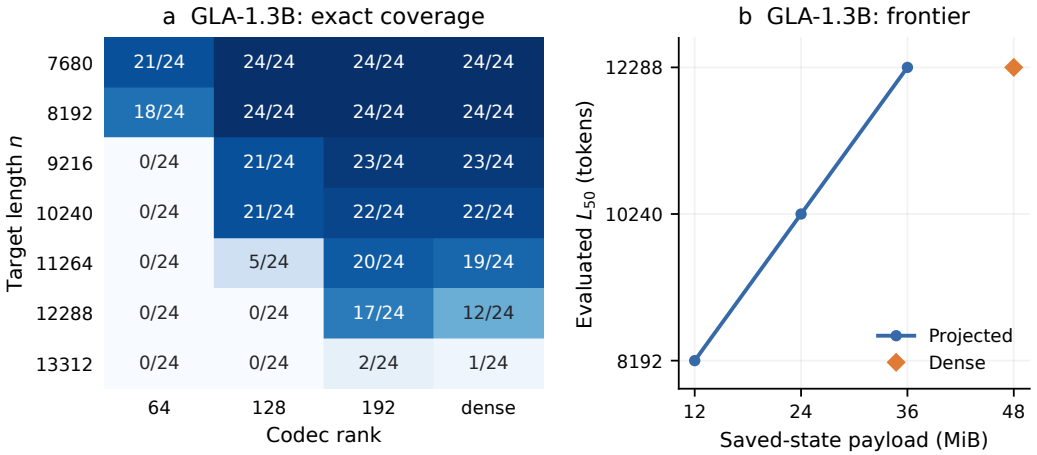


Figure 6: **The memory frontier replicates on GLA-1.3B.** At fixed prompt budget $m = 512$, 24 fresh targets are paired across four channels and seven continuation lengths, with up to 10K optimization updates. **Left:** exact saved-state coverage. **Right:** the evaluated L_{50} rises from 8192 to 10240 to 12288 tokens as projected payload increases from 12 to 24 to 36 MiB; the dense 48-MiB state also reaches 12288. Every channel falls below 50% coverage within the tested grid. Lines connect evaluated conditions only.

Table 4: GLA-1.3B strict coverage at $m = 512$, out of 24 targets.

n	rank 64 (12 MiB)	rank 128 (24 MiB)	rank 192 (36 MiB)	dense (48 MiB)
7680	21/24	24/24	24/24	24/24
8192	18/24	24/24	24/24	24/24
9216	0/24	21/24	23/24	23/24
10240	0/24	21/24	22/24	22/24
11264	0/24	5/24	20/24	19/24
12288	0/24	0/24	17/24	12/24
13312	0/24	0/24	2/24	1/24
L_{50}	8192	10240	12288	12288

The first and second execution cohorts each contain 12 fresh targets. The three projected-channel L_{50} values are $(8192, 10240, 12288)$ in both cohorts. Dense identity reaches 12288 in the first cohort and 11264 in the second, with combined coverage exactly $12/24$ at 12288. Thus its combined majority endpoint is particularly sensitive to the chosen threshold and finite target set.

E.2.4 Prompt-budget interaction

The matched comparison fixes $n = 4096$ and uses the same 24 target identities across $m = 256, 512, 1024$ and four channels. Each condition uses two paired initializations, a 20K-update schedule, and strict saved-state evaluation at 6K, 10K, and 20K. Target-token hashes agree across all three prompt budgets and channels. The $m = 1024$ arm was a prespecified prompt-budget replication of $m = 512$; the $m = 256$ arm was a later matched extension using the same targets, frozen bases, and search schedule. Neither adds an independent target population. Initializations are paired across channels within a prompt budget; tensors at different prompt lengths have different shapes. Table 5 counts a target if any declared restart and checkpoint succeeds.

Table 5: Matched GLA-340M prompt-budget comparison at $n = 4096$, out of 24 targets. The later $m = 256$ extension completes the same-target comparison.

m	rank 32	rank 64	rank 96	dense
256	0/24	0/24	0/24	0/24
512	0/24	20/24	24/24	24/24
1024	16/24	24/24	24/24	24/24

Both budgets matter: increasing the prompt can make a fixed channel more effective, while increasing retained memory gives no observed gain at $m = 256$. These are discovered-coverage comparisons under the stated search, not certificates that a failed target is inaccessible.

