TAYLORSSM: DEPENDENCE-PRESERVING ROBUSTNESS CERTIFICATION FOR SELECTIVE STATE-SPACE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robustness certification is well studied for feedforward networks, Transformers, and recurrent networks. In selective state-space models, the same input perturbations affect both recurrent states and the input-dependent rules for updating and reading them. Bounding these effects separately can weaken certificates, motivating verification methods that exploit their shared dependence. To our knowledge, we present the first end-to-end robustness verifier tailored to Mamba-style classifiers. We propose TaylorSSM as the verifier's core method for constructing sound polynomial bounds that preserve shared input dependence. These bounds support direct certification and enable joint analysis across blocks to certify larger perturbation regions. We certify cases unresolved by IBP and GenBaB's nonlinear branch-and-bound. Our intermediate bounds also strengthen affine bound propagation and tighten GenBaB's bounds before branching. These results highlight the value of combining contributions from shared perturbations before bounding their joint effect.

1 INTRODUCTION

State-space models use recurrent representations of sequence history to balance information retention against computation and memory costs. Selective architectures such as Mamba (Gu & Dao, 2024) and its successors (Dao & Gu, 2024; Lahoti et al., 2026) adapt memory updates to the input. However, predictive success on evaluated inputs does not guarantee prediction stability under input perturbations. Robustness certification provides guarantees over prescribed perturbation sets, with established methods for feedforward networks (Xu et al., 2020), Transformers (Shi et al., 2020; Bonaert et al., 2021), recurrent networks (Ko et al., 2019; Ryou et al., 2021), and general nonlinear computation graphs (Shi et al., 2025). For selective state-space models, the same input perturbations affect both recurrent states and the input-dependent rules for updating and reading them. Bounding these effects separately can combine worst cases that no single perturbation realizes. The challenge is to preserve these dependencies at manageable computational cost.

Formally, given a fixed sequence classifier f and a length-T sequence $X_0 \in \mathbb{R}^{T\times d}$ of d-dimensional embeddings with class label y, we seek to verify

$$\min_{\xi\in[-1,1]^s} [f_y(X_0 + \mathcal{P}\xi) - f_k(X_0 + \mathcal{P}\xi)] > 0, \qquad \forall k \neq y. \tag{1}$$

Here f_k is class k's score, and $\mathcal{P}$ scales perturbations by ε_{in} and inserts them at q selected positions, giving $s = qd$ variables. Although transitions are linear in the state, the objective is generally nonconvex in the perturbations. We seek sound lower bounds; nonpositive bounds are inconclusive.

To our knowledge, we present the first end-to-end robustness verifier tailored to Mamba-style classifiers under box-constrained embedding perturbations. We propose TaylorSSM to enclose network quantities using quadratic polynomials and rigorous bounds on their approximation errors. Each polynomial and its error bound form a quadratic enclosure. The construction builds on Taylor models (Makino & Berz, 2003) and symbolic-remainder propagation in POLAR (Huang et al., 2022).

The resulting margin enclosure yields a direct lower bound on the classification margin. We also develop Full-QC, which uses intermediate ranges to build quadratic constraints on local nonlinear

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.

deviations in the original network (Fazlyab et al., 2022; Xue et al., 2024) and bound their joint effect on the margin. Our main contributions are:

1. **Dependence-preserving quadratic bounds.** TaylorSSM combines shared contributions before bounding them at recurrent readouts and in root mean square (RMS) normalization followed by a fixed linear projection. The readout construction uses accumulated transitions and avoids a polynomial per hidden-state coordinate.
2. **Structured joint certification.** Full-QC exploits recurrent and normalization structure for joint quadratic-constraint certification across blocks, without forming a dense global response matrix.
3. **A sound end-to-end verifier.** We establish soundness for the original classifiers, accounting for approximation errors and numerical errors in the verifier. Experiments demonstrate certification gains and show that intermediate bounds strengthen affine propagation and tighten GenBaB's bounds before branching.

2 Related work

Bound propagation. POPQORN (Ko et al., 2019) bounds recurrent gate–state interactions; PROVER (Ryou et al., 2021) constructs and refines polyhedral abstractions of recurrent updates. GenBaB (Shi et al., 2025) combines linear bound propagation with nonlinear branch-and-bound.

Polynomial enclosures and remainders. Taylor models (Makino & Berz, 2003) and polynomial zonotopes (Kochdumper & Althoff, 2021; Kochdumper et al., 2023) retain nonlinear dependence on shared inputs. POLAR (Huang et al., 2022) symbolically propagates linear remainder images to reduce wrapping (overestimation from repeated enclosures). Hessian-based methods bound outputs (Sharifi & Fazlyab, 2024; Entesari & Fazlyab, 2026), while DiffReach (Shen & Chou, 2026) connects Taylor-model reachability to CROWN through affine bounds. TaylorSSM specializes polynomial and symbolic-remainder representations to unrolled selective recurrences (Dao & Gu, 2024) and RMS normalization followed by a fixed linear projection.

Quadratic constraints. DeepSDP (Fazlyab et al., 2022) verifies networks using quadratic constraints and semidefinite conditions. Sparse and chordal decompositions (Newton & Papachristodoulou, 2021; Xue et al., 2024) and implicit matrix computations (Dathathri et al., 2020) reduce semidefinite verification costs; SDP-CROWN (Chiu et al., 2025) incorporates SDP-derived coupling into linear bound propagation. Full-QC's temporal and RMS recursions preserve a fixed QC system's linear responses and weighted quadratic form, including coupling across blocks.

