LIPMAMBA: LIPSCHITZ-CONSTRAINED SELECTIVE STATE-SPACE MODELS WITH STATE-RETENTION AND PAC-BAYESIAN BOUNDS AGAINST HIDDEN-STATE POISONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective state-space models (SSMs) such as Mamba are efficient linear-time alternatives to attention, but their input-dependent discretisation creates an architecture-specific attack surface: short adversarial token sequences can drive the step size Δ_t toward saturation so that $\bar{\mathbf{A}}_t = \exp(\Delta_t \mathbf{A}) \to 0$, erasing prior context (*hidden-state poisoning*, HiSPA). Existing defences are detectors that offer no analytical guarantee. We introduce LipMamba, a selective-SSM parameterisation with spectrally-normalised projections, an eigenvalue-bounded state matrix and a discretisation step clamped to $[\Delta_{\min}, \Delta_{\max}]$, and prove three results with complete proofs: (i) an explicit Lipschitz constant for the selective (input-dependent) scan on the bounded-input domain, which identifies the lower clamp $\Delta_{\min} > 0$ as the condition that keeps the state bounded and thereby makes precise the error explosion previously observed in input-dependent SSMs; (ii) a state-retention bound giving an analytically computed trigger length $\ell^\star$ below which no trigger can reduce the hidden-state norm below a chosen fraction of its pre-trigger value, which removes the $\bar{\mathbf{A}}_t \to 0$ mechanism of HiSPA; and (iii) a PAC-Bayesian bound on adversarial risk whose gap term is governed by the same constant. We are explicit about scope: the analytical global constant grows geometrically with depth and yields a vacuous global certificate, so per-input radii are reported under a local Lipschitz estimate and are empirical; the retention bound constrains norm, not content, and for the constraint values used here it certifies triggers of one token at 50% retention and of up to six tokens against near-complete erasure. On RoBench-25 and HarmBench, LipMamba-370M raises poisoned accuracy under HiSPA at trigger length 24 from 23.4% (Mamba) to 85.3%, retains 80.4% accuracy at empirical local-Lipschitz radius 0.18, and adds 3.5–4.9% perplexity overhead over the public Mamba checkpoints on WikiText-103. Ablations attribute the gain chiefly to spectral normalisation of the selective projections and to step clamping.

1 INTRODUCTION

Selective SSMs, most prominently Mamba (Gu & Dao, 2024) and Mamba-2 (Dao & Gu, 2024), achieve Transformer-competitive perplexity at linear inference cost by predicting the discretisation parameters $(\Delta_t, \mathbf{B}_t, \mathbf{C}_t)$ from the input at every token. This selectivity is also an attack surface. Mercier et al. (2026b) identified *hidden-state poisoning* (HiSPA): because $\mathbf{A}$ has negative real spectrum by HiPPO initialisation (Gu et al., 2020), a trigger that drives $\Delta_t = \mathrm{softplus}(\mathbf{W}_\Delta \mathbf{x}_t)$ to large values makes $\bar{\mathbf{A}}_t = \exp(\Delta_t \mathbf{A}) \approx 0$, and a handful of attacker-controlled tokens multiplicatively attenuate every prior contribution to the state. Optimised triggers collapse Mamba–Transformer hybrids on long-context retrieval while pure Transformers are unaffected. The existing defences, CLASP (Mercier et al., 2026a) and SpectralGuard (Bonetto, 2026), are detectors that mitigate empirically but bound nothing; all three works are unrefereed 2026 preprints.

Lipschitz-bounded architectures (Tsuzuku et al., 2018; Leino et al., 2021; Wang & Manchester, 2023) turn a logit margin into a perturbation radius, and PAC-Bayes bounds (McAllester, 2003;

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.

Pérez-Ortiz et al., 2021) certify risk over a posterior, but neither has been applied to the selective recurrence: prescribed-L_2 SSMs (Massai et al., 2025; Barbara et al., 2025) cover the linear time-invariant core only, and Qi et al. (2024) showed that input-dependent SSMs are prone to *error explosion* through the discretisation without a parameterisation that removes it.

What this paper delivers, and what it does not.

- **The first closed-form Lipschitz constant for a selective (input-dependent) recurrence** (Theorem 3), in terms of the spectral bounds, the eigenvalue range of $\mathbf{A}$ and a state bound H (Lemma 2). Its content is structural: the lower clamp $\Delta_{\min} > 0$ is exactly what bounds the state, the mechanism behind the error explosion of Qi et al. (2024). As a number it is vacuous at depth, and we say so (Remark 1).
- **A parameterisation that enforces every constant by construction** (Section 4), composed with a GloRo-style head (Leino et al., 2021), at 3.5–4.9% perplexity cost.
- **State-retention bounds** (Theorem 4, Corollary 5) that remove the HiSPA collapse mechanism and certify the trigger length $\ell^\star$ for any retention level, for the norm and for the component along the pre-trigger state. For the trained constants $\ell^\star$ is one token at 50% retention and six against near-complete erasure (Remark 2).
- **A PAC-Bayes adversarial-risk bound** (Theorem 6) coupling the Lipschitz constant with the KL term; we claim no non-vacuous instantiation, because the available constants are either vacuous (global) or not upper bounds (local).
- **An empirical study** on RoBench-25 and HarmBench with a component ablation (Section 6); robustness beyond the certified trigger lengths is measured, not certified.

Code, configurations and the scripts that regenerate every table and figure are provided anonymously (`https://anonymous.4open.science/r/LipMamba-ICLR27`).

2 Related Work

Selective SSMs. HiPPO (Gu et al., 2020), S4 (Gu et al., 2022) and S5 (Smith et al., 2023) established structured state spaces; Mamba (Gu & Dao, 2024) introduced the selective S6 layer and Mamba-2 (Dao & Gu, 2024) the state-space duality; Jamba (Lieber et al., 2024) and Zamba (Glorioso et al., 2024) interleave SSM and attention layers. Robustness studies of visual SSMs (Du et al., 2024; Malik et al., 2025) and SSM-specific backdoors (Wu & Zhang, 2025) show that the selective mechanism is itself an attack surface. Most relevant is Qi et al. (2024), who derive output-error bounds for fixed and input-dependent SSMs, observe that the latter suffer error explosion from the discretisation, and propose Adaptive Scaling, a heuristic rescaling applied during adversarial training that mitigates the explosion empirically; our Theorem 3 gives the constant explicitly and identifies the step clamp as the architectural condition that removes it by construction rather than by training.

Hidden-state poisoning and LLM attacks. Language-model attacks include data and weight poisoning (Gu et al., 2017; Chen et al., 2017; Kurita et al., 2020), adversarial suffixes (Zou et al., 2023b), sleeper agents (Hubinger et al., 2024), activation steering (Zou et al., 2023a; Wang & Shu, 2024) and memory poisoning (Zou et al., 2025; Chen et al., 2024), benchmarked with a hidden-state modality by BackdoorLLM (Li et al., 2025). HiSPA (Mercier et al., 2026b) is inference-time, architecture-specific and black-box; its defences CLASP (Mercier et al., 2026a) (XGBoost on block-output embeddings) and SpectralGuard (Bonetto, 2026) (a spectral-drift detector) are empirical.

Lipschitz architectures and certification. Spectral normalisation (Miyato et al., 2018), Lipschitz-margin training (Tsuzuku et al., 2018), orthogonal and unitary parameterisations (Trockman & Kolter, 2021; Arjovsky et al., 2016), direct Lipschitz-bounded parameterisations (Wang & Manchester, 2023; Pauli et al., 2024), Lipschitz RNNs (Erichson et al., 2021) and prescribed-L_2 SSMs (Massai et al., 2025; Barbara et al., 2025) enforce bounds by construction; GloRo (Leino et al., 2021) and LiResNet (Hu et al., 2023) build the certificate into the head; randomized smoothing (Cohen et al., 2019) certifies probabilistically. For Transformers, attention is not globally Lipschitz (Kim et al., 2021), LipsFormer (Qi et al., 2023) repaired this, and Newhouse et al. (2025) found that the global bound needed to match baseline accuracy at 145M parameters is astronomically large, a caution we take seriously in Remark 1. Certified defences against *training-set* poisoning (Levine & Feizi, 2021; Wang et al., 2022; Rosenfeld et al., 2020) address a different threat.

PAC-Bayes. Non-vacuous bounds for deep networks (Dziugaite & Roy, 2017; Neyshabur et al., 2018; Pérez-Ortiz et al., 2021), adversarial-risk PAC-Bayes (Viallard et al., 2021; Mustafa et al., 2024), LLM-scale bounds via compression (Lotfi et al., 2024) and bounds for stable RNNs (Eringis et al., 2024) are the ingredients of Theorem 6. A PAC-Bayes-inspired objective was used for a poisoning-resilient selective SSM in intrusion detection (Anaedevha et al., 2026), without Lipschitz analysis, step clamp or state bound.

3 Preliminaries and Threat Model

A continuous-time system $\mathbf{h}'(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}u(t)$, $y(t) = \mathbf{C}\mathbf{h}(t)$ discretised by zero-order hold at step Δ gives $\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}u_t$ with $\bar{\mathbf{A}} = \exp(\Delta\mathbf{A})$. In Mamba's S6 layer the parameters are input-dependent,

$$\begin{aligned} &\Delta_t = \mathrm{softplus}(\mathbf{W}_\Delta \mathbf{x}_t + \mathbf{b}_\Delta), \qquad \mathbf{B}_t = \mathbf{W}_B \mathbf{x}_t, \qquad \mathbf{C}_t = \mathbf{W}_C \mathbf{x}_t, \\ &\mathbf{h}_t = \bar{\mathbf{A}}_t \mathbf{h}_{t-1} + \bar{\mathbf{B}}_t \mathbf{x}_t, \qquad \mathbf{y}_t = \mathbf{C}_t^\top \mathbf{h}_t, \end{aligned} \tag{1}$$

with $\bar{\mathbf{A}}_t = \exp(\Delta_t \mathbf{A})$, $\bar{\mathbf{B}}_t = \Delta_t \mathbf{B}_t$ (Mamba's simplified Euler rule for the input matrix, in place of the exact zero-order-hold expression $(\Delta_t\mathbf{A})^{-1}(\exp(\Delta_t\mathbf{A}) - \mathbf{I})\Delta_t\mathbf{B}_t$; all bounds below are for the rule actually implemented), and $\mathbf{A} = -\mathrm{diag}(\exp(\mathbf{a}_{\log}))$ so that $\lambda_i(\mathbf{A}) < 0$.