Crossed control. A separate eight-target comparison crosses two prompt budgets and three lengths, using common 6144-token streams, one matched initialization per cell, and the same 20K schedule (Table 6). At $m = 512, n = 4096$, all eight discordant target pairs favor rank 64 or dense over rank 32. Each unadjusted two-sided exact McNemar p-value is 0.0078125; Holm correction over the full 18-comparison family gives 0.140625. This secondary surface is descriptive, not a multiplicity-corrected interaction discovery.

E.2.5 Dense-state prompt and search-budget controls

This control holds the dense GLA-340M channel fixed at 12 MiB and uses 20 targets per length, paired across prompt budgets and update endpoints. Targets at different lengths are distinct, not common prefixes. The 6K-, 10K-, and 20K-update conditions are separate runs from the same initialization within each prompt–target pair, rather than checkpoints along one trajectory. Only the final native endpoint of each run determines success.

Each run saves all 24 FP32 recurrent tensors after prompt ingestion, deletes the prompt, and restores the state for greedy generation from the public delimiter without a supplied target token. This control

Table 6: Crossed GLA-340M prompt-budget control, out of eight targets.

m	n	rank 32	rank 64	dense
256	2048	6/8	8/8	8/8
256	4096	0/8	0/8	0/8
256	6144	0/8	0/8	0/8
512	2048	8/8	8/8	8/8
512	4096	0/8	8/8	8/8
512	6144	0/8	0/8	0/8

Table 7: Dense GLA-340M budget control: exact autonomous matches out of 20 paired targets at each length. The last column counts target identities successful in at least one of the three budgeted runs, not independent successes.

m	n	6K updates	10K updates	20K updates	Any run
256	4096	0/20	0/20	0/20	0/20
512	4096	15/20	20/20	18/20	20/20
1024	4096	20/20	19/20	20/20	20/20
512	8192	0/20	0/20	0/20	0/20
1024	8192	0/20	0/20	0/20	0/20
1024	16384	0/20	0/20	0/20	0/20

uses in-process restoration rather than the separate target-blind worker; a second fresh restoration reproduces the output.

At length 4096, increasing m from 256 to 512 changes discovered coverage from $0/20$ to $20/20$; increasing to 1024 adds no exact targets at either 4096 or 8192. No intermediate lengths between 4096 and 8192 were tested, so this is a no-gain region on the evaluated grid, not a precise frontier. Nor does failure imply lack of optimization progress: at length 8192, the mean best matching prefix across endpoints rises from 1840.70 to 4952.35 tokens when m increases from 512 to 1024, without completing a target. Native exactness is nonmonotone across update budgets; 112 runs are native-exact versus 113 under uncached teacher forcing, and the discrepant run remains a native failure. The control does not establish convergence or identify the active theoretical bound.

E.2.6 Equal-size geometry and search controls

Figure 7 brings together the primary equal-size comparison and the independent longer-search control. Table 8 reports the primary equal-size comparison. A target counts if either of its two trajectories is exact at any declared checkpoint. Exact two-sided 95% Clopper–Pearson intervals treat targets, not checkpoints or restarts, as the sample units.

Table 8: GLA-340M geometry control at $m = 512, n = 4096$, out of 24 fresh targets.

Channel	Payload	Coverage
Informed rank 32	3 MiB	0/24
Random rank 32	3 MiB	0/24
Informed rank 64	6 MiB	23/24
Random rank 64	6 MiB	1/24
Dense	12 MiB	24/24