Robustness of state-space models. Control-theoretic analyses study stability and perturbation sensitivity in Mamba (Halloran et al., 2025; Zubić & Scaramuzza, 2026; Cao et al., 2026). Class-score perturbation bounds imply robustness when sufficiently small relative to nominal margins. Bouri & Saoud (2025) study word-substitution robustness through growth-bound matrices, including for S4 models. Our certificates lower-bound classification margins over continuous embedding boxes for fixed selective SSMs, covering the full computation through normalization and the classification head.

3 TaylorSSM: Dependence-preserving enclosures

TaylorSSM represents intermediate quantities as quadratic polynomials in the same input perturbations, with rigorous bounds on the approximation errors (remainders). When the same remainder contributes through several paths, we combine its fixed linear contributions before bounding their sum. We specialize these rules to recurrent readouts and RMS normalization followed by a fixed linear projection; Algorithm 1 handles the full network.

3.1 Model structure and verification setting

We begin with a Mamba-1 channel j, whose state $h_{t,j} \in \mathbb{R}^n$ and readout $o_{t,j}$ satisfy

$$h_{t,j} = \mathrm{diag}(e^{a_j \Delta_{t,j}}) h_{t-1,j} + B_t \Delta_{t,j} u_{t,j}, \qquad o_{t,j} = C_t^T h_{t,j} + D_j u_{t,j}.$$

Here $u_{t,j}$ is the channel's SiLU activation after a causal convolution. The positive step size $\Delta_{t,j}$ and injection and readout vectors B_t, C_t depend on the input; the negative rate vector a_j and scalar D_j are fixed learned parameters. For each channel, this is the single-injection case of the general recurrence

$$h_t = F_t h_{t-1} + \sum_a B_{t,a} U_{t,a}, \qquad o_t = d_t + C_t^T h_t.$$

The correspondence is $F_t = \mathrm{diag}(e^{a_j \Delta_{t,j}})$, $B_{t,1} = B_t$, $U_{t,1} = \Delta_{t,j} u_{t,j}$, and $d_t = D_j u_{t,j}$. The general form allows vector- or matrix-valued states and multiple injection terms; d_t is computed without the recurrent state. The recurrence coefficients and signals may depend on block inputs up to time t; network weights, nonlinear functions, and initial states remain fixed throughout verification.

Mamba-2 uses matrix-valued states, while the Mamba-3-style recurrence also includes rotations and delayed injections. Supported transitions are either exponentials of an input-dependent scalar times a fixed diagonal matrix, or scalar decays combined with rotations on fixed, disjoint coordinate planes. Within each recurrence they commute, so their products follow from accumulated decay parameters and rotation angles. Injection terms may use a fixed number of past feature values. The causal network combines these recurrences with fixed affine maps, causal convolutions, pointwise nonlinearities, gates, skip connections, and RMS normalization with positive denominators. Appendix C specifies the Mamba-style instances.

For the task in Eq.(1), write $\mathcal{B} = [-1, 1]^s$ and $\mathcal{X} = \{X_0 + \mathcal{P}\xi : \xi \in \mathcal{B}\}$.

Index positions by $t = 0, \ldots, T-1$. If p is the earliest perturbed position, $H = T - p$ is the affected suffix length. The unperturbed prefix fixes the recurrent entry states and past feature values needed at the suffix boundary. We certify the original network under real-arithmetic semantics, with the verifier's numerical errors included in the bounds.

3.2 Quadratic representation and shared dependence

Quadratic enclosures. All times, layers, and branches use the same input perturbation variables $\xi \in \mathcal{B}$, with fixed coordinate ordering and center $\xi = 0$. For a scalar v, an enclosure consists of a quadratic polynomial P_v and a set $\mathcal{R}_v$ enclosing the values of the remainder function r_v:

$$v(\xi) = \underbrace{v_0 + a_v^T \xi + \sum_{i \le j} q_{v,ij} \xi_i \xi_j}_{P_v(\xi)} + r_v(\xi), \qquad r_v(\xi) \in \mathcal{R}_v \quad (\xi \in \mathcal{B}). \tag{2}$$

Denote the constant, linear, and quadratic parts of P_v by v_0, v_1, v_2. The constant v_0 need not equal the exact nominal value $v(0)$; their difference is included in $r_v(0)$. Vector and matrix quantities use this representation componentwise.

Shared remainders. For a shared remainder, $\mathsf{L}_1 r(\xi) + \mathsf{L}_2 r(\xi) = (\mathsf{L}_1 + \mathsf{L}_2) r(\xi)$. We store each remainder's identity, enclosure, and fixed linear maps, combining maps by identity before bounding. Repeated uses therefore share one value rather than independent worst cases. In general, $r_v(\xi) = \sum_o \mathsf{L}_{v,o} r_o(\xi)$ with $r_o(\xi) \in \mathcal{R}_o$ on $\mathcal{B}$. Distinct remainder functions are enclosed separately. Nonlinear operations introduce new remainders whose repeated fixed linear images can be combined (Appendix A.1.1).

Second-order composition. To enclose a smooth map F, collect its enclosed operands as $w = w_0 + w_1 + w_2 + r_w$. Here $DF[\cdot]$ and $D^2F[\cdot, \cdot]$ are Jacobian and componentwise Hessian actions at w_0. Retaining terms through degree two in ξ gives

$$P_F = F(w_0) + DF[w_1] + DF[w_2] + \tfrac{1}{2} D^2F[w_1, w_1]. \tag{3}$$

The output remainder includes the linear inherited term $DF(w_0)[r_w]$. We also enclose interactions involving operand remainders, discarded higher-degree products, the third-order Taylor remainder, and coefficient-arithmetic errors. This rule requires F to be C^3 on a neighborhood of every segment from w_0 to $w(\xi)$, with derivative bounds valid on all these segments. At nonsmooth thresholds, a polynomial approximation with a validated error over the full operand range provides containment instead (Appendix A.1.2).