Dimensions. $\mathbf{x}_t \in \mathbb{R}^D$ (D channels); $\mathbf{B}_t, \mathbf{C}_t \in \mathbb{R}^N$ are shared across channels; $\Delta_t \in \mathbb{R}^D$ and $\mathbf{A} \in \mathbb{R}^{D\times N}$ are per channel; the state $\mathbf{h}_t \in \mathbb{R}^{D\times N}$ has rows $\mathbf{h}_t^{(d)} = \bar{\mathbf{A}}_t^{(d)}\mathbf{h}_{t-1}^{(d)} + \Delta_t^{(d)}\mathbf{B}_t x_t^{(d)}$ and the output $\mathbf{y}_t \in \mathbb{R}^D$ has $y_t^{(d)} = \mathbf{C}_t^\top \mathbf{h}_t^{(d)}$. $\|\mathbf{h}_t\|_2$ denotes the Frobenius norm over (D, N) and $\|\mathbf{x}_t\|_2, \|\mathbf{y}_t\|_2$ Euclidean norms over channels; every bound below is stated for these norms and carries no factor of D (Section B). Section K lists every symbol. Each $\bar{\mathbf{A}}_t^{(d)}$ is real diagonal with entries $e^{\Delta_t^{(d)} a_i} \in (0, 1)$, so its spectral radius and extreme singular values coincide with $\max_i$ and $\min_i e^{\Delta_t^{(d)} a_i}$; we use this repeatedly.

Definition 1 (Hidden-state poisoning (Mercier et al., 2026b)). A trigger $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_\ell)$ of tokens inserted at position t_0 is an (α, ℓ)-poisoning attack if the post-trigger state satisfies $\|\mathbf{h}_{t_0+\ell}\|_2 / \|\mathbf{h}_{t_0}\|_2 \le \alpha$ with $\alpha \ll 1$.

Threat model. The attacker knows the architecture and modifies neither weights nor training data. We consider two access levels: a *query-only* adversary, who observes outputs (Z-HiSPA), and a *white-box* adversary with gradient access (M-HiSPA, PGD). Our bounds address two perturbation surfaces separately: *discrete token insertion* (Definition 1; Theorem 4 and Corollary 5) and *continuous ℓ_2-bounded perturbation* of the scan input (Theorems 3 and 6); token insertions are not ϵ-small in any continuous norm. Definition 1 concerns state *norm*; a large-norm state overwritten with adversarial content is not excluded by it (Remark 2).

Problem. Find a parameterisation $\boldsymbol{\theta} \in \Theta_{\mathrm{Lip}}$ with $\|\mathbf{W}_B\|_2 \le s_B$, $\|\mathbf{W}_C\|_2 \le s_C$, $\|\mathbf{W}_\Delta\|_2 \le s_\Delta$, $\lambda_i(\mathbf{A}) \in [-\lambda_{\max}, -\lambda_{\min}]$ and $\Delta_t \in [\Delta_{\min}, \Delta_{\max}]$, that minimises clean risk with clean-perplexity overhead below 5%, subject to (a) an explicit Lipschitz constant for continuous input perturbations, (b) a state-retention length $\ell^\star$ for token insertions, and (c) a PAC-Bayes bound on adversarial risk under continuous perturbations.

4 The LipMamba Parameterisation

Spectral projection. Writing $\sigma_{\max}(\cdot)$ and $\sigma_{\min}(\cdot)$ for the largest and smallest singular values, so that $\|\mathbf{W}\|_2 = \sigma_{\max}(\mathbf{W})$, we set at every step $\bar{\mathbf{W}}_\bullet = \mathbf{W}_\bullet \cdot \min(1, s_\bullet / \sigma_{\max}(\mathbf{W}_\bullet))$ for $\bullet \in \{B, C, \Delta, \mathrm{out}\}$, which guarantees $\|\bar{\mathbf{W}}_\bullet\|_2 \le s_\bullet$. As in the spectral normalisation of Miyato et al. (2018), $\sigma_{\max}$ is estimated by one power-iteration step per training step, averaged over steps; because power iteration approaches $\sigma_{\max}$ from below, the estimate is inflated by a 2% safety margin before rescaling.

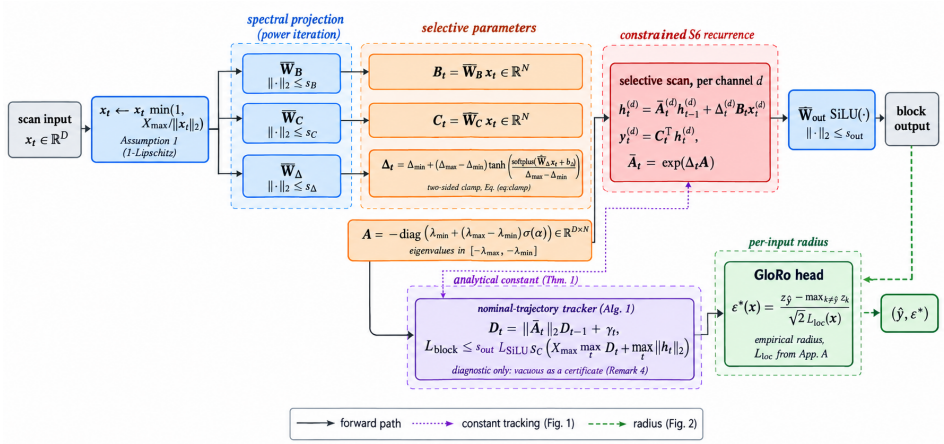

Figure 1: The LipMamba constrained scan. Left to right: the bounded-input projection (Assumption 1); spectral normalisation of $\mathbf{W}_B, \mathbf{W}_C, \mathbf{W}_\Delta$; the two-sided step clamp (2) and the eigenvalue-reparameterised $\mathbf{A}$; the per-channel selective recurrence; and the spectrally normalised readout. Below: the nominal-trajectory tracker of Algorithm 1 (dotted), which is a diagnostic of the analytical constant and not a certificate, and the GloRo head (dashed), whose radius uses the local estimate L_{loc} (Remark 1).

Eigenvalue reparameterisation. $\mathbf{A} = -\text{diag}(\lambda_{\min} + (\lambda_{\max} - \lambda_{\min})\sigma(\boldsymbol{\alpha}))$ with $\boldsymbol{\alpha} \in \mathbb{R}^{D \times N}$ an unconstrained trainable parameter and $\sigma(\cdot)$ the elementwise logistic sigmoid (distinct from the singular values $\sigma_{\max}, \sigma_{\min}$ and the posterior scale σ of (3)), so $\lambda_i(\mathbf{A}) \in [-\lambda_{\max}, -\lambda_{\min}]$ for every value of $\boldsymbol{\alpha}$ and gradient descent cannot leave the constraint set.

Two-sided step clamp. The softplus step is unbounded above and can approach zero; we bound it on both sides,

$$\Delta_t = \Delta_{\min} + (\Delta_{\max} - \Delta_{\min}) \tanh\Big(\frac{\text{softplus}(\bar{\mathbf{W}}_\Delta \mathbf{x}_t + \mathbf{b}_\Delta)}{\Delta_{\max} - \Delta_{\min}}\Big) \in [\Delta_{\min}, \Delta_{\max}), \tag{2}$$

applied per channel, which is smooth and 1-Lipschitz in its pre-activation. Consequently $\rho_{\max} := e^{-\Delta_{\min}\lambda_{\min}} < 1$ upper-bounds $\left\|\bar{\mathbf{A}}_t\right\|_2 = \sigma_{\max}(\bar{\mathbf{A}}_t)$ and $\rho_{\min} := e^{-\Delta_{\max}\lambda_{\max}} > 0$ lower-bounds $\sigma_{\min}(\bar{\mathbf{A}}_t)$ for every input. Both extremes matter: since $\left\|\bar{\mathbf{A}}_t \mathbf{h}\right\|_2 \le \sigma_{\max}(\bar{\mathbf{A}}_t)\, \|\mathbf{h}\|_2$, $\rho_{\max} < 1$ makes every step a contraction and bounds the state (Lemma 2); since $\left\|\bar{\mathbf{A}}_t \mathbf{h}\right\|_2 \ge \sigma_{\min}(\bar{\mathbf{A}}_t)\, \|\mathbf{h}\|_2$, $\rho_{\min} > 0$ caps how much one token can shrink the state, which is what HiSPA exploits (Theorem 4). We set $\Delta_{\min} = 10^{-3}$, the lower end of Mamba's step-initialisation range, so that the lower clamp binds only outside the pretrained step distribution; the price is that $\rho_{\max}$ is close to one and the state bound below is loose (Remark 1).

Bounded inputs. The analysis assumes $\|\mathbf{x}_t\|_2 \le X_{\max}$ at the input of the scan. We enforce it by the Euclidean projection $\mathbf{x}_t \mapsto \mathbf{x}_t \min(1, X_{\max}/\|\mathbf{x}_t\|_2)$; the projection onto a convex set is 1-Lipschitz and changes none of the constants below. On a pretrained checkpoint it rescales activations above $X_{\max}$, a perturbation that fine-tuning absorbs; its cost is included in Fig. 4.

Readout and head. The scan output is $\mathbf{y}_t^{\text{out}} = \bar{\mathbf{W}}_{\text{out}}\, \text{SiLU}(\mathbf{y}_t)$ with $L_{\text{SiLU}} \approx 1.0998$. For classification with logits $\mathbf{z} \in \mathbb{R}^K$, prediction $\hat{y} = \arg\max_k z_k$ and a Lipschitz constant L of the map from input to logits, a GloRo head (Leino et al., 2021) reports the *radius* $\epsilon^\star(\mathbf{x}) = (z_{\hat{y}} - \max_{k \ne \hat{y}} z_k)/(\sqrt{2}L)$: every difference $z_{\hat{y}} - z_k$ is $\sqrt{2}L$-Lipschitz ($\|\mathbf{e}_{\hat{y}} - \mathbf{e}_k\|_2 = \sqrt{2}$), so no input perturbation of norm below $\epsilon^\star(\mathbf{x})$ can close the margin and change $\hat{y}$. Training appends a $(K{+}1)$-th logit $\tilde{z}_{K+1} = \max_{k \ne \hat{y}} z_k + \sqrt{2}L\epsilon_{\text{train}}$ before the cross-entropy, so that classifying correctly against it forces $\epsilon^\star(\mathbf{x}) \ge \epsilon_{\text{train}}$; $\epsilon_{\text{train}} = 0.18$ is the target radius. For the true/false questions of RoBench-25 ($K = 2$) the logits are the language-model logits of the answer tokens "true" and "false" at the answer position. Which L is used is the subject of Remark 1.

Block topology. Theorem 3 covers the scan with SiLU and $\bar{\mathbf{W}}_{\text{out}}$; the retained Mamba components (input projection, convolution, gate, RMSNorm, residual) only enlarge the full-block constant, which strengthens Remark 1 (Section G).