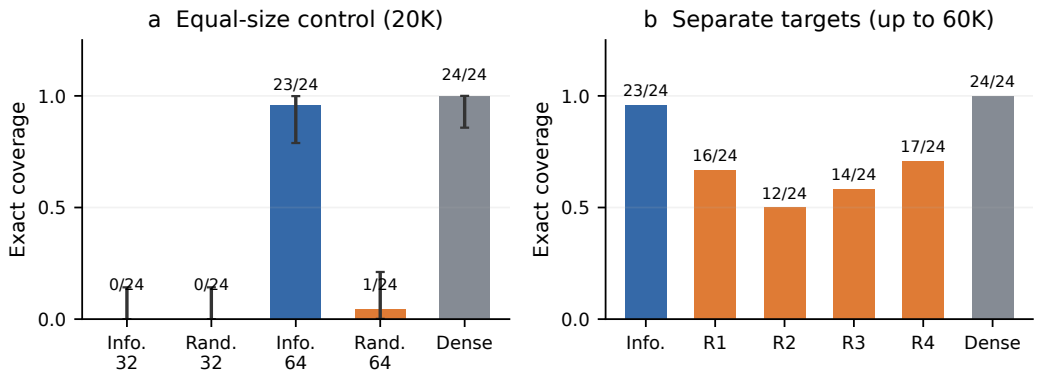


Figure 7: **Equal-size GLA channels can behave differently. Left:** informed and random rank-64 channels retain the same 6-MiB payload but reach $23/24$ and $1/24$ targets under the primary 20K schedule. Error bars are two-sided 95% Clopper–Pearson intervals. **Right:** on a separate target set, four random bases improve substantially with search through 60K updates. Each panel counts a target if any declared restart and checkpoint succeeds; the panels are not a paired comparison.

The four prespecified paired comparisons are shown in Table 9. For rank-64 informed versus random, one target succeeds in both arms, 22 only in the informed arm, none only in the random arm, and one in neither. The coverage difference is 0.9167, with Newcombe paired 95% interval $[0.6893, 0.9651]$ (Newcombe, 1998). The two-sided exact McNemar test, Holm-adjusted across the declared four-test family, gives $p = 1.43 \times 10^{-6}$. The rank-32 floor and the small dense-versus-informed difference are unresolved comparisons, not evidence of equivalence.

The same-prompt diagnostic selects the 45 of 48 dense trajectories that are strict-exact at 20K. None remains exact under any of the four tested compressed channels without reoptimization. Its unit is a selected trajectory, not an independent target: the $23/24$ informed-channel result instead counts targets after channel-specific reoptimization. Thus that result is not obtained simply by compressing a dense-optimized prompt.

The separate longer-search experiment uses 24 new targets, two initializations, four random rank-64 bases, and informed and dense controls. Table 10 distinguishes checkpoint-specific coverage from the union across the schedule. In particular, the fourth random basis has $17/24$ in the union but $16/24$ at the final checkpoint.

Table 9: Declared paired geometry tests. Discordant counts list first-only and second-only successes. Holm adjustment uses all four comparisons.

Comparison	Discordant counts	Raw p	Holm p
Informed 64 vs. informed 32	$(23, 0)$	$2.38\,10^{-7}$	$9.54\,10^{-7}$
Informed 32 vs. random 32	$(0, 0)$	1	1
Informed 64 vs. random 64	$(22, 0)$	$4.77\,10^{-7}$	$1.43\,10^{-6}$
Dense vs. informed 64	$(1, 0)$	1	1

Table 10: Longer GLA-340M search at $m = 512, n = 4096$, out of 24 targets. Each entry unions the two restarts at the indicated checkpoint.

Channel	20K	40K	60K	Any checkpoint
Informed rank 64	19	21	23	23
Random basis 1	0	9	16	16
Random basis 2	1	6	12	12
Random basis 3	1	10	14	14
Random basis 4	0	12	16	17
Dense	24	24	24	24

E.3 SWA NATIVE-MEMORY COMPARISON

We fix one Qwen3-0.6B weight set and a prompt budget of $m = 1000$, and vary the attention window over $W \in \{16, 32, 64, 128, 256, 512\}$. Table 11 gives the refined 50% crossings. The lower endpoint is the longest tested length with at least $6/12$ exact native generations; the upper endpoint is the shortest tested length with fewer than $6/12$. These brackets describe the evaluated search, not confidence intervals or proofs of inaccessibility at greater lengths. Full attention reaches $11/12$ at the largest tested length, 2048, and remains right-censored.