3.3 READOUT CONSTRUCTION FROM ACCUMULATED TRANSITIONS

We unroll the recurrence to construct readout enclosures without propagating a polynomial for each hidden-state coordinate. Reindex the affected suffix by $t = 0, \dots, H-1$, with enclosed entry state h_{-1}. Writing the intervening transitions as $\Phi_{r:t} = F_t \cdots F_{r+1}$, with $\Phi_{t:t} = I$, the readout is

$$o_t = d_t + C_t^T \Phi_{-1:t} h_{-1} + \sum_{r=0}^{t} \sum_{a} \big(C_t^T \Phi_{r:t} B_{r,a}\big) U_{r,a}. \tag{4}$$

Accumulated transitions. Accumulation can expose cancellation before the exponential remainder is formed. For $\xi \in [-1, 1]$ and steps $\Delta_1 = \delta_0 + \varepsilon\xi$, $\Delta_2 = \delta_0 - \varepsilon\xi$, the product $e^{-\Delta_2} e^{-\Delta_1} = e^{-2\delta_0}$ has no perturbation dependence. Separate expansions followed by degree-two truncation of their product give the same constant polynomial in exact arithmetic. Yet replacing the remainder functions by marginal ranges independent of each other and ξ can, over repeated steps, preclude a positive-margin certificate (Proposition 4).

We first sum the per-step parameter enclosures in the common variables ξ, then use these accumulated parameters in the contracted readout map below. For the supported transition families, fixed commuting generators Λ_j and per-step parameter vectors γ_v give $F_v = \exp(\sum_j \gamma_{v,j} \Lambda_j)$. Their products satisfy $\Phi_{r:t} = \Psi(\theta_{r:t})$, where $\theta_{r:t} = \sum_{v=r+1}^{t} \gamma_v$ and $\Psi(\theta) = \exp(\sum_j \theta_j \Lambda_j)$. The accumulated parameters are cumulative step sizes for diagonal transitions, or accumulated decay parameters and rotation angles for fixed-plane rotations.

Contraction at the readout. Each history term combines an injection, the intervening transitions, and a readout. We enclose this combined map $\mathcal{K}(C, B, \theta) = C^T \Psi(\theta) B$ with operands $w = (C, B, \theta)$. The prefix contribution uses $B = h_{-1}$.

For diagonal transitions with n state coordinates, vector operands C, B, fixed rates a_ν, and accumulated step δ, the contracted map is

$$\mathcal{K}(C, B, \delta) = \sum_{\nu=1}^{n} C_\nu B_\nu e^{a_\nu \delta}. \tag{5}$$

We sum signed contributions over state coordinates in $\mathcal{K}$ and its derivatives before applying Eq.(3). The same contractions combine linear inherited remainder contributions before bounding. We then assemble the readout enclosure via Eq.(4). Appendix A.3.1 gives the diagonal polynomial expansion and complete remainder enclosure.

The quadratic terms retain interactions between the injection/readout factors and the accumulated step, as well as between step variations at different times. With smooth operations and exact Taylor coefficients and arithmetic, stepwise quadratic propagation yields the same readout polynomial. Construction order can still change remainder bounds and costs (Appendix A.2).

Theorem 1 (Polynomial enclosure of the selective scan)**.** *For the transition families in Section 3.1, assume valid polynomial enclosures on $\mathcal{B}$ of the prefix state, C_t, $B_{r,a}$, $U_{r,a}$, d_t, and accumulated parameters, together with derivative bounds on the full operand segments. The readout enclosure rule, valid multiplication by $U_{r,a}$, and summation in Eq.(4) enclose every o_t on $\mathcal{B}$.*

Computational cost. For each injection–readout pair with $C \in \mathbb{R}^{n \times r_C}$ and $B \in \mathbb{R}^{n \times r_B}$, evaluating $\mathcal{K}$ and contracting the required transition-parameter derivatives costs $O(n r_C r_B)$ arithmetic, before assembling polynomial coefficients in ξ. No dense Hessian over operand coordinates is formed. Each represented scalar stores $s(s+1)/2$ quadratic coefficients; all readouts require $O(H^2)$ injection–readout pairs (Appendix A.2).

3.4 PROJECTED RMS WITH SHARED REMAINDERS

RMS normalization provides another opportunity to combine shared contributions: the same inherited input remainder enters both its numerator and denominator. We treat normalization together with the fixed linear projection that follows it.

For normalization over d_{rms} coordinates, absorb the learned weights and projection into a fixed map W and write $F(x) = Wxf(\upsilon)$, where $\upsilon = d_{\text{rms}}^{-1}\|x\|^2 + \eta$, $f(\upsilon) = \upsilon^{-1/2}$, and $\eta > 0$. Given a valid input enclosure $x = x_0 + x_1 + x_2 + r$, put $\upsilon_0 = d_{\text{rms}}^{-1}\|x_0\|^2 + \eta$. $\eta > 0$ ensures smoothness on every expansion segment. The output remainder contains the linear inherited contribution

$$DF(x_0)[r] = \left[f(\upsilon_0)W + \frac{2f'(\upsilon_0)}{d_{\text{rms}}}(Wx_0)x_0^T \right] r. \tag{6}$$

The matrices come from the numerator and denominator, respectively, and act on the same r; we add them before bounding.

Size of the cancellation. For r parallel to x_0, the numerator and denominator contributions partially cancel whenever nonzero, reflecting RMS normalization's suppression of input-scale changes. If η is small relative to $d_{\text{rms}}^{-1}\|x_0\|^2$, combining their maps can give much tighter bounds than separate bounding (Appendix A.4).