Training objective. Let $\boldsymbol{\theta} = \{\mathbf{W}_B, \mathbf{W}_C, \mathbf{W}_\Delta, \boldsymbol{\alpha}, \mathbf{W}_{\text{out}}\}$ collect the constrained parameters of all blocks and $d \approx 8.7 \times 10^7$ (at 130M) their number. Following PAC-Bayes, $Q = \mathcal{N}(\boldsymbol{\theta}, \sigma^2 I_d)$ is a Gaussian posterior centred at $\boldsymbol{\theta}$ and $P = \mathcal{N}(\boldsymbol{\theta}_{\text{prior}}, \sigma_0^2 I_d)$ a prior fit on a held-out clean split (Pérez-Ortiz et al., 2021). On the training sample S of size n we minimise over $\boldsymbol{\theta}$ (σ, σ_0 fixed)

$$\mathcal{L}(\boldsymbol{\theta}) = \hat{\mathcal{L}}_S^{\text{adv}}(\boldsymbol{\theta}) + \tfrac{1}{2} L_\varphi L_{\text{loc}}(\boldsymbol{\theta})\, \epsilon_{\text{train}} + \beta \sqrt{\frac{\text{KL}(Q\|P) + \ln(2\sqrt{n}/\delta)}{2n}}, \tag{3}$$

where $\hat{\mathcal{L}}_S^{\text{adv}}$ is the average over S of the cross-entropy with the appended GloRo logit (the empirical *adversarial* surrogate, as opposed to the clean empirical risk $\hat{\mathcal{L}}_S$ of Theorem 6); L_φ is the Lipschitz constant, in the logits, of the loss φ (a property of the loss alone, at most $\sqrt{2}$ for softmax cross-entropy); L_{loc} is the local Lipschitz estimate of Section E; the square root is the PAC-Bayes complexity term with confidence $\delta = 0.05$, weighted by $\beta = 1$ so that it enters at the scale at which it appears in (8); and $\sigma = 0.04$, $\sigma_0 = 0.10$. Equation (3) is a training heuristic modelled on, not equal to, (8): because $\hat{\mathcal{L}}_S^{\text{adv}}$ already enforces the margin at ϵ_{train}, the explicit gap term is halved, which the theorem, stated for the clean risk, does not do (Section H). Algorithm 1 (Section A) gives the forward pass with online tracking of the analytical constant.

5 Theoretical Properties

Assumption 1 (Bounded inputs and constrained parameters)**.** *Inputs satisfy* $\|\mathbf{x}_t\|_2 \le X_{\max}$ *for all* t*; the projections satisfy* $\left\|\bar{\mathbf{W}}_\bullet\right\|_2 \le s_\bullet$*;* $\lambda_i(\mathbf{A}) \in [-\lambda_{\max}, -\lambda_{\min}]$ *with* $\lambda_{\max} \ge \lambda_{\min} > 0$*; and* $\Delta_t \in [\Delta_{\min}, \Delta_{\max}]$ *with* $\Delta_{\min} > 0$ *by* (2)*. We write* $\rho_{\max} = e^{-\Delta_{\min}\lambda_{\min}}$*,* $\rho_{\min} = e^{-\Delta_{\max}\lambda_{\max}}$ *and* $c := s_B \Delta_{\max} X_{\max}^2$ *for the per-step injection bound* $\left\|\bar{\mathbf{B}}_t \mathbf{x}_t\right\|_2 \le \Delta_t \left\|\mathbf{B}_t\right\|_2 \|\mathbf{x}_t\|_2 \le c$*.*

Lemma 2 (Bounded state)**.** *Under Assumption 1,* $\|\mathbf{h}_t\|_2 \le H := c/(1-\rho_{\max})$ *(Frobenius norm over all channels and states, hence also for every channel row) for all* t *and all inputs, uniformly over the initial state* $\mathbf{h}_0 = \mathbf{0}$*.*

The lemma (proved in Section B) is the reason the clamp has a lower bound: with $\Delta_{\min} = 0$, $\rho_{\max} = 1$ and no such H exists, which is the error-explosion regime of Qi et al. (2024).

Theorem 3 (Lipschitz constant of the selective scan)**.** *Under Assumption 1, for any two input sequences* $\mathbf{x}, \mathbf{x}'$ *in the bounded domain, writing* $\|\cdot\|$ *for the sequence-level* ℓ_2 *norm,*

$$\|f_{\text{block}}(\mathbf{x}) - f_{\text{block}}(\mathbf{x}')\| \le L_{\text{block}} \|\mathbf{x} - \mathbf{x}'\|, \qquad L_{\text{block}} = s_{\text{out}} L_{\text{SiLU}}\, s_C \Big(\frac{X_{\max}\,\gamma}{1-\rho_{\max}} + H \Big), \tag{4}$$

where $\gamma := 2 s_B \Delta_{\max} X_{\max} + s_\Delta \big(\lambda_{\max} H + s_B X_{\max}^2 \big)$ *collects the per-step input sensitivity of the state through* $\bar{\mathbf{B}}_t$*, through* Δ_t *in* $\bar{\mathbf{A}}_t$*, and through* Δ_t *in* $\bar{\mathbf{B}}_t$*. An* N*-block network on the same domain has constant* $L_{\text{SSM}} \le \prod_{i=1}^N L_{\text{block}}^{(i)}$*.*

The proof (Section B) unrolls the coupled recurrence channel by channel, bounds $\left\|\bar{\mathbf{A}}_t^{(d)} - \bar{\mathbf{A}}_t^{\prime(d)}\right\|_2 \le \lambda_{\max} |\Delta_t^{(d)} - \Delta_t^{\prime(d)}|$ using the diagonal structure, sums over channels with Cauchy–Schwarz (which is why the constant carries no factor of D), and controls the geometric convolution by Young's inequality. The constant is an upper bound; we make no tightness claim. The product over blocks abstracts a cascade of scans; the additional factors of the full block topology are discussed in Section 4.

Remark 1 (Scope: global versus local). Theorem 3 is a global constant on the bounded-input domain and compounds across blocks. For the constants of Section 6 ($\Delta_{\min} = 10^{-3}$, $\lambda_{\min} = 0.05$), $1/(1-\rho_{\max}) = 2 \times 10^4$, $H = 10^4$ and $L_{\text{block}} = 1.1 \times 10^8$, so the worst-case product over 24 blocks is 10^{193}; the nominal-trajectory tracker of Algorithm 1 still reaches $\sim 10^{10}$ (Fig. 5, Section F). The GloRo radius under either constant is below 10^{-10} and vacuous, consistent with Newhouse et al. (2025). The per-input radii of Section 6 therefore use a *local* estimate $L_{\text{loc}}(\mathbf{x})$ from gradient-norm

maximisation over a ball around $\mathbf{x}$ (Section E). It is a *lower* estimate of the true local constant, so the radii are *empirical local-Lipschitz radii*, not certificates, and L_{loc} cannot be substituted into Theorem 3 or Theorem 6; a provable depth-independent local bound is left open. The value of Theorem 3 is structural: it identifies which constraints control sensitivity and which one ($\Delta_{\min}$) prevents unbounded state growth.

Theorem 4 (State-retention bound)**.** *Under Assumption 1, for any trigger of length ℓ injected at t_0,*

$$\|\mathbf{h}_{t_0+\ell}\|_2 \;\geq\; \rho_{\min}^{\ell}\,\|\mathbf{h}_{t_0}\|_2 - c\,\frac{1-\rho_{\min}^{\ell}}{1-\rho_{\min}}. \tag{5}$$

Consequently, with $\kappa := c/\big((1-\rho_{\min})\,\|\mathbf{h}_{t_0}\|_2\big)$, the retention ratio satisfies $\|\mathbf{h}_{t_0+\ell}\|_2 \,/\, \|\mathbf{h}_{t_0}\|_2 \geq \alpha_{\min}$ for every trigger of length

$$\ell \leq \ell^{\star} := \frac{\log\big((\alpha_{\min}+\kappa)/(1+\kappa)\big)}{\log \rho_{\min}}, \tag{6}$$

and $\ell^{\star} > 0$ whenever $\alpha_{\min} < 1$. At least one token is certified ($\ell^{\star} \geq 1$) if and only if $\rho_{\min}(1+\kappa) - \kappa \geq \alpha_{\min}$, i.e. $\|\mathbf{h}_{t_0}\|_2 \geq c/(\rho_{\min}-\alpha_{\min})$ when $\rho_{\min} > \alpha_{\min}$, and never when $\rho_{\min} \leq \alpha_{\min}$.

The proof (Section C) uses $\big\|\bar{\mathbf{A}}_t\mathbf{h}\big\|_2 \geq \sigma_{\min}(\bar{\mathbf{A}}_t)\,\|\mathbf{h}\|_2 \geq \rho_{\min}\,\|\mathbf{h}\|_2$, valid because $\bar{\mathbf{A}}_t$ is diagonal, and the reverse triangle inequality. The denominator $1+\kappa$ in (6) is necessary because the injection term is present on both sides when the bound is solved for ℓ.

Corollary 5 (Directional retention)**.** *Under Assumption 1, for any trigger of length ℓ injected at t_0, with $S_\ell := c\,(1-\rho_{\max}^{\ell})/(1-\rho_{\max}) \leq c\,\ell$ the total injected norm,*

$$\big\langle \mathbf{h}_{t_0+\ell}, \mathbf{h}_{t_0}\big\rangle \;\geq\; \|\mathbf{h}_{t_0}\|_2\,\big(\rho_{\min}^{\ell}\,\|\mathbf{h}_{t_0}\|_2 - S_\ell\big), \qquad \cos\angle(\mathbf{h}_{t_0+\ell}, \mathbf{h}_{t_0}) \;\geq\; \frac{\rho_{\min}^{\ell} - S_\ell/\,\|\mathbf{h}_{t_0}\|_2}{1 + S_\ell/\,\|\mathbf{h}_{t_0}\|_2}, \tag{7}$$

where $\langle\cdot,\cdot\rangle$ is the Frobenius inner product. In particular the component of the post-trigger state along the pre-trigger state is at least $\alpha\,\|\mathbf{h}_{t_0}\|_2$ whenever $\rho_{\min}^{\ell}\,\|\mathbf{h}_{t_0}\|_2 - S_\ell \geq \alpha\,\|\mathbf{h}_{t_0}\|_2$; for $\ell = 1$ this is exactly the one-token condition of Theorem 4.

The proof (Section C) uses that the product of the diagonal transition matrices over the trigger has entries in $[\rho_{\min}^{\ell}, 1)$, so its quadratic form is at least $\rho_{\min}^{\ell}$ times the squared norm, while the injected terms contribute at most S_ℓ. The corollary says more than Theorem 4: a trigger cannot *remove* the pre-trigger state, only *dilute* it with injected content of norm at most S_ℓ. For the constants of Section 6 and $\|\mathbf{h}_{t_0}\|_2 = 4$ the certified cosine after one token is 0.43.