Fixed evaluation protocol. The extension preserves the original Qwen3-0.6B checkpoint, tokenizer, target recipe, and initialization rule. All 28 layers use the native sliding-window implementation with window W during both prompt ingestion and continuation; the same-weight control uses full attention in all 28 layers. The checkpoint's rotary positional encoding configuration is unchanged, and restoration preserves the native cache's cumulative position rather than resetting positions at the cut. The runtime uses PyTorch 2.8.0 and Transformers 5.10.2 with scaled dot-product attention, BF16 frozen weights, and FP32 prompts. The 12-target construction uses seed 2026081622. At each length, the 12 target identities and initialization seeds are paired across evaluated windows and the full-attention control. The recipe includes the requested length, so pairing at one length does not imply common target prefixes across lengths. Each target uses one initialization, AdamW with the hyperparameters above, at most 5000 updates, post-update evaluation every 50 updates, and patience 2000. Generation starts from the fixed public newline token 198 after one-shot ingestion, serialization, prompt deletion, and restoration of the native cache. No target token is supplied to this generation path.

Retained payload. The saved-state tensor payloads are 1,720,544, 3,555,552, 7,225,568, 14,565,600, 29,245,664, and 58,605,792 bytes for windows 16 through 512, respectively. At the cut, each of 28 sliding-window layers retains $W - 1$ positions in both its key and value tensors, together with its 8-byte window tensor. The resulting ledger count is $28[2 \cdot 8 \cdot (W - 1) \cdot 128 \cdot 2 + 8]$ bytes. The same-weight full-attention cache contains 114,688,000 tensor bytes at $m = 1000$. These counts describe the logical tensor payload exposed to continuation, not the serialized file size or total working memory; full-attention payload continues to grow with prompt length.

Adaptive length selection. The search expands to bracket the 50% crossing and then refines it on a two-token grid. After refining W16, a nonnegative-slope affine model fitted to the measured crossing intervals proposes probes for larger windows; ineligible or inconsistent proposals fall back

Table 11: Same-weight Qwen3-0.6B refined crossings at $m = 1000$. Counts are out of 12 targets. Full-attention controls are paired at each endpoint; its frontier remains unreached at 2048.

W	Lower n	SWA	Full	Upper n	SWA	Full
16	64	8/12	12/12	72	2/12	12/12
32	126	7/12	12/12	136	3/12	12/12
64	252	10/12	10/12	262	5/12	12/12
128	476	8/12	12/12	484	4/12	12/12
256	862	9/12	11/12	872	3/12	12/12
512	1844	6/12	11/12	1856	4/12	12/12

to local bisection or outward expansion by a factor of 1.2. Predictions choose measurements, not endpoints, and all 12 targets and paired controls are completed at each selected length. Validated cells and shared controls are reused, not pooled with new repetitions.

The final widths are at most ten tokens for W16–W256. W512 reaches the ten-new-cell stopping limit with a 12-token bracket. Figure 8 retains every evaluated SWA cell and its paired full-attention count. Observed nonmonotonic success fractions are retained. The brackets show an approximately proportional response but do not admit one exact affine law. The descriptive curve in Figure 3 fits $n = cW^{\alpha}$ to arithmetic bracket midpoints by unweighted least squares in logarithmic coordinates, giving $c \approx 5.03$ and $\alpha \approx 0.940$. It is restricted to the measured window range and does not identify an architectural limit.