3.5 Full-network construction and certification

Algorithm 1 applies the readout and RMS rules together with the rules for the remaining network operations. Under the conditions in Section 3.2 and Appendix A, each rule preserves Eq.(2). Starting from valid input and prefix enclosures, induction over the dependency order establishes containment on $\mathcal{B}$ for every represented quantity.

Algorithm 1 TaylorSSM Enclosure Construction

Input: Network f, X_0, and $\mathcal{P}$ as in Section 3.1. Use enclosure arithmetic on $\mathcal{B}$; $\text{rng}_{\mathcal{B}}$ denotes a valid range enclosure. Each g is a recurrence with all suffix readouts, a projected RMS map, or one other operation.

Initialize $\mathcal{E}$ with $X_0 + \mathcal{P}\xi$ and prefix enclosures.
For each subcomputation g in the affected suffix, in dependency order:
  Read operand enclosures from $\mathcal{E}$.
  If g is a recurrence with readouts: $\theta_{r:t} \leftarrow \sum_{j=r+1}^{t} \gamma_j$ for $0 \le t < H$, $-1 \le r \le t$.
    Sum over state coordinates in $\mathcal{K}$ and its derivatives before applying Eq.(3).
    Enclose the remainders (Section 3.3); assemble $(P_{o_t}, \mathcal{R}_{o_t})$ via Eq.(4).
  Else if g is projected RMS: form P_F via Eq.(3).
    Combine the maps in Eq.(6); bound their shared image and remaining terms to obtain $\mathcal{R}_F$ (Appendix A.4).
  Else if g is a fixed affine map:
    Transform polynomial coefficients; combine $\mathsf{L}_{v,o}$ for each shared r_o before bounding.
  Else: Construct output polynomials and remainder enclosures with the degree-two rules (Appendix A).
  For each output v of g:
    If refinement is enabled and a valid $\mathcal{G}_v \supseteq v(\mathcal{B})$ is available:
      $\mathcal{R}_v \leftarrow \mathcal{R}_v \cap \big(\mathcal{G}_v - \text{rng}_{\mathcal{B}}(P_v)\big)$.
    Store $(P_v, \mathcal{R}_v)$ in $\mathcal{E}[v]$ with the shared r_o and maps $\mathsf{L}_{v,o}$.
Return: $\mathcal{E}$ with $v(\xi) - P_v(\xi) \in \mathcal{R}_v$ for every represented v and all $\xi \in \mathcal{B}$.

Direct certification. Let $\phi(X) = f_y(X) - f_k(X)$ for a fixed competing class $k \neq y$. To obtain a direct margin bound, we combine the classification head's signed polynomial and shared-remainder contributions before bounding. Let $\underline{c}_\phi, L_{2,\phi}, \underline{r}_\phi$ lower-bound the margin's constant, quadratic, and remainder terms, and let A_ϕ upper-bound $\|a_\phi\|_1$ for its linear coefficient vector a_ϕ. Since $a_\phi^T\xi \ge -\|a_\phi\|_1 \ge -A_\phi$ on $\mathcal{B}$, combining the component bounds gives

$$\underline{m}_{\text{dir}} = \underline{c}_\phi - A_\phi + L_{2,\phi} + \underline{r}_\phi \le \inf_{X \in \mathcal{X}} \phi(X).$$

Positive $\underline{m}_{\text{dir}}$ for every competing class on the same box certifies classification robustness. Appendix A.5 details the component bounds.

Exported bounds. Intermediate enclosures supply value intervals by bounding $P_v + r_v$ and affine bounds in ξ by enclosing $v_2 + r_v$ while retaining $v_0 + v_1$. These bounds support Full-QC's joint analysis of the original network (Section 4) and can also be reused by other verifiers.

4 JOINT QUADRATIC CERTIFICATION

The direct bound treats the polynomial and remainder separately, which can combine worst cases from different perturbations. Even exact polynomial minimization can fail to certify with a loose remainder bound. Full-QC uses TaylorSSM's intermediate ranges to relate nonlinear deviations to their inputs through quadratic constraints (QCs), then bounds their joint effect on the original margin.

4.1 FROM POLYNOMIAL BOUNDS TO LOCAL CONSTRAINTS

For a product $w = uv$, as in a gate or recurrent update, write $u = c_u + d_u$ and $v = c_v + d_v$ about fixed centers, using the intermediate bounds to obtain $|d_u| \le \bar{d}_u$ and $|d_v| \le \bar{d}_v$. For positive $\bar{d}_u, \bar{d}_v$, define the residual coordinate $\epsilon_{\text{prod}} = d_u d_v / (\bar{d}_u \bar{d}_v)$. The product output then satisfies

$$w = c_u c_v + c_v d_u + c_u d_v + \bar{d}_u \bar{d}_v \, \epsilon_{\text{prod}},$$
$$\epsilon_{\text{prod}}^2 \le \frac{d_u^2}{\bar{d}_u^2}, \qquad \epsilon_{\text{prod}}^2 \le \frac{d_v^2}{\bar{d}_v^2}.$$

The relaxation replaces the exact residual definition by the two QCs, retaining the affine relation and sharing ϵ_{prod} across all uses of w. The QCs retain input dependence: $d_u = 0$ forces $\epsilon_{\text{prod}} = 0$, unlike an independent residual interval. A zero operand radius eliminates the residual. Here residuals represent deviations from fixed affine relations; TaylorSSM remainders measure deviations from quadratic polynomials (Section 3.2).

Other assigned residuals likewise have affine relations and local QCs. These relations may leave nonlinear terms or center and coefficient mismatches unaccounted for. We enclose these remaining contributions as additive discrepancies and propagate them through the network (Appendix E.1).