Remark 2 (What the retention bounds certify). Theorem 4 bounds the *norm* of the state and Corollary 5 the component along the pre-trigger state. Neither is a behavioural guarantee: an adversary can add injected content orthogonal to $\mathbf{h}_{t_0}$ of norm up to $c\,\ell$, and whether the readout $\mathbf{C}_t^{\top}\mathbf{h}_t$ then recovers the original context depends on $\mathbf{C}_t$, which the bounds do not control. What they do establish is that the collapse mechanism of HiSPA ($\bar{\mathbf{A}}_t \to 0$) is unavailable: in unconstrained Mamba one saturated step erases the state, whereas under the clamp each token shrinks it by at most $\rho_{\min}$ plus an injection of norm at most c. $\ell^{\star}$ depends on $\|\mathbf{h}_{t_0}\|_2$ through κ and is per-input. For the constants of Section 6 ($s_B = 1$, $\Delta_{\max} = 0.5$, $\lambda_{\max} = 1$, $X_{\max} = 1$, hence $\rho_{\min} = 0.607$ and $c = 0.5$), (6) gives, at $\alpha_{\min} = 0.5$, $\ell^{\star} = 0.95$ for $\|\mathbf{h}_{t_0}\|_2 = 4$ and $\ell^{\star} \to \log\alpha_{\min}/\log\rho_{\min} = 1.39$ as $\|\mathbf{h}_{t_0}\|_2 \to \infty$: a single-token trigger is certified to retain half the norm whenever $\|\mathbf{h}_{t_0}\|_2 > 4.7$, and nothing longer. Against the near-complete erasure of Definition 1, $\alpha = 0.05$, the same formula gives $\ell^{\star} = 2.6$ at $\|\mathbf{h}_{t_0}\|_2 = 4$ and $\ell^{\star} \to 6.0$: no trigger of up to six tokens can erase 95% of the norm. Certifying $\ell^{\star} \geq 24$ at $\alpha_{\min} = 0.5$ would require $\Delta_{\max}\lambda_{\max} \leq 0.029$. The bound certifies the *mechanism* and the first few tokens of Fig. 3, not its empirical range.

Theorem 6 (PAC-Bayes bound on adversarial risk)**.** *Let the loss $\varphi(\cdot, y) \in [0,1]$ be L_φ-Lipschitz in the logits, $Q = \mathcal{N}(\boldsymbol{\theta}, \sigma^2 I_d)$, P a prior independent of the sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \sim \mathcal{D}^n$, $\hat{\mathcal{L}}_S(\boldsymbol{\theta}) = \frac{1}{n}\sum_i \varphi(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)$ the clean empirical risk, and $\mathcal{L}_{\mathrm{adv}}(\boldsymbol{\theta};\epsilon) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \sup_{\|\boldsymbol{\xi}\|_2 \leq \epsilon} \varphi(f_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\xi}), y)$ the adversarial risk. If $f_{\boldsymbol{\theta}}$ is $L(\boldsymbol{\theta})$-Lipschitz on the perturbation domain, then with probability at least $1-\delta$ over S,*

$$\mathbb{E}_{\boldsymbol{\theta}\sim Q}\big[\mathcal{L}_{\mathrm{adv}}(\boldsymbol{\theta};\epsilon)\big] \leq \mathbb{E}_{\boldsymbol{\theta}\sim Q}\big[\hat{\mathcal{L}}_S(\boldsymbol{\theta})\big] + \mathbb{E}_{\boldsymbol{\theta}\sim Q}\big[L_\varphi L(\boldsymbol{\theta})\big]\epsilon + \sqrt{\frac{\mathrm{KL}(Q\|P) + \ln(2\sqrt{n}/\delta)}{2n}}. \tag{8}$$

The proof (Section D) applies McAllester's bound (McAllester, 2003) to the clean loss and the Lipschitz property to the adversarial gap, which is why the gap term enters with coefficient one and the clean $\hat{\mathcal{L}}_S$, not $\hat{\mathcal{L}}_S^{\mathrm{adv}}$, appears on the right; the data-dependent prior of Pérez-Ortiz et al. (2021) keeps the KL term small. The theorem requires a valid upper bound $L(\boldsymbol{\theta})$; the only one available is the global constant of Theorem 3, under which the gap term is vacuous, and L_{loc} is not an upper bound (Remark 1), so we report no numerical value of (8). What it establishes is the form of the dependence: the adversarial excess risk is controlled by the constant that Theorem 3 expresses in the constraint budgets, and its use here is as the training objective (3).

6 Experiments

Benchmarks. *RoBench-25* is the long-context retrieval benchmark released with HiSPA (Mercier et al., 2026b): 120 abstracts of accepted NeurIPS-2025 papers with 240 true/false questions, evaluated with and without trigger insertion between abstract and question; we use the authors' anonymised release, for evaluation only. *HarmBench* (Mazeika et al., 2024) (510 behaviours) provides the safety evaluation.

HarmBench protocol. The models are base language models; the attack success rate (ASR), the fraction of behaviours whose completion HarmBench-CLS judges harmful, is computed on the *scoreable* subset of greedy completions and is a same-model comparison whose mechanism is that the trigger can no longer erase the safety instruction from the state (protocol and caveats in Section I).

Baselines and attacks. Unconstrained Mamba (Gu & Dao, 2024) and Mamba-2 (Dao & Gu, 2024) and the hybrid Jamba-Mini-1.7 (Lieber et al., 2024) (12B active / 52B total) are evaluated as released, without adversarial fine-tuning; the detector defences CLASP (Mercier et al., 2026a) and SpectralGuard (Bonetto, 2026) are applied to the unconstrained Mamba. Attacks are the query-only Z-HiSPA and the white-box M-HiSPA triggers of Mercier et al. (2026b) at $\ell \in \{4, \dots, 48\}$, and PGD-40 in embedding space at $\epsilon \in [0.05, 0.30]$.

Metrics. Clean accuracy (ACC); poisoned accuracy under HiSPA at $\ell = 24$ (PACC); ASR on HarmBench; local-Lipschitz robust accuracy at radius ϵ (LL-Acc@ϵ), the fraction of inputs correctly classified with $\epsilon^\star(\mathbf{x}) \geq \epsilon$ under L_{loc} (Remark 1); ECE with 15 bins (Guo et al., 2017); and per-token latency. Three seeds $\{42, 137, 2026\}$; Friedman test with Holm post-hoc (Demšar, 2006), blocks being the 240 RoBench-25 questions under each seed.

Implementation. PyTorch 2.1 with a reference selective scan implementing (2). LipMamba-130M and -370M are initialised from the public `state-spaces/mamba-130m` and `state-spaces/mamba-370m` checkpoints (pre-trained on The Pile, GPT-NeoX tokenizer) by mapping every tensor into the constrained parameterisation and projecting it onto the constraint set, then adapted with (3) on the WikiText-103 training split in 1024-token blocks, with next-token cross-entropy and a margin term on true/false retrieval questions built from WikiText-103 paragraphs, so that RoBench-25 is never seen in training. Constraints: $s_B = s_C = s_{\mathrm{out}} = 1$, $s_\Delta = 0.5$, $\Delta_{\min} = 10^{-3}$, $\Delta_{\max} = 0.5$, $\lambda_{\min} = 0.05$, $\lambda_{\max} = 1$, $X_{\max} = 1$. Optimiser, derived constants and the perplexity protocol are in Section J.

6.1 Main results

Table 1 reports the headline comparison. Under HiSPA at $\ell = 24$, LipMamba-370M retains 85.3% accuracy against 23.4% for unconstrained Mamba and 71.4% for the strongest detector; the Friedman test over the seven methods rejects equality ($p < 10^{-8}$) and Holm-corrected pairwise comparisons confirm the gap to every baseline ($p < 0.001$ against SpectralGuard). The detectors act after the fact and can only flag; LipMamba changes the dynamics so that the collapse mechanism is unavailable. The local-Lipschitz robust accuracy at $\epsilon = 0.18$ is 80.4%, against 19.2% for Mamba, and ECE improves from 0.058 to 0.034. Because the baselines are base checkpoints while LipMamba is adapted with (3), part of the gap is attributable to adversarial fine-tuning rather than to the constraints; the 130M ablation of Table 2 separates the two. Full LipMamba-130M reaches 82.7%

Table 1: Results on RoBench-25 and HarmBench, mean over three seeds (per-seed values in the released logs). PACC at HiSPA $\ell = 24$ under the published Z-/M-HiSPA triggers. ASR on the scoreable subset of HarmBench completions. LL-Acc@0.18 is robust accuracy under the *local* Lipschitz estimate (Remark 1); it is not a certificate. Detector baselines have no margin head and hence no radius. **Best**, second best.

Method	ACC	PACC	ASR↓	LL-Acc@0.18	ECE↓	Lat./tok
Mamba (unconstrained) (2024)	89.7	23.4	92.1	19.2	0.058	0.5 ms
Mamba-2 (2024)	90.3	25.8	89.4	21.8	0.052	0.4 ms
Jamba-Mini-1.7 (2024)	91.2	18.6	88.7	12.4	0.061	1.2 ms
CLASP (2026a)	89.7	67.1	38.5	—	—	1.0 ms
SpectralGuard (2026)	89.4	71.4	35.6	—	—	0.9 ms
LipMamba-130M	90.8	82.7	32.4	77.1	0.038	0.6 ms
LipMamba-370M	**91.9**	**85.3**	**30.9**	**80.4**	**0.034**	0.8 ms

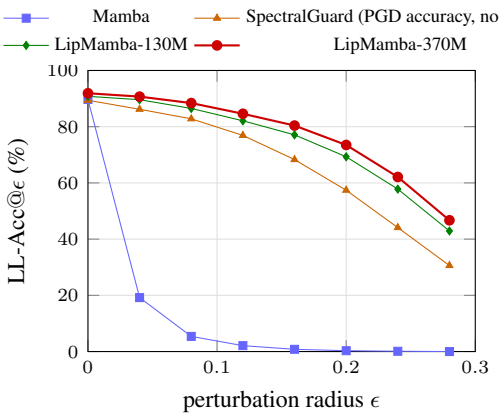

Figure 2: Local-Lipschitz robust accuracy versus radius on RoBench-25 for the models with a margin head; radii are computed under L_{loc} and are empirical (Remark 1). The SpectralGuard curve is empirical accuracy under PGD of the plotted magnitude (no margin head, no radius), for reference only. LipMamba-370M keeps $\geq 70\%$ up to $\epsilon = 0.20$; unconstrained Mamba collapses by $\epsilon = 0.04$.