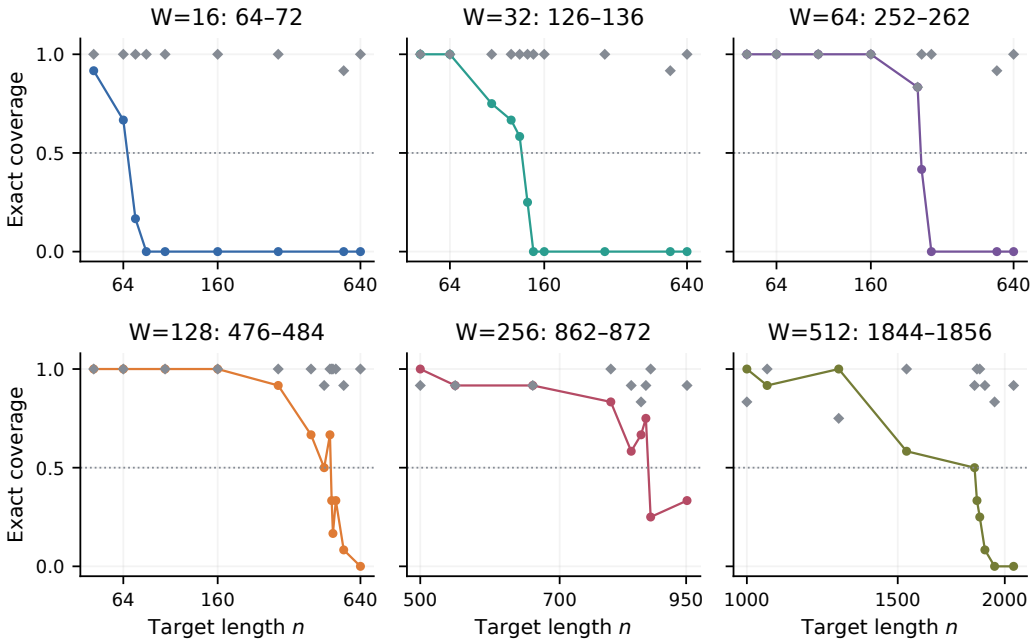


Figure 8: **Complete observed SWA coverage.** Colored circles show native SWA exactness and gray diamonds the paired full-attention control at each tested length. Lines join observed SWA cells only as a visual guide; lengths are not matched across all windows. The dotted line marks $6/12$. Panel titles give the final measured crossing brackets.

A separate retention-only control keeps the W64 write/read mechanism fixed while retaining either 32 or 64 prompt positions; its $5/5$ exact outcomes in both arms do not establish memory sufficiency.

E.4 Mamba: prompt response at fixed recurrent memory

Mamba provides a second recurrent architecture in which to examine the interaction between prompt budget and fixed retained memory. Historical and extension sweeps vary prompt length on Mamba-130M, 370M, and 790M while leaving each checkpoint's recurrent-state dimensions unchanged. These sweeps measure teacher-forced recovery. Separate Mamba-130M saved-state experiments establish the complete memory interface and examine autonomous generation through it. The curves provide complementary evidence about prompt–memory interaction; the controlled changes to memory size remain the GLA and SWA comparisons.

E.4.1 Teacher-forced prompt-length response

The legacy objective supplies the first target token and scores the remaining $n - 1$ tokens under teacher forcing. It therefore differs from whole-target generation across the causal cut. We consistently include equality at 50%: a length qualifies with at least ten successes among the planned 20 targets. Figure 9 and Table 12 show the measured crossings for all three checkpoints, with histor-