Full-QC also uses the affine bounds from Section 3.5 to refine layer-input domains, while keeping each input equal to the preceding layer's output. Every original network execution satisfies these relations and QCs under one consistent assignment of residuals and discrepancies (Appendix B.1).

4.2 CERTIFICATE PRINCIPLE

We obtain a margin lower bound by bounding the squared output response and enclosing discrepancies. Fix the affine coefficients, positive QC weights, and positive residual scales (Appendix G). Write each node deviation as $d = \widehat{d} + b$: $\widehat{d}$ is linear in ξ and the assigned residuals, while b is the propagated discrepancy. Thus the original margin is $\phi = c + \ell_\phi + b_\phi$, with stored constant c, linear output response ℓ_ϕ, and discrepancy b_ϕ.

The local QCs control the residuals through their input signals. For the product above, positive weights λ_u, λ_v and scale ω give

$$p_{\text{prod}} = \left(\sqrt{\lambda_u} \frac{\widehat{d}_u}{\bar{d}_u}, \sqrt{\lambda_v} \frac{\widehat{d}_v}{\bar{d}_v} \right)^T, \qquad e_{\text{prod}} = \omega \epsilon_{\text{prod}}.$$

If both operand discrepancies vanish, the QCs imply $\|p_{\text{prod}}\|^2 \ge (\lambda_u + \lambda_v)\epsilon_{\text{prod}}^2$. Choosing $\omega^2 < \lambda_u + \lambda_v$ therefore makes $\|p_{\text{prod}}\|^2 - e_{\text{prod}}^2 \ge 0$ on original executions. For every residual, its squared scale must be smaller than the sum of its QC weights. Stacking the weighted QC input responses in p and scaled residuals in e gives the quadratic energy $J = \|p\|^2 - \|e\|^2$.

We use this energy to separate a quadratic optimization from the original network constraints. First, find $U \ge 0$ such that $\ell_\phi^2 + J \le U$ throughout the input box for *all real residual assignments*. Allowing free residuals enlarges the feasible set; the term $-\|e\|^2$ can keep the resulting maximization finite. Section 4.3 describes the computation and validation.

Second, use the QCs on original executions to bound $-J \le V$. When discrepancies vanish in all QC input signals, the product argument gives $J \ge 0$, so $V = 0$ suffices. Otherwise, the QCs constrain signals that include b, whereas p contains only their linear responses. Propagated discrepancy bounds give $V \ge 0$ limiting how far J can fall below zero (Appendix E.2).

Adding the two inequalities cancels J. With a verified $b_\phi \geq \underline{b}_\phi$ on the box, this gives

$$\ell_\phi^2 = (\ell_\phi^2 + J) - J \leq U + V, \qquad \phi \geq c + \underline{b}_\phi - \sqrt{U+V} = \underline{m}_{\mathrm{QC}}.$$

Guarantee (informal; Theorem 10). If every original network execution on $\mathcal{X}$ satisfies the local relations and QCs, and $U, V, \underline{b}_\phi$ satisfy the stated bounds with numerical errors enclosed, then $\underline{m}_{\mathrm{QC}}$ lower-bounds the original network's margin throughout $\mathcal{X}$. A positive bound for every competing class certifies Eq.(1).

4.3 Structured computation and numerical validation

To compute U, we must bound $\ell_\phi^2 + J$ for all residual assignments. We organize the calculation into causally ordered blocks. In the running Mamba-1 instance, RMS blocks supply normalized signals to channel-wise temporal blocks; fixed projections and the skip connection form the next layer input.

Within each block, we complete the square to express its energy as a quadratic form in the block inputs minus a squared residual norm. This invertible change of residual coordinates keeps the output linear in those inputs and transformed residuals. Temporal blocks complete squares along the affected suffix; RMS blocks exploit low-rank residual coupling. When composing blocks, we keep reused signals as common variables, preserving cross terms in the joint energy.

After composition, if the quadratic part of $\ell_\phi^2 + J$ in the free residuals is negative definite, we maximize over them by square completion, leaving a quadratic form in ξ. Bounding this form by a diagonal quadratic with nonnegative coefficients gives U as their sum, since $|\xi_i| \leq 1$ on $\mathcal{B}$.

Forward and adjoint passes evaluate operators without forming a dense global response matrix (Appendix B). Floating-point search proposes QC weights and numerical trials (Appendix G). Acceptance requires checks of local and global definiteness, the input-box bound, and final certificate arithmetic, all with numerical errors enclosed (Appendices D and E).

5 Experiments

5.1 Experimental setup

We use Yelp Polarity (binary sentiment classification) as the main benchmark and AG News (four-class topic classification) for multiclass certification. Both classifiers are trained from scratch with four Mamba-1-style layers, residual width $d = 384$, inner width $m = 768$, and state dimension $n = 16$. Yelp test accuracy is 89.38%. The injection and readout matrices B_t, C_t (selectors) are computed before SiLU to keep the selector maps affine in the convolution output; Section 5.5 evaluates the standard post-SiLU variant. Appendix F.1.1 gives training and tokenization details.

Inputs contain $T = 128$ tokens, including left padding and a final classification token. Perturbations vary all embedding coordinates at selected content positions within an ℓ_∞ box of radius $\varepsilon_{\mathrm{in}}$, leaving all other embeddings fixed. At $\varepsilon_{\mathrm{in}} = .1$, token boxes have a maximum ℓ_2 radius of about 10% of the embedding norm and an ℓ_∞ radius of about 3% of the nearest distinct non-padding, non-CLS embedding distance in that norm (median ratios; Appendix F.1.2).