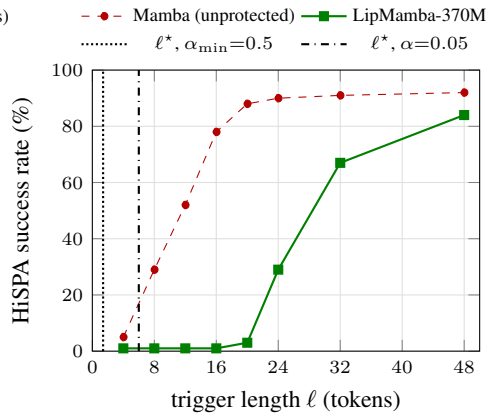

Figure 3: HiSPA success rate versus trigger length (M-HiSPA, RoBench-25). Vertical lines: the supremum over inputs of the length certified by Theorem 4 for the trained constants at 50% retention (1.39) and against 95% erasure (6.0; Remark 2). The plateau of LipMamba beyond these lengths is measured, not certified.

PACC against 23.4% for unconstrained Mamba, a gap of 59.3 points; the same constrained architecture trained *without* the adversarial term reaches 61.9%, so 38.5 points (61.9 − 23.4) come from the architecture alone and 20.8 (82.7 − 61.9) from the objective.

6.2 Global constant, local radii and trigger length

Both analytical curves compound geometrically with depth (Fig. 5, Section F), so the useful quantity is the local estimate of Fig. 2. In Fig. 3 the attack succeeds on unconstrained Mamba from $\ell = 8$ but stays below 5% for LipMamba until $\ell \approx 20$; Theorem 4 accounts for the first token at 50% retention and the first six against erasure, the rest is empirical. The constraints cost 3.5–4.9% perplexity (Fig. 4).

6.3 Ablation

Spectral normalisation of the selective projections is the single most important component (Table 2): without it PACC drops by 46.5 points and no radius survives, consistent with Theorem 3, where s_B, s_C, s_Δ multiply every term. The step clamp is second (−41 PACC points), and the $\Delta_{\max}$ sweep trades clean accuracy for retention ($\Delta_{\max} = 0.25$ gives $\rho_{\min} = 0.78$ and certified $\ell^\star = 2.0$ at

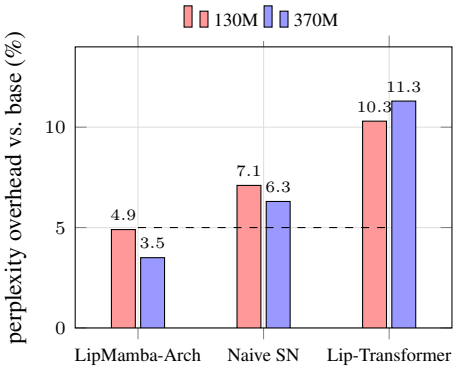

Figure 4: Clean-perplexity overhead on WikiText-103 (validation, GPT-NeoX tokenizer) relative to the public `state-spaces/mamba-130m` and `-370m` checkpoints, including the input projection. Naive spectral normalisation of every matrix costs 6–7%; constraining only the selective projections and the step costs 3.5–4.9% (dashed line: the 5% budget). Lip-Transformer is a Transformer of matched parameter count with spectrally normalised attention and MLP matrices, in the manner of Newhouse et al. (2025), adapted with the same recipe from a base checkpoint of the same size.

Table 2: Component ablation on RoBench-25 (LipMamba-130M, three seeds). LL-Acc@0.18 as in Table 1; last column: median $\epsilon^\star(\mathbf{x})$ over correct inputs.

Configuration	ACC	PACC	LL-Acc@0.18	median $\epsilon^\star$
LipMamba-130M (full)	90.8	**82.7**	**77.1**	0.15
Architectural				
w/o spectral norm on $\mathbf{W}_B, \mathbf{W}_C$	91.4	36.2	0.0	—
w/o step clamp (vanilla softplus)	91.0	41.7	8.4	0.04
w/o eigenvalue reparameterisation (free $\mathbf{A}$)	90.7	64.3	32.5	0.07
w/o GloRo head	90.9	73.8	0.0	—
Objective				
w/o PAC-Bayes term	91.2	58.4	41.6	0.09
w/o adversarial training	90.6	61.9	56.2	0.12
data-independent prior $\mathcal{N}(0, I)$	90.8	78.4	67.3	0.13
Sensitivity to $\Delta_{\max}$				
$\Delta_{\max} = 0.25$	89.7	84.1	80.6	0.16
$\Delta_{\max} = 0.50$ (default)	90.8	82.7	77.1	0.15
$\Delta_{\max} = 1.00$	91.4	71.2	58.7	0.10
$\Delta_{\max} = 2.00$	91.6	52.4	28.6	0.05

50%, 12 against erasure). The eigenvalue reparameterisation and the data-dependent prior contribute smaller but consistent gains.

7 Discussion and Limitations

Proved versus measured. Theorem 3 is structural but vacuous at depth, so the radii are empirical and Theorem 6 is not instantiated (Remark 1); Theorem 4 and Corollary 5 certify state norm and direction for a few tokens, not the readout (Remark 2). Extending them to the readout requires controlling $\mathbf{C}_t^\top$ on the injected subspace; certifying the observed range requires $\Delta_{\max}\lambda_{\max} \leq 0.03$, whose perplexity cost is untested. **Attacks and baselines.** The attacks are the published HiSPA triggers and PGD, so the numbers are robustness against those attacks, not worst-case (Athalye et al., 2018; Carlini et al., 2019); HiSPA, CLASP, SpectralGuard and RoBench-25 are unrefereed 2026 preprints. The detectors certify nothing and the unconstrained models are not adversarially fine-tuned, so Table 1 compares dynamics and defence type, with the adversarial-training share isolated by the ablation. Results are at 130–370M parameters.

8 Conclusion

LipMamba constrains the selective recurrence so that its sensitivity and state norm admit explicit bounds, and states what they do and do not certify. The lower step clamp bounds the state, the upper clamp removes the HiSPA collapse mechanism, and spectral normalisation of the selective projections matters most empirically. Making the local radius provable and extending the retention bounds beyond a few tokens and from norm to content are the main open problems.

Reproducibility Statement

All constraint values, the derived constants and optimisation hyper-parameters are listed in Section 6; the forward pass with online constant tracking is Algorithm 1; the local Lipschitz estimator is specified in Section E. Complete proofs of Lemma 2, Theorem 3, Theorem 4 and Theorem 6 are in Sections B to D, with all assumptions collected in Assumption 1. RoBench-25, HarmBench, JailbreakBench and WildJailbreak are public; the base checkpoints are the public `state-spaces/mamba-130m` and `-370m`. An anonymised repository (`https://anonymous.4open.science/r/LipMamba-ICLR27`) contains the constrained scan, the certificate and PAC-Bayes code, the HiSPA attack harness, baseline configurations, a script that prints every derived constant of Assumption 1–Theorem 4 for a configuration, and the scripts that regenerate every table and figure from the seeds listed above.

Ethics Statement

This work is defensive. The attacks used are publicly documented; no new attack is introduced. Benchmarks contain no personal data. Because the guarantees are scoped as described in Section 7 (norm, not content; a few tokens, not the empirical range; local, not global), the method should not be deployed as a sole safeguard for a safety-critical system.

AI Use Statement

Large-language-model assistants were used to reformat the manuscript to the ICLR template, partially for copy-editing of author-written text, to draft the boilerplate of the three statements, and, under author review, to write parts of the released evaluation harness (the scripts that recompute the constants of Assumption 1–Theorem 4 and the baseline presets). They were not used to generate research ideas, the experimental design, the proofs, or any reported result. All AI-assisted text were reviewed by the authors and every citation was verified manually against its source. The authors take full responsibility for the content of this work.

References

Roger Nick Anaedevha, Alexander G. Trofimov, and Yuri V. Borodachev. MambaShield: Selective state-space models for poisoning-resilient sequential modeling. *Expert Systems with Applications*, 2026. doi: 10.1016/j.eswa.2026.131175.

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *Proc. ICML*, 2016.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proc. ICML*, 2018.

Nicholas H. Barbara, Ruigang Wang, and Ian R. Manchester. R2DN: Scalable parameterization of contracting and Lipschitz recurrent deep networks. *arXiv preprint arXiv:2504.01250*, 2025.

Davi Bonetto. SpectralGuard: Detecting memory collapse attacks in state space models. *arXiv preprint arXiv:2603.12414*, 2026.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, et al. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Proc. NeurIPS Datasets and Benchmarks*, 2024.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In *Proc. NeurIPS*, 2024.

Jeremy Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proc. ICML*, 2019.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proc. ICML*, 2024.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Chengbin Du, Yanxi Li, and Chang Xu. Understanding robustness of visual state space models for image classification. *arXiv preprint arXiv:2403.10935*, 2024.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proc. UAI*, 2017.

N. Benjamin Erichson, Omri Azencot, Alejandro Queiruga, Liam Hodgkinson, and Michael W. Mahoney. Lipschitz recurrent neural networks. In *Proc. ICLR*, 2021.

Deividas Eringis, John Leth, Zheng-Hua Tan, Rafal Wisniewski, and Mihaly Petreczky. PAC-Bayes generalisation bounds for dynamical systems including stable RNNs. In *Proc. AAAI*, 2024.

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

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

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Proc. NeurIPS*, 2020.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *Proc. ICLR*, 2022.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying vulnerabilities in the machine learning model supply chain. In *NeurIPS Workshop on Machine Learning and Computer Security*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proc. ICML*, 2017.

Kai Hu, Andy Zou, Zifan Wang, Klas Leino, and Matt Fredrikson. Unlocking deterministic robustness certification on ImageNet. In *Proc. NeurIPS*, 2023.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, et al. WildTeaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. In *Proc. NeurIPS*, 2024.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *Proc. ICML*, 2021.

Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pre-trained models. In *Proc. ACL*, 2020.

Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-robust neural networks. In *Proc. ICML*, 2021.

Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. In *Proc. ICLR*, 2021.

Yige Li, Hanxun Huang, Yunhan Zhao, Xingjun Ma, and Jun Sun. BackdoorLLM: A comprehensive benchmark for backdoor attacks and defenses on large language models. In *Proc. NeurIPS*, 2025.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. ICLR*, 2019.

Sanae Lotfi, Marc Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In *Proc. ICML*, 2024.

Hashmat Shadab Malik, Fahad Shamshad, Muzammal Naseer, Karthik Nandakumar, Fahad Shahbaz Khan, and Salman Khan. Towards evaluating the robustness of visual state space models. In *Proc. CVPR Workshops*, 2025.

Leonardo Massai, Muhammad Zakwan, and Giancarlo Ferrari-Trecate. L2RU: A structured state space model with prescribed L_2-bound. In *Proc. IEEE Conference on Decision and Control (CDC)*, 2025. arXiv:2503.23818.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proc. ICML*, 2024.

David A. McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

Alexandre Le Mercier, Thomas Demeester, and Chris Develder. CLASP: Defending hybrid large language models against hidden state poisoning attacks. *arXiv preprint arXiv:2603.12206*, 2026a.

Alexandre Le Mercier, Chris Develder, and Thomas Demeester. Hidden state poisoning attacks against Mamba-based language models. *arXiv preprint arXiv:2601.01972*, 2026b.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proc. ICLR*, 2018.