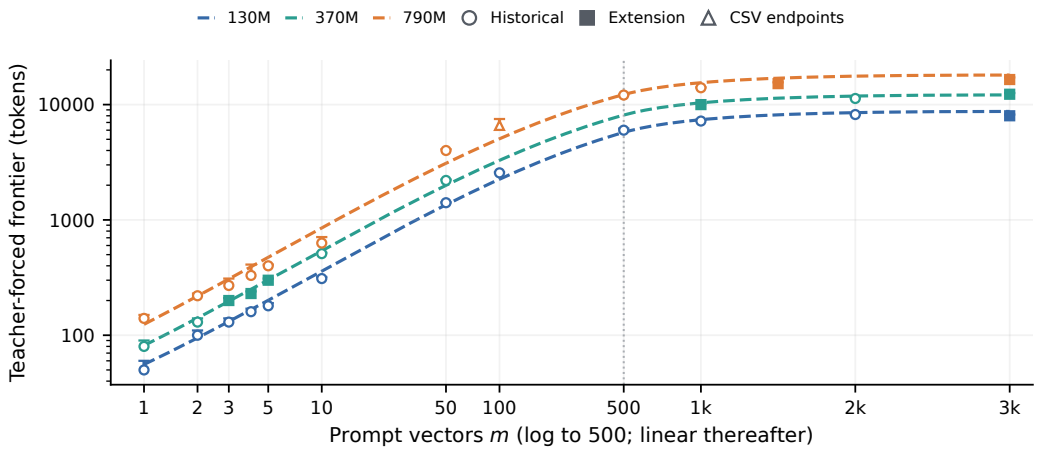


Figure 9: **Prompt response at fixed recurrent-state dimensions.** Each color denotes a frozen Mamba checkpoint. Markers give qualifying lengths and vertical segments extend to the tested non-qualifying endpoint; these are search brackets, not confidence intervals. Open circles are historical results, filled squares the fixed-5000-update extension, and the open triangle completed CSV endpoints from an unfinished historical search. The vertical axis is logarithmic. The horizontal axis is logarithmic through $m = 500$ and continues linearly thereafter, with a continuous slope at the dotted divider. This retains small-budget resolution while giving the larger budgets more horizontal space. Dashed curves are descriptive monotone cubic fits in logarithmic coordinates, restricted to the measured range. They combine bracket summaries for visualization; trial counts remain separate, and the comparison does not isolate the effect of model size.

ical and extension cohorts identified separately. The two cohorts share the objective and optimizer family, but not a fully matched optimization schedule or target population.

For each checkpoint, the displayed trend minimizes squared deviations from the log of each arithmetic bracket midpoint, with equal weight per bracket. We represent the fitted log length as a cubic Bernstein polynomial of the normalized log prompt budget and constrain its coefficients to be non-decreasing. Monotonicity is a smoothing assumption, not a conclusion from these mixed-cohort measurements. The curves do not extrapolate beyond $m = 3000$ or estimate a limiting capacity. In particular, a flattening fitted curve is not an independent test of a memory ceiling.

Extension protocol. The extension uses frozen FP32 weights and FP32 prompts, with TF32 disabled. AdamW uses learning rate 0.01, $(\beta_1, \beta_2) = (0.9, 0.9)$, and weight decay 0.01. Each target has one initialization, at most 5000 updates, and early stopping only on exact teacher-forced recovery. The actual scored prompt is retained before its optimizer update. Targets use the ordered 100K-word vocabulary, seed 2026091801, and deterministic maximum-length streams of 24000 tokens whose prefixes are paired across prompt budgets and lengths within a checkpoint. Target tokens, hashes, and initialization seeds are recorded. They are not claimed identical to historical targets.

The runtime uses PyTorch 2.8.0 and Transformers 5.10.2, with gradient checkpointing and a 512-token chunked output-head loss. Accelerated convolution and selective-scan implementations are fixed throughout the extension and checked against the legacy objective and its prompt gradient. These are numerical consistency checks, not a claim of bitwise identical optimization trajectories. Historical runs can differ in precision, patience, and update cap; their trials are not pooled with the new cohort.

The checkpoints are `state-spaces/mamba-130m-hf`, `state-spaces/mamba-370m-hf`, and `state-spaces/mamba-790m-hf`.

The high-prompt search expands by a factor of 1.2 and bisects on a 100-token grid until its width is at most the larger of 200 tokens and 2.5% of the qualifying length, subject to eight tested lengths per curve. The low-prompt 370M searches use successive ten-token probes. All observed extension cells, including unresolved intermediate cells, are retained when determining the reported brackets.

Table 12: Selected Mamba teacher-forced crossings. H: historical; E: extension; C: completed CSV endpoints from an unfinished historical search. Counts are observed successes/trials, not imputed 20-target rates. The two partial E endpoints have decisions fixed for the planned 20 targets.

Model	m	Qualifying n	Count	Below-half n	Count	Cohort
130M	1	50	16/20	60	8/20	H
130M	2	100	13/20	110	7/20	H
130M	3	130	11/20	140	7/20	H
130M	4	160	13/20	170	9/20	H
130M	5	180	15/20	190	5/20	H
130M	10	310	11/20	320	9/20	H
130M	50	1410	12/20	1420	8/20	H
130M	100	2560	11/20	2600	9/20	H
130M	500	6000	10/20	6200	3/20	H
130M	1000	7200	10/20	7400	9/20	H
130M	2000	8200	11/20	8240	7/20	H
130M	3000	8000	19/20	8200	9/20	E
370M	1	80	11/20	90	4/20	H
370M	2	130	11/20	140	5/20	H
370M	3	200	10/20	210	9/20	E
370M	4	230	11/20	240	9/20	E
370M	5	300	12/20	310	8/20	E
370M	10	510	13/20	520	8/20	H
370M	50	2200	11/20	2240	7/20	H
370M	1000	10000	16/20	10200	9/20	E
370M	2000	11300	10/20	11500	5/20	H
370M	3000	12300	10/20	12600	7/20	E
790M	1	140	11/20	150	7/20	H
790M	2	220	14/20	230	7/20	H
790M	3	270	10/20	310	7/20	H
790M	4	330	10/20	410	0/20	H
790M	5	400	10/20	410	7/20	H
790M	10	630	10/20	710	3/20	H
790M	50	4000	13/20	4100	8/20	H
790M	100	6600	11/20	7500	4/20	C
790M	500	12100	11/20	12200	8/20	H
790M	1000	14000	11/20	14100	9/20	H
790M	1500	15200	12/19	15600	9/20	E
790M	3000	16500	16/20	17200	3/16	E