We compare direct margin bounds with Full-QC using shared TaylorSSM enclosures. Baselines include interval bound propagation (IBP), CROWN-IBP (Zhang et al., 2020) (forward intervals with backward affine propagation), and GenBaB's nonlinear branch-and-bound (Shi et al., 2025). SSM-Affine$_0$ and SSM-Affine$_1$ use identical affine propagation without and with our intermediate bounds, respectively, isolating their effect (Appendix H). Our methods and both SSM-Affine variants enclose numerical errors with outward-rounded interval arithmetic; the other baselines use ordinary floating-point arithmetic. The main panel uses an NVIDIA RTX 3090 GPU and AMD EPYC 7402 CPU.

5.2 Certified robustness

We evaluate 256 single-position tasks: four equally spaced content positions in each of 64 correctly classified Yelp reviews. Table 1 uses common tasks and time limits; SSM-Affine$_1$, direct certification, and Full-QC share a radius-search schedule (Appendix F.2). The attained radius is the largest tested radius with a positive margin lower bound.

Table 1: Attained radii on 256 tasks. Threshold counts require positive bounds at or above that radius. Any-radius counts and summaries use each method's search schedule; summaries include only successful tasks. Daggers mark ordinary floating point; other rows use outward rounding.

Method	.01	.05	.1	Any tested radius	Median $[Q_{25}, Q_{75}]$	[Min, Max]
IBP†	0	0	0	256/256	$(3.79\,[2.46, 8.59]) \times 10^{-7}$	$[2.37 \times 10^{-8}, 9.31 \times 10^{-6}]$
CROWN-IBP†	0	0	0	256/256	$(5.14\,[3.40, 10.9]) \times 10^{-7}$	$[9.31 \times 10^{-8}, 1.07 \times 10^{-5}]$
SSM-Affine$_0$	0	0	0	256/256	$(6.98\,[5.83, 10.0]) \times 10^{-4}$	$[3.05 \times 10^{-4}, 3.05 \times 10^{-3}]$
SSM-Affine$_1$	231	216	46	231/256	.07 [.05, .08]	[.025, .1]
Direct bound	231	216	49	231/256	.07 [.05, .08]	[.025, .1]
Full-QC	208	207	126	208/256	.1 [.08, .1]	[.01, .12]

Intermediate bounds bring affine propagation close to the radius coverage of the direct bound. At .1, Full-QC raises coverage from 49/256 with direct certification to 126/256; at .01, direct certification covers more (231/256 versus 208/256), with a union of 235/256. For the 231 tasks in this union with validated attack upper bounds, the larger certified radius per task reaches at least 26% of the true robustness radius on average (Appendix F.2).

On a separate 16-task panel at .1 with the same review-sampling rule, neither GenBaB configuration yields a positive final bound within four hours and 96 GiB of host memory per task. The direct bound and Full-QC certify 2/16 and 11/16 tasks, respectively (Appendix F.4.2).

5.3 Enclosure construction

On four tasks (two each at $H = 2, 12$) and radii $.01, .1$, paired comparisons vary one mechanism with the output-bound method fixed (Table 2). Remainder combination sums linear contributions of the same inherited remainder in RMS before bounding; without quadratic retention, new quadratic terms enter the remainder immediately. At .1, retaining quadratic terms gives the largest median paired increase in the margin lower bound.

History contraction saves about 67–68% of construction time at $H = 2$ and 43–44% at $H = 12$ versus stepwise propagation (median paired speedups at the two radii; Table 6). For one review at $H = 16$, median times per output bound over three radii are about 453 s (direct bound) and 1,802 s (Full-QC; Appendix F.6.1).

Table 2: Enclosure ablation at $\varepsilon_{\text{in}} = .1$. Bounds show disabled $\to$ enabled median margin lower bounds; Δ is the median paired increase. Quadratic retention uses four pairs (one per task) with other mechanisms enabled; other comparisons retain quadratic terms over 16 pairs (Appendix F.3.1). Flips count nonpositive $\to$ positive / reverse pairs; all affine flips share one $H = 12$ task.

	Direct bound			Affine + enclosure intervals		
Mechanism	Median bound	Δ	Flips	Median bound	Δ	Flips
Quadratic retention	$-12.1 \to -0.82$	10.509	0/0	$-7.9 \to -0.286$	7.667	1/0
History contraction	$-1.48 \to -1.23$	0.266	0/0	$-0.544 \to -0.439$	0.107	1/0
Remainder combination	$-1.55 \to -0.993$	0.544	0/0	$-0.61 \to -0.347$	0.299	3/0
Range refinement	$-1.51 \to -1.22$	0.0956	0/0	$-0.485 \to -0.477$	0.0188	1/0

5.4 Bound reuse and joint certification

Table 3 compares methods on common boxes for 32 Yelp tasks ($H \in \{2, 12, 14, 16\}$). Adding enclosure intervals and then compatible affine bounds to SSM-Affine$_0$ yields SSM-Affine$_1$. At .1, Full-QC certifies 19/32 tasks versus 3/32 for the direct bound and 10/32 for SSM-Affine$_1$; direct certification covers more at .025 and .05. Median bound components help explain this change with radius: Full-QC has a relatively large square-root penalty at small radii, while the remainder term in the direct bound becomes much more negative as the radius grows (Table 10; Figure 1).

Table 3: Certified tasks on 32 Yelp boxes; SSM-Affine$_0$ bounds are negative or nonfinite.