Waleed Mustafa, Philipp Liznerski, Dennis Wagner, Puyu Wang, and Marius Kloft. Non-vacuous generalization bounds for adversarial risk in stochastic neural networks. In *Proc. AISTATS*, 2024.

Laker Newhouse, R. Preston Hess, Franz Cesista, Andrii Zahorodnii, Jeremy Bernstein, and Phillip Isola. Training transformers with enforced Lipschitz constants. *arXiv preprint arXiv:2507.13338*, 2025.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. In *Proc. ICLR*, 2018.

Patricia Pauli, Ruigang Wang, Ian R. Manchester, and Frank Allgöwer. LipKernel: Lipschitz-bounded convolutional neural networks via dissipative layers. *arXiv preprint arXiv:2410.22258*, 2024.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research*, 22(227):1–40, 2021.

Biqing Qi, Yiang Luo, Junqi Gao, Pengfei Li, Kai Tian, Zhiyuan Ma, and Bowen Zhou. Exploring adversarial robustness of deep state space models. In *Proc. NeurIPS*, 2024.

Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, and Lei Zhang. LipsFormer: Introducing Lipschitz continuity to vision transformers. In *Proc. ICLR*, 2023.

Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and J. Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *Proc. ICML*, 2020.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *Proc. ICLR*, 2023.

Asher Trockman and J. Zico Kolter. Orthogonalizing convolutional layers with the Cayley transform. In *Proc. ICLR*, 2021.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Proc. NeurIPS*, 2018.

Paul Viallard, Eric Guillaume Vidot, Amaury Habrard, and Emilie Morvant. A PAC-Bayes analysis of adversarial robustness. *arXiv preprint arXiv:2102.08649*, 2021.

Haoran Wang and Kai Shu. Trojan activation attack: Red-teaming large language models using activation steering for safety-alignment. In *Proc. ACM CIKM*, 2024.

Ruigang Wang and Ian R. Manchester. Direct parameterization of Lipschitz-bounded deep networks. In *Proc. ICML*, 2023.

Wenxiao Wang, Alexander Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In *Proc. ICML*, 2022.

Cheng-Yi Wu and Yu Zhang. BadViM: Backdoor attack against vision Mamba. *arXiv preprint arXiv:2507.00577*, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023a.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023b.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *Proc. USENIX Security Symposium*, 2025.

A FORWARD PASS WITH CONSTANT TRACKING

Algorithm 1 LipMamba forward pass with online tracking of the analytical constant

Require: scan inputs $(\mathbf{x}_1, \dots, \mathbf{x}_L)$, $\mathbf{x}_t \in \mathbb{R}^D$; parameters $\{\mathbf{W}_B, \mathbf{W}_C, \mathbf{W}_\Delta, \mathbf{b}_\Delta, \boldsymbol{\alpha}, \mathbf{W}_{\text{out}}\}$, bounds $(s_B, s_C, s_\Delta, s_{\text{out}})$, $(\lambda_{\min}, \lambda_{\max})$, $(\Delta_{\min}, \Delta_{\max})$, $X_{\max}$; all operations per channel d with $\mathbf{h}_t \in \mathbb{R}^{D\times N}$

1: $\bar{\mathbf{W}}_\bullet \leftarrow \mathbf{W}_\bullet \min(1, s_\bullet/\sigma^{\text{PI}}_{\max}(\mathbf{W}_\bullet))$ for $\bullet \in \{B, C, \Delta, \text{out}\}$
2: $\mathbf{A} \leftarrow -\text{diag}(\lambda_{\min} + (\lambda_{\max} - \lambda_{\min})\sigma(\boldsymbol{\alpha}))$; $\rho_{\max} \leftarrow e^{-\Delta_{\min}\lambda_{\min}}$; $H \leftarrow s_B \Delta_{\max} X^2_{\max}/(1-\rho_{\max})$
3: $\mathbf{h}_0 \leftarrow \mathbf{0}$; $D_0 \leftarrow 0$ {D_t tracks the state-sensitivity bound}
4: **for** $t = 1$ to L **do**
5: $\quad \mathbf{x}_t \leftarrow \mathbf{x}_t \min(1, X_{\max}/\|\mathbf{x}_t\|_2)$ {bounded-input domain}
6: $\quad \Delta_t \leftarrow \Delta_{\min} + (\Delta_{\max} - \Delta_{\min}) \tanh(\text{softplus}(\bar{\mathbf{W}}_\Delta \mathbf{x}_t + \mathbf{b}_\Delta)/(\Delta_{\max} - \Delta_{\min}))$
7: $\quad \mathbf{B}_t \leftarrow \bar{\mathbf{W}}_B \mathbf{x}_t$; $\mathbf{C}_t \leftarrow \bar{\mathbf{W}}_C \mathbf{x}_t$; $\bar{\mathbf{A}}_t \leftarrow \exp(\Delta_t \mathbf{A})$; $\bar{\mathbf{B}}_t \leftarrow \Delta_t \mathbf{B}_t$
8: $\quad \mathbf{h}_t \leftarrow \bar{\mathbf{A}}_t \mathbf{h}_{t-1} + \bar{\mathbf{B}}_t \mathbf{x}_t$; $\mathbf{y}_t \leftarrow \bar{\mathbf{W}}_{\text{out}} \text{SiLU}(\mathbf{C}_t^\top \mathbf{h}_t)$
9: $\quad \gamma_t \leftarrow 2 s_B \Delta_t X_{\max} + s_\Delta(\lambda_{\max}\|\mathbf{h}_{t-1}\|_2 + s_B X^2_{\max})$; $D_t \leftarrow \|\bar{\mathbf{A}}_t\|_2 D_{t-1} + \gamma_t$ {data-dependent refinement of (4)}
10: **end for**
11: $L_{\text{block}} \leftarrow s_{\text{out}} L_{\text{SiLU}} s_C (X_{\max} \max_t D_t + \max_t \|\mathbf{h}_t\|_2)$; logits $\mathbf{z} \leftarrow \text{Head}(\mathbf{y}_L)$
12: **return** $(\mathbf{y}_{1:L})$, L_{block}, margin $z_{\hat{y}} - \max_{k\neq\hat{y}} z_k$

Replacing the observed $\|\bar{\mathbf{A}}_t\|_2$ and $\|\mathbf{h}_t\|_2$ by their worst-case values $\rho_{\max}$ and H recovers the input-free constant of Theorem 3; the tracked value is therefore always at most the worst case, and is the quantity plotted as the middle curve of Fig. 5. It is a *nominal-trajectory path-sensitivity estimate*, not a certified data-dependent operator norm: lines 9–10 evaluate the cross-terms of the proof of Theorem 3 at the unperturbed Δ_t and $\mathbf{h}_{t-1}$ rather than over all perturbed trajectories in a ball, and D_t is a row sum of the time-varying kernel $M_{t,k} = \gamma_k \prod_{j=k+1}^{t} \|\bar{\mathbf{A}}_j\|_2$, which bounds the $\ell_\infty \to \ell_\infty$ gain, whereas the sequence-level ℓ_2 operator norm would also require the column sums (Schur's test). A certified data-dependent bound would replace both by their suprema over the ball, which for the constants used here returns essentially the worst case.

B Proofs of Lemma 2 and Theorem 3

Diagonal structure. For each channel d, $\mathbf{A}^{(d)} = \mathrm{diag}(a_1, \dots, a_N)$ with $a_i \in [-\lambda_{\max}, -\lambda_{\min}]$ and $\Delta_t^{(d)} \in [\Delta_{\min}, \Delta_{\max}]$, so $\bar{\mathbf{A}}_t^{(d)} = \mathrm{diag}(e^{\Delta_t^{(d)} a_i})$ with $e^{\Delta_t^{(d)} a_i} \in [\rho_{\min}, \rho_{\max}]$. Hence $\left\|\bar{\mathbf{A}}_t^{(d)}\right\|_2 \le \rho_{\max}$ and $\sigma_{\min}(\bar{\mathbf{A}}_t^{(d)}) \ge \rho_{\min}$. Moreover, for two steps Δ, Δ' in the range, $|e^{\Delta a_i} - e^{\Delta' a_i}| \le |a_i|\,|\Delta - \Delta'| \le \lambda_{\max}|\Delta - \Delta'|$ by the mean-value theorem, since $a_i < 0$ gives $|a_i e^{s a_i}| \le |a_i|$ for $s > 0$; so $\left\|\bar{\mathbf{A}}_t^{(d)} - \bar{\mathbf{A}}_t^{\prime(d)}\right\|_2 \le \lambda_{\max}|\Delta_t^{(d)} - \Delta_t^{\prime(d)}|$. We write $\delta_t = \mathbf{x}_t - \mathbf{x}'_t \in \mathbb{R}^D$ with components $\delta_t^{(d)}$, and note $|x_t^{(d)}| \le \|\mathbf{x}_t\|_2 \le X_{\max}$, $\sum_d |x_t^{(d)}|^2 \le X_{\max}^2$, $\|\Delta_t - \Delta'_t\|_2 \le s_\Delta \|\delta_t\|_2$ (the clamp is 1-Lipschitz per channel and $\left\|\bar{\mathbf{W}}_\Delta\right\|_2 \le s_\Delta$).