The 790M searches stopped at the accepted wider brackets in Table 12; missing outcomes are not converted into successes or failures.

For 790M at $(m, n) = (1500, 15200)$, the 12 observed successes already guarantee at least ten among the planned 20. At $(3000, 17200)$, three successes among 16 observed trials allow at most seven among 20. These are decision-resolved partial cells, not complete success-rate estimates. The $3/6$ observed at $(3000, 16800)$ remains undecided and does not tighten the bracket.

Interpretation at fixed memory. The historical 130M curve rises from a qualifying length of 6000 at $m = 500$ to 7200 at $m = 1000$ and 8200 at $m = 2000$. The new $m = 3000$ bracket, 8000–8200, remains in the same range. This flattening is consistent with diminishing gains at fixed memory, but the cohorts differ and the comparison does not establish either a decline or a strict ceiling. The larger checkpoints still improve: within the extension, the 370M qualifying length rises from 10000 at $m = 1000$ to 12300 at $m = 3000$, and the 790M value rises from 15200 at $m = 1500$ to 16500 at $m = 3000$. The diminishing gains provide complementary evidence about prompt and memory acting together, not identification of memory as the active bound. Optimization limitations and fixed-model expressivity remain alternative explanations. Changing checkpoints also changes weights and state dimensions, so these are not controlled parameter-scaling measurements.

The historical 790M $m = 100$ bracket uses completed CSV endpoints from an unfinished search and is marked separately. Ambiguous or incomplete searches without decision-resolved endpoints remain excluded and are not treated as target failures.

E.4.2 Strict Mamba-130M saved-state experiments

A complete, fixed saved state. The strict experiments use one frozen Mamba-130M checkpoint, BF16 model computation, and FP32 optimized prompt vectors. Each of its 24 layers retains a convolutional tensor of shape $[1, 1536, 4]$ and a recurrent tensor of shape $[1, 1536, 16]$. Together, these 48 tensors contain 1,474,560 bytes, or 1.40625 MiB. This is the prompt-dependent tensor payload, excluding fixed model weights and serialization overhead. Its dimensions and byte count remain unchanged across 16 audited prompt lengths from 1 through 2000.

After one-shot prompt ingestion, the cache is serialized and restored in the same process, and the optimized prompt and optimizer are deleted. Generation starts from a fixed newline delimiter and supplies each predicted token as the next input until the entire declared target length has been generated. Target tokens are used for scoring, not supplied to this generation path. The audit checks all convolutional and recurrent tensors, cache metadata, finite values, serialization agreement, and a fresh reset. A separate tokenwise-ingestion diagnostic can produce different BF16 states, so one-shot ingestion remains the fixed evaluation schedule.

All 16 state-audit jobs pass these integrity checks. They use only one or 20 optimization updates and are technical tests rather than a coverage experiment. A separate positive control crosses $m \in \{16, 64\}$ and $n \in \{4, 8\}$ with two targets and nominal update caps of 500 and 2000. All 16 jobs generate their targets exactly, stopping after 25–75 updates. The paired budget arms stop at the same successful checkpoints, so this establishes that the state-only pathway works rather than a benefit from the larger nominal cap.