Method and supplied bounds	.01	.025	.05	.1
SSM-Affine$_0$	0/32	0/32	0/32	0/32
Affine + enclosure intervals	32/32	31/32	13/32	7/32
SSM-Affine$_1$	32/32	32/32	21/32	10/32
Direct bound	32/32	32/32	24/32	3/32
Full-QC	32/32	26/32	21/32	19/32

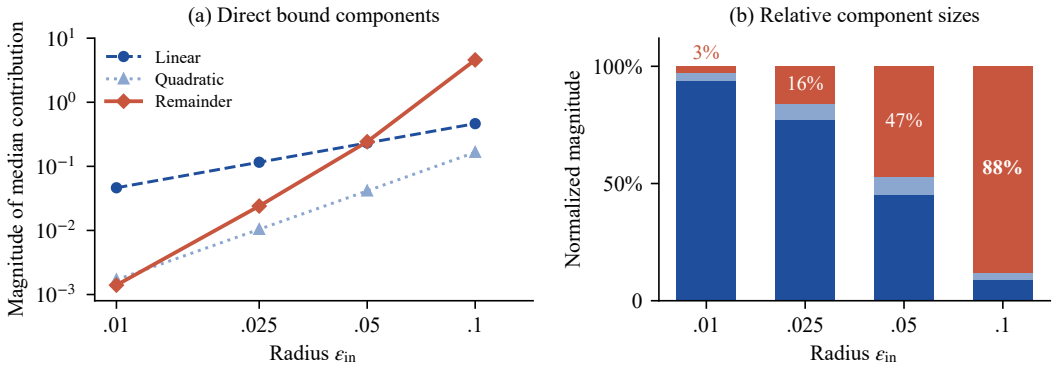

Figure 1: Direct bound components on 32 Yelp tasks. (a) Magnitudes of median contributions. (b) The same magnitudes normalized to sum to one at each radius; labels give remainder shares.

To test whether better polynomial minimization could certify more tasks, we evaluate the fixed direct lower-bound functions exactly at feasible perturbations. Values are nonpositive on 27/32 tasks at .1, so their minima cannot be positive. Exact minimization alone therefore cannot certify these tasks with unchanged remainder and coefficient-error bounds (Appendix F.4.3).

On the separate GenBaB panel at .1, we compare TaylorSSM and analytic ranges in the same FP64 computation before branching. TaylorSSM yields four positive $H = 2$ root bounds, versus none with native or analytic ranges. Relative to analytic ranges, all eight $H = 2$ and three completed longer-horizon bounds improve; the latter remain nonpositive (Appendix F.4.2).

5.5 Additional Evaluations

Multiple margins and positions. Both methods certify all 16 AG News inputs at .01; at .05, direct certification covers 9/16 and Full-QC 10/16. On Yelp at .01, both methods certify fewer four-position than two-position boxes (32 per size; Table 12).

Other selective recurrences. On separately trained post-SiLU Mamba-1, Mamba-2, and Mamba-3-style checkpoints with horizons up to $H = 32$, both methods produce certificates, with Full-QC achieving higher aggregate coverage on each model (Table 13; Appendix F.5.2).

6 Discussion and Conclusion

Readout-level construction avoids per-state polynomials, but each represented scalar still requires $O(s^2)$ storage. The $O(H^2)$ history–readout work also limits scalability. Reducing remainder construction and bounding to $O(H)$ time using updatable history summaries remains open.

Full-QC's symmetric error constraints and squared-response bound lose directional information about margin changes. Retaining it could tighten margin lower bounds.

Our end-to-end verifier certifies Mamba-style classifiers under box-constrained embedding perturbations. Direct certification and Full-QC give complementary coverage across radii; intermediate bounds also strengthen affine propagation and tighten GenBaB's bounds before branching.

Reproducibility Statement

Appendices A–E provide the enclosure rules, model definitions, and certificate derivations and proofs. Appendix F gives training, task selection, evaluation protocols, and additional results; Appendices G and H specify certificate search and the affine reference. The supplementary source code contains the proposed verifier and the experimental and evaluation pipeline.

AI Usage Statement

Generative AI tools, including ChatGPT, Claude, DeepSeek, and Codex, supported research development, software implementation, experimental design and analysis, literature exploration, translation, and manuscript preparation.

Research design and mathematical development. The authors initially defined the research problem, the TaylorSSM construction, and its direct certification procedure; Full-QC emerged through discussions with AI. For Theorem 1, Lemma 2, and Proposition 4, its role was checking algebra and edge cases and organizing remainders and exposition in author-developed arguments. AI contributed intermediate derivations and proof sketches for history/readout contraction and its coefficient and complexity arguments (Theorem 3), recurrence transformations (Corollary 1), formulations for shared residuals, common witnesses, and energy accounting (Proposition 5), and recentered source and original-network margin bounds (Proposition 9 and Theorem 10). It contributed substantially to deriving and drafting the temporal and low-rank square-completion arguments (Propositions 6 and 7), including recursions and response–energy identities, and the structured composition proof (Theorem 8), including preservation of residual directions and energy and matrix-free operator actions. The authors finalized the objects, statements, and assumptions, revising, rejecting, or re-deriving proposed arguments as needed.

Software development. The authors specified mathematical invariants, architecture, and interfaces, and directed debugging and redesign. AI also proposed algorithmic details and participated in iterative implementation. It was used to write, modify, debug, and test code for TaylorSSM, direct certification, Full-QC, certificate search, numerical validation, and reference-method interfaces.

Experiments and reporting. The authors designed the dataset configuration and task framework; ablations and the sizes of evaluation task sets were developed through discussions with AI. AI assisted with software for model training and data handling, perturbation tasks, radius searches, attack generation and validation, ablations, resource measurements, result aggregation, and table and figure generation. It also assisted with summarizing and interpreting experimental findings, checking the resulting claims against the reported results and evaluation scope, exploring related literature, and translating and editing research and manuscript text. AI also assisted with language polishing and simulated peer review of the manuscript. AI was not used to generate, rewrite, or annotate the experimental dataset content. The authors determined which results were sufficiently validated for inclusion.

Validation and responsibility. The authors checked mathematical derivations by hand and used additional AI tools for review. Core code not written by the authors was read and tested. Numerical verification routines were assessed primarily through tests. Tests covered analytic identities, enclosure arithmetic, structured operators, and original-graph consistency, as described in Appendix E.3 and the supplementary code documentation. The authors take responsibility for the final mathematical claims, implementation, experimental protocol, reported results, and scientific conclusions.

References

Gregory Bonaert, Dimitar I. Dimitrov, Maximilian Baader, and Martin Vechev. Fast and precise certification of transformers. In *ACM SIGPLAN International Conference on Programming Language Design and Implementation*, pp. 466–481, 2021. doi: 10.1145/3453483.3454056.

Mohammed Bouri and Adnane Saoud. Bridging robustness and generalization against word substitution attacks in NLP via the growth bound matrix approach. In *Findings of the Association*

for Computational Linguistics: ACL 2025, pp. 12118–12137, 2025. doi: 10.18653/v1/2025.findings-acl.627.

Jyot Buch, Murat Arcak, and Peter Seiler. An efficient algorithm to compute norms for finite horizon, linear time-varying systems. *IEEE Control Systems Letters*, 5(5):1597–1602, 2021. doi: 10.1109/LCSYS.2020.3042075.

Liang Cao, Weide Liu, Zhuo Chen, and Yan Qin. A control-theoretic view of Mamba on stability and robustness. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=kLFm50q9kO.

Hong-Ming Chiu, Hao Chen, Huan Zhang, and Richard Y. Zhang. SDP-CROWN: Efficient bound propagation for neural network verification with tightness of semidefinite programming. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10449–10468, 2025. URL https://proceedings.mlr.press/v267/chiu25a.html.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Sumanth Dathathri, Krishnamurthy Dvijotham, Alexey Kurakin, Aditi Raghunathan, Jonathan Uesato, Rudy R. Bunel, Shreya Shankar, Jacob Steinhardt, Ian Goodfellow, Percy S. Liang, and Pushmeet Kohli. Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5318–5331, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/397d6b4c83c91021fe928a8c4220386b-Abstract.html.

Taha Entesari and Mahyar Fazlyab. Hierarchical end-to-end Taylor bounds for complete neural network verification. In *Learning for Dynamics and Control Conference*, volume 331 of *Proceedings of Machine Learning Research*, pp. 1070–1087, 2026. URL https://proceedings.mlr.press/v331/entesari26a.html.

Mahyar Fazlyab, Manfred Morari, and George J. Pappas. Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming. *IEEE Transactions on Automatic Control*, 67(1):1–15, 2022. doi: 10.1109/TAC.2020.3046193.

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

John Timothy Halloran, Manbir S. Gulati, and Paul F. Roysdon. Mamba state-space models are Lyapunov-stable learners. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=wzsYQYs3dO.

Chao Huang, Jiameng Fan, Xin Chen, Wenchao Li, and Qi Zhu. POLAR: A polynomial arithmetic framework for verifying neural-network controlled systems. In *Automated Technology for Verification and Analysis*, pp. 414–430, 2022. doi: 10.1007/978-3-031-19992-9_27.

Ching-Yun Ko, Zhaoyang Lyu, Lily Weng, Luca Daniel, Ngai Wong, and Dahua Lin. POPQORN: Quantifying robustness of recurrent neural networks. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3468–3477, 2019. URL https://proceedings.mlr.press/v97/ko19a.html.

Niklas Kochdumper and Matthias Althoff. Sparse polynomial zonotopes: A novel set representation for reachability analysis. *IEEE Transactions on Automatic Control*, 66(9):4043–4058, 2021. doi: 10.1109/TAC.2020.3024348.

Niklas Kochdumper, Christian Schilling, Matthias Althoff, and Stanley Bak. Open- and closed-loop neural network verification using polynomial zonotopes. In *NASA Formal Methods*, pp. 16–36, 2023. doi: 10.1007/978-3-031-33170-1_2.

Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HwCvaJOiCj.

Kyoko Makino and Martin Berz. Taylor models and other validated functional inclusion methods. *International Journal of Pure and Applied Mathematics*, 4(4):379–456, 2003. URL https://www.bmtdynamics.org/pub/papers/TMIJPAM03/TMIJPAM03.pdf.

Matthew Newton and Antonis Papachristodoulou. Exploiting sparsity for neural network verification. In *Learning for Dynamics and Control Conference*, volume 144 of *Proceedings of Machine Learning Research*, pp. 715–727, 2021. URL https://proceedings.mlr.press/v144/newton21a.html.

Siegfried M. Rump. Verified bounds for singular values, in particular for the spectral norm of a matrix and its inverse. *BIT Numerical Mathematics*, 51(2):367–384, 2011. doi: 10.1007/s10543-010-0294-0.

Wonryong Ryou, Jiayu Chen, Mislav Balunovic, Gagandeep Singh, Andrei Dan, and Martin Vechev. Scalable polyhedral verification of recurrent neural networks. In *Computer Aided Verification*, pp. 225–248, 2021. doi: 10.1007/978-3-030-81685-8_10.

Peter Seiler, Robert Moore, Chris Meissen, Murat Arcak, and Andrew Packard. Finite horizon robustness analysis of LTV systems using integral quadratic constraints. *Automatica*, 100:135–143, 2019. doi: 10.1016/j.automatica.2018.11.009.

Sina Sharifi and Mahyar Fazlyab. Provable bounds on the Hessian of neural networks: Derivative-preserving reachability analysis. arXiv preprint arXiv:2406.04476, 2024. URL https://arxiv.org/abs/2406.04476.

Keyi Shen and Glen