Proof of Lemma 2. Per channel, $\mathbf{h}_t^{(d)} = \sum_{k \le t} \big(\prod_{j=k+1}^{t} \bar{\mathbf{A}}_j^{(d)}\big) \Delta_k^{(d)} \mathbf{B}_k x_k^{(d)}$, so $\left\|\mathbf{h}_t^{(d)}\right\|_2 \le \frac{c}{X_{\max}} \sum_{k \le t} \rho_{\max}^{t-k} |x_k^{(d)}|$. By Cauchy–Schwarz, $\big(\sum_k \rho_{\max}^{t-k} |x_k^{(d)}|\big)^2 \le \big(\sum_k \rho_{\max}^{t-k}\big)\big(\sum_k \rho_{\max}^{t-k} |x_k^{(d)}|^2\big)$, and summing over d, $\|\mathbf{h}_t\|_2^2 \le \frac{c^2}{X_{\max}^2} \cdot \frac{1}{1-\rho_{\max}} \sum_k \rho_{\max}^{t-k} \|\mathbf{x}_k\|_2^2 \le \frac{c^2}{(1-\rho_{\max})^2} = H^2$. The same bound holds a fortiori for each row $\mathbf{h}_t^{(d)}$. For $D = 1$ this is the scalar recursion $\|\mathbf{h}_t\|_2 \le \rho_{\max} \|\mathbf{h}_{t-1}\|_2 + c$ iterated from $\mathbf{h}_0 = \mathbf{0}$. □

Proof of Theorem 3. Let $\mathbf{x}, \mathbf{x}'$ be two sequences in the domain and $\mathbf{e}_t^{(d)} = \mathbf{h}_t^{(d)} - \mathbf{h}_t^{\prime(d)}$. Subtracting the two recurrences channel-wise, $\mathbf{e}_t^{(d)} = \bar{\mathbf{A}}_t^{(d)} \mathbf{e}_{t-1}^{(d)} + (\bar{\mathbf{A}}_t^{(d)} - \bar{\mathbf{A}}_t^{\prime(d)}) \mathbf{h}_{t-1}^{\prime(d)} + (\Delta_t^{(d)} - \Delta_t^{\prime(d)}) \mathbf{B}_t x_t^{(d)} + \Delta_t^{\prime(d)} (\mathbf{B}_t - \mathbf{B}'_t) x_t^{(d)} + \Delta_t^{\prime(d)} \mathbf{B}'_t \delta_t^{(d)}$. We bound the Frobenius norm of each term over channels. (i) $\big(\sum_d \left\|\bar{\mathbf{A}}_t^{(d)} \mathbf{e}_{t-1}^{(d)}\right\|_2^2\big)^{1/2} \le \rho_{\max} \|\mathbf{e}_{t-1}\|_2$. (ii) By the diagonal structure, $\left\|(\bar{\mathbf{A}}_t^{(d)} - \bar{\mathbf{A}}_t^{\prime(d)}) \mathbf{h}_{t-1}^{\prime(d)}\right\|_2 \le \lambda_{\max}|\Delta_t^{(d)} - \Delta_t^{\prime(d)}| \left\|\mathbf{h}_{t-1}^{\prime(d)}\right\|_2$; summing squares over d and using $\left\|\mathbf{h}_{t-1}^{\prime(d)}\right\|_2 \le \left\|\mathbf{h}'_{t-1}\right\|_2 \le H$ (Lemma 2) gives at most $\lambda_{\max} H \|\Delta_t - \Delta'_t\|_2 \le \lambda_{\max} H s_\Delta \|\delta_t\|_2$. (iii) With $\|\mathbf{B}_t\|_2 \le s_B X_{\max}$, the first injection term satisfies $\big(\sum_d |\Delta_t^{(d)} - \Delta_t^{\prime(d)}|^2 \|\mathbf{B}_t\|_2^2 |x_t^{(d)}|^2\big)^{1/2} \le s_B X_{\max} \cdot X_{\max} \|\Delta_t - \Delta'_t\|_2 \le s_\Delta s_B X_{\max}^2 \|\delta_t\|_2$, and the second $\big(\sum_d (\Delta_{\max} \left\|\bar{\mathbf{W}}_B \delta_t\right\|_2 |x_t^{(d)}|)^2\big)^{1/2} \le \Delta_{\max} s_B \|\delta_t\|_2 X_{\max}$. (iv) $\big(\sum_d (\Delta_{\max} \|\mathbf{B}'_t\|_2 |\delta_t^{(d)}|)^2\big)^{1/2} \le \Delta_{\max} s_B X_{\max} \|\delta_t\|_2$. Collecting (ii)–(iv) gives $\|\mathbf{e}_t\|_2 \le \rho_{\max} \|\mathbf{e}_{t-1}\|_2 + \gamma \|\delta_t\|_2$ with $\gamma = 2 s_B \Delta_{\max} X_{\max} + s_\Delta(\lambda_{\max} H + s_B X_{\max}^2)$, independent of D, and unrolling, $\|\mathbf{e}_t\|_2 \le \gamma \sum_{k \le t} \rho_{\max}^{t-k} \|\delta_k\|_2$. The right-hand side is a convolution of the sequence $(\|\delta_k\|_2)_k$ with the time-invariant kernel $(\rho_{\max}^j)_{j \ge 0}$; by Young's inequality its ℓ_2 norm over t is at most the ℓ_1 norm of the kernel times the ℓ_2 norm of the sequence, i.e. $\|\mathbf{e}\| \le \frac{\gamma}{1-\rho_{\max}} \|\mathbf{x} - \mathbf{x}'\|$.

For the readout, $y_t^{(d)} - y_t^{\prime(d)} = \mathbf{C}_t^\top \mathbf{e}_t^{(d)} + (\mathbf{C}_t - \mathbf{C}'_t)^\top \mathbf{h}_t^{\prime(d)}$, so $\|\mathbf{y}_t - \mathbf{y}'_t\|_2 \le s_C X_{\max} \|\mathbf{e}_t\|_2 + s_C \|\delta_t\|_2 \|\mathbf{h}'_t\|_2 \le s_C X_{\max} \|\mathbf{e}_t\|_2 + s_C H \|\delta_t\|_2$, and over the sequence $\|\mathbf{y} - \mathbf{y}'\| \le s_C \big(X_{\max} \frac{\gamma}{1-\rho_{\max}} + H\big) \|\mathbf{x} - \mathbf{x}'\|$. Composing with SiLU ($L_{\mathrm{SiLU}}$-Lipschitz) and $\bar{\mathbf{W}}_{\mathrm{out}}$ ($\|\cdot\|_2 \le s_{\mathrm{out}}$) gives (4). The cascade bound follows from $L(f \circ g) \le L(f) L(g)$, each scan's output staying in a bounded domain by Lemma 2 and the input projection. □

C Proofs of Theorem 4 and Corollary 5

Proof of Theorem 4. By the reverse triangle inequality and the diagonal structure (applied per channel and summed in Frobenius norm), $\|\mathbf{h}_t\|_2 \ge \left\|\bar{\mathbf{A}}_t \mathbf{h}_{t-1}\right\|_2 - \left\|\bar{\mathbf{B}}_t \mathbf{x}_t\right\|_2 \ge \rho_{\min} \|\mathbf{h}_{t-1}\|_2 - c$, where $\left\|\bar{\mathbf{B}}_t \mathbf{x}_t\right\|_2 \le c$ is the per-step injection bound of Assumption 1 (over channels, $\sum_d (\Delta_t^{(d)} \|\mathbf{B}_t\|_2 |x_t^{(d)}|)^2 \le (\Delta_{\max} s_B X_{\max})^2 X_{\max}^2 = c^2$). Let $u_j = \|\mathbf{h}_{t_0+j}\|_2$. Then $u_j \ge \rho_{\min} u_{j-1} - c$, and by induction $u_\ell \ge \rho_{\min}^\ell u_0 - c \sum_{j=0}^{\ell-1} \rho_{\min}^j = \rho_{\min}^\ell u_0 - c(1 - \rho_{\min}^\ell)/(1 - \rho_{\min})$, which is (5). Dividing by u_0 and writing $\kappa = c/((1 - \rho_{\min}) u_0)$, the retention ratio is at least $\rho_{\min}^\ell (1 + \kappa) - \kappa$. This is $\ge \alpha_{\min}$ iff $\rho_{\min}^\ell \ge (\alpha_{\min} + \kappa)/(1 + \kappa)$, i.e. iff $\ell \log \rho_{\min} \ge \log((\alpha_{\min} + \kappa)/(1 + \kappa))$. Since $\log \rho_{\min} < 0$ and $(\alpha_{\min} + \kappa)/(1 + \kappa) < 1$ for $\alpha_{\min} < 1$, both

sides are negative and the condition is $\ell \le \ell^\star$ with $\ell^\star$ as in (6), which is strictly positive. Setting $\ell = 1$ in $\rho_{\min}^{\ell}(1+\kappa) - \kappa \ge \alpha_{\min}$ gives the one-token condition $\rho_{\min}(1+\kappa) - \kappa \ge \alpha_{\min}$, i.e. $\kappa \le (\rho_{\min} - \alpha_{\min})/(1 - \rho_{\min})$, i.e. $u_0 \ge c/(\rho_{\min} - \alpha_{\min})$ when $\rho_{\min} > \alpha_{\min}$. □

Proof of Corollary 5. Unrolling the recurrence over the trigger, $\mathbf{h}_{t_0+\ell} = \mathbf{P}\mathbf{h}_{t_0} + \mathbf{r}$, where per channel $\mathbf{P}^{(d)} = \prod_{j=1}^{\ell} \bar{\mathbf{A}}_{t_0+j}^{(d)}$ is diagonal with entries in $[\rho_{\min}^{\ell}, 1)$, and $\mathbf{r} = \sum_{j=1}^{\ell} \big(\prod_{i=j+1}^{\ell} \bar{\mathbf{A}}_{t_0+i}\big) \bar{\mathbf{B}}_{t_0+j} \mathbf{x}_{t_0+j}$ collects the injected terms; each has norm at most c and is multiplied by $\ell - j$ subsequent contractions of norm at most $\rho_{\max}$, so $\|\mathbf{r}\|_2 \le c \sum_{j=0}^{\ell-1} \rho_{\max}^{j} = S_\ell \le c\ell$. Since $\mathbf{P}$ is diagonal with entries at least $\rho_{\min}^{\ell}$, its quadratic form satisfies $\langle \mathbf{P}\mathbf{h}_{t_0}, \mathbf{h}_{t_0} \rangle = \sum_{d,i} P_i^{(d)} (h_{t_0,i}^{(d)})^2 \ge \rho_{\min}^{\ell} \|\mathbf{h}_{t_0}\|_2^2$, and by Cauchy–Schwarz $\langle \mathbf{r}, \mathbf{h}_{t_0} \rangle \ge -S_\ell \|\mathbf{h}_{t_0}\|_2$; adding the two gives the first inequality of (7). (The recursive argument of Theorem 4, which discounts each injection by later contractions $\ge \rho_{\min}$, does not transfer to the inner product, because $\langle \bar{\mathbf{A}}\mathbf{h}, \mathbf{g} \rangle \ge \rho_{\min} \langle \mathbf{h}, \mathbf{g} \rangle$ requires $h_i g_i \ge 0$ coordinatewise; this is why the additive term here is S_ℓ rather than $c(1 - \rho_{\min}^{\ell})/(1 - \rho_{\min})$.) For the second inequality, $\|\mathbf{h}_{t_0+\ell}\|_2 \le \|\mathbf{P}\mathbf{h}_{t_0}\|_2 + \|\mathbf{r}\|_2 \le \|\mathbf{h}_{t_0}\|_2 + S_\ell$ because the entries of $\mathbf{P}$ are below one, and dividing the first inequality by $\|\mathbf{h}_{t_0}\|_2 \|\mathbf{h}_{t_0+\ell}\|_2$ gives the cosine bound. The final claim is the first inequality divided by $\|\mathbf{h}_{t_0}\|_2$; at $\ell = 1$, $S_1 = c$ and the condition $\rho_{\min} \|\mathbf{h}_{t_0}\|_2 - c \ge \alpha \|\mathbf{h}_{t_0}\|_2$ is the one-token condition of Theorem 4. □