Prompt-length pilot and stopping-rule follow-up. The pilot tests $m \in \{16, 64, 256, 512\}$ and $n \in \{160, 320, 640, 1280\}$ with two targets per length, paired across prompt budgets. Prompt optimization uses AdamW with learning rate 0.01, $\beta_1 = \beta_2 = 0.9$, and weight decay 0.01, with evaluations every 25 updates. It allows up to 2000 updates but stops when uncached teacher-forced predictions first become exact. At the selected endpoints, 24 of the 32 jobs are teacher-forced exact, while 10 are exact under saved-state generation.

The follow-up reuses eight selected pilot cells with the same targets and initialization seeds: $m \in \{16, 64\}$ at $n \in \{160, 320\}$ and $m \in \{256, 512\}$ at $n \in \{320, 640\}$. It runs all 2000 updates, evaluates native generation at updates 250, 500, 1000, and 2000, and also evaluates the checkpoint with the lowest recorded teacher-forced loss. Table 13 summarizes these diagnostics. Denominators count target–condition jobs, not independent targets. Because only selected cells were rerun, the pilot and follow-up totals are not an unbiased comparison over the same grid.

Both length-640 targets succeed at some evaluated checkpoint for both $m = 256$ and $m = 512$, where the pilot had found no native successes. Exactness can also be lost: both $(64, 320)$ targets

Table 13: Mamba-130M strict-state diagnostics. Each cell contains two targets, reused across prompt budgets. The follow-up covers eight selected pilot cells. Best loss is selected by teacher-forced loss; Any unions the fixed and selected checkpoints within a job.

Experiment	Evaluation rule	Exact jobs
Pilot (16 cells)	Uncached teacher forcing at stopping	24/32
Pilot (16 cells)	Native generation at stopping	10/32
Follow-up (8 cells)	Native at final update 2000	11/16
Follow-up (8 cells)	Native at best loss	13/16
Follow-up (8 cells)	Native at any evaluated checkpoint	14/16

succeed at their selected checkpoints but fail at update 2000. The pilot's apparent boundary is therefore not a stable Mamba frontier.

Cached teacher-forced and autonomous exact-match outcomes agree at every evaluated follow-up checkpoint. At update 2000, however, uncached teacher forcing succeeds in 14 jobs while both cached evaluation and autonomous generation succeed in 11. The distinction is therefore between computation pathways and checkpoint selection, not a general failure of teacher forcing as an evaluation principle. The follow-up establishes successful whole-target generation through a fixed Mamba state and motivates evaluating the declared native pathway directly.

Evidence scope and exclusions. The teacher-forced curves and strict diagnostics contribute different evidence. The former show prompt-length responses in a non-attention architecture; the latter demonstrate a complete fixed-state interface and exact generation through it. These experiments hold the native Mamba state dimensions fixed within each checkpoint. A fresh-target confirmation with a controlled change to those dimensions would be a separate memory-size experiment.

Historical trial counts are not pooled with the extension, and technical state audits are not used as coverage estimates.

E.5 Statistical conventions and reproducibility scope

The reported proportions summarize a specified finite target population and search protocol. Binomial intervals are descriptive uncertainty summaries over independently constructed target streams; they are not simultaneous confidence bands over all lengths, ranks, or prompt budgets. The primary geometry test uses paired target outcomes and its declared four-test family. No p-value is inferred from treating optimizer checkpoints or restarts as independent observations.

The evaluated L_{50} is

$$\widehat{L}_{50}(m, c) = \max\{n \in \mathcal{N}_{\text{tested}} : \widehat{C}(m, n, c) \geq 1/2\},$$

when the set is nonempty. Reaching the longest tested length is right censoring. Falling below 50% at a subsequent grid point brackets a response under this experiment; it neither proves global failure at every longer length nor identifies a theoretical capacity. Targets, selected subspaces, optimizer schedules, and decoder conventions are fixed within each stated comparison. Pilots and adaptive searches select feasible regions and new lengths; these finite grids are not unbiased samples of all model configurations.

The supplementary code provides a minimal reference implementation of the causal-cut evaluation framework, target construction, saved-state serialization and restoration, and representative memory interventions. It is intended to make the experimental structure inspectable and support new runs rather than reproduce every historical campaign or archived result. The appendix reports the primary evaluated conditions and result tables; raw manifests, validation records, and saved trajectories remain in the project archive.