D Proof of Theorem 6

For fixed $\boldsymbol{\theta}$ and any $(\mathbf{x}, y)$, $\sup_{\|\boldsymbol{\xi}\|_2 \le \epsilon} \varphi(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\xi}), y) \le \varphi(f_{\boldsymbol{\theta}}(\mathbf{x}), y) + L_\varphi \sup_{\|\boldsymbol{\xi}\|_2 \le \epsilon} \|f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\xi}) - f_{\boldsymbol{\theta}}(\mathbf{x})\|_2 \le \varphi(f_{\boldsymbol{\theta}}(\mathbf{x}), y) + L_\varphi L(\boldsymbol{\theta}) \epsilon$. Taking expectations over $\mathcal{D}$, $\mathcal{L}_{\mathrm{adv}}(\boldsymbol{\theta}; \epsilon) \le R(\boldsymbol{\theta}) + L_\varphi L(\boldsymbol{\theta}) \epsilon$, where $R(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},y)} \varphi(f_{\boldsymbol{\theta}}(\mathbf{x}), y)$ is the clean risk (a different object from the Lipschitz constant $L(\boldsymbol{\theta})$). McAllester's bound (McAllester, 2003) states that with probability at least $1 - \delta$ over S, simultaneously for all Q, $\mathbb{E}_Q R \le \mathbb{E}_Q \hat{\mathcal{L}}_S + \sqrt{(\mathrm{KL}(Q\|P) + \ln(2\sqrt{n}/\delta))/(2n)}$. Taking $\mathbb{E}_Q$ of the pointwise inequality and adding the two gives (8). The data-dependent prior of Pérez-Ortiz et al. (2021) is fit on a split disjoint from S, so P is independent of S as required. □

E Local Lipschitz estimator

For an input $\mathbf{x}$ and radius r, $L_{\mathrm{loc}}(\mathbf{x})$ is the maximum of $\|\nabla_{\mathbf{x}'} (z_{\hat{y}} - z_k)(\mathbf{x}')\|_2 / \sqrt{2}$ over $k \ne \hat{y}$ and over $\mathbf{x}'$ in the ball of radius r around $\mathbf{x}$, found by 20 steps of projected gradient ascent from 8 random starts. This is a lower estimate of the true local constant; radii computed with it are therefore empirical. Reported radii use $r = 0.3$ in embedding space.

F ANALYTICAL CONSTANT VERSUS DEPTH

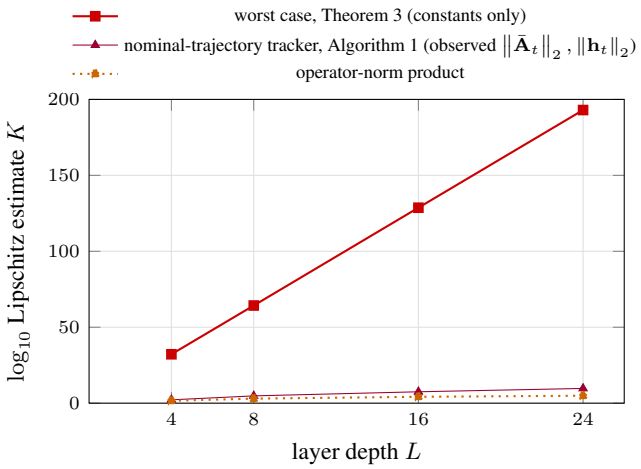

Figure 5: Analytical constant of Theorem 3 composed across depth. The worst-case product from the constants alone grows as $10^{8.04L}$ (10^{193} at $L = 24$); the nominal-trajectory tracker of Algorithm 1 on RoBench-25 inputs reaches $10^{9.7}$ (a path-sensitivity estimate, not a certified operator norm, see Section A); the bare operator-norm product is $10^{4.9}$. The global GloRo radius under either analytical curve is vacuous (Remark 1); the constraints control the growth rate but do not remove it.

G BLOCK TOPOLOGY BEYOND THE SCAN

The Mamba block that we fine-tune retains, around the constrained scan, the input projection to $2D$ channels, the depthwise causal convolution of width 4, the multiplicative gate $\mathbf{y}_t \odot \mathrm{SiLU}(\mathbf{z}_t)$, RMSNorm and the residual connection of the public checkpoints. On the bounded domain each is Lipschitz: a linear map (spectral norm), a fixed-kernel convolution (bounded by the kernel's ℓ_1 norm), SiLU (L_{SiLU}), and the gate, a product of two bounded Lipschitz maps $f \cdot g$ with constant at most $\|f\|_\infty L_g + \|g\|_\infty L_f$, where the sup-norms are finite by Lemma 2 and the input projection; the residual adds one. The full-block constant is the product of these factors and is therefore larger than L_{block} of Theorem 3, and the cascade bound $L_{\mathrm{SSM}} \le \prod_i L^{(i)}_{\mathrm{block}}$ abstracts a cascade of scans. We do not carry the extra factors through because the only conclusion they affect, that the global constant is vacuous at depth (Remark 1), becomes stronger; the component the constraints target, and the one Theorem 4 and Corollary 5 concern, is the scan, whose state is unaffected by the surrounding components within a block.

H NOTES ON THE TRAINING OBJECTIVE

Three remarks keep (3) honest. First, on the $\frac{1}{2}$. Theorem 6 bounds the adversarial risk by the *clean* empirical risk $\hat{\mathcal{L}}_S$ plus the full gap $L_\varphi L\epsilon$. Training instead uses the *adversarial* surrogate $\hat{\mathcal{L}}^{\mathrm{adv}}_S$, whose GloRo logit already penalises any margin below $\sqrt{2}L\epsilon_{\mathrm{train}}$, i.e. it already pays for part of the gap. Adding the full gap on top would count that part twice, so the explicit term is weighted by $\frac{1}{2}$; the value is a design choice (the "w/o PAC-Bayes term" row of Table 2 removes both regularisers), and (3) is a training heuristic, not the right-hand side of Theorem 6. The weight $\beta = 1$ on the complexity term keeps it at the scale of the bound. Second, the objective is evaluated at the posterior mean $\boldsymbol{\theta}$ without Monte-Carlo sampling from Q; the embedding, convolution, gate and normalisation parameters are trained but not randomised. With $\sigma \neq \sigma_0$ the variance part of the KL, $\frac{d}{2}(\sigma^2/\sigma_0^2 - 1 + \ln(\sigma_0^2/\sigma^2)) \approx 0.50\,d$, is a constant that does not depend on $\boldsymbol{\theta}$, so only the mean term $\|\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{prior}}\|_2^2/(2\sigma_0^2)$ shapes training; for $d \approx 8.7 \times 10^7$ this constant alone exceeds one when substituted into (8), which is a further reason no numerical instantiation of Theorem 6 is claimed. Third, as a training signal L_{loc} needs only to track local sensitivity, whereas as a certificate it would need to be an upper bound, which it is not (Remark 1).

I HarmBench protocol

The models are base language models, not instruction-tuned assistants. Each behaviour is presented as a completion prompt preceded by a fixed safety instruction; the HiSPA trigger is inserted between the instruction and the behaviour; 256 tokens are sampled greedily and scored by HarmBench-CLS. Because 130M–370M models frequently produce degenerate completions (repetition, empty or off-language text) that a harm classifier cannot meaningfully score, ASR is computed on the *scoreable* subset (completions of at least eight words without a dominant repeated trigram) and the scoreable fraction is reported with it in the released logs. The number is therefore a same-model comparison: the mechanism by which the constraints lower ASR is that a HiSPA trigger can no longer erase the safety instruction from the state, so the constrained model continues the instruction rather than the behaviour; it is not alignment, and the absolute ASR of a base model is not comparable with instruction-tuned leaderboards. Other jailbreak suites (Chao et al., 2024; Jiang et al., 2024) target instruction-tuned models and are not used.

J Training details

AdamW (Loshchilov & Hutter, 2019) with learning rate 2×10^{-4}, weight decay 0.1, gradient clipping at norm 1.0 and a cosine schedule, 100 epochs over the block set; the prior P is fit on a disjoint 5% clean split; $4\times$A100 80GB. The constraint values of Section 6 give $c = 0.5$, $\rho_{\max} = 0.99995$, $\rho_{\min} = 0.607$, $H = 10^4$ and $L_{\text{block}} = 1.1 \times 10^8$. Perplexity overhead (Fig. 4) is measured on the WikiText-103 validation split (raw, GPT-NeoX tokenizer, block size 1024) against the unmodified base checkpoint evaluated through the same code path, so that the only difference between the two numbers is the constraint set.

K Notation

Symbol	Meaning
$\mathbf{x}_t \in \mathbb{R}^D$, $\mathbf{h}_t \in \mathbb{R}^{D\times N}$, $\mathbf{y}_t \in \mathbb{R}^D$	scan input, state (D channels $\times$ N states), scan output
$\mathbf{A}$, $\boldsymbol{\alpha} \in \mathbb{R}^{D\times N}$	diagonal state matrix and its unconstrained parameter
Δ_t, $\bar{\mathbf{A}}_t$, $\bar{\mathbf{B}}_t$	step size, discretised state matrix $\exp(\Delta_t \mathbf{A})$, input matrix $\Delta_t \mathbf{B}_t$
$\sigma_{\max}(\cdot)$, $\sigma_{\min}(\cdot)$	largest / smallest singular value; $\|\cdot\|_2 = \sigma_{\max}(\cdot)$
$\sigma(\cdot)$	logistic sigmoid (only in the reparameterisation of $\mathbf{A}$)
$s_B, s_C, s_\Delta, s_{\text{out}}$	spectral budgets of the projections
$\lambda_{\min}, \lambda_{\max}$, $\Delta_{\min}, \Delta_{\max}$, $X_{\max}$	eigenvalue, step and input bounds (Assumption 1)
$\rho_{\max}$, $\rho_{\min}$, c, H, γ	derived constants (Assumption 1, Lemma 2, Theorem 3)
ℓ, t_0, $\boldsymbol{\tau}$	trigger length, insertion position, trigger tokens (Definition 1)
α, $\alpha_{\min}$, κ, $\ell^\star$, S_ℓ	retention levels and constants (Theorem 4, Corollary 5)
L_{block}, L_{SSM}, $L_{\text{loc}}(\mathbf{x})$, L	block, network, local and generic Lipschitz constants
$\mathbf{z}$, $\hat{y}$, $\epsilon^\star(\mathbf{x})$, ϵ_{train}	logits, prediction, radius, training radius
φ, L_φ	loss and its Lipschitz constant in the logits
S, n, $\hat{\mathcal{L}}_S$, $\hat{\mathcal{L}}_S^{\text{adv}}$, $\mathcal{L}_{\text{adv}}$	sample, size, clean / adversarial empirical risk, adversarial risk
$\boldsymbol{\theta}$, d, Q, P, σ, σ_0, β, δ	parameters, their count, posterior, prior, their scales, weight, confidence
$\boldsymbol{\xi}$, ϵ	continuous input perturbation and its radius (Theorem 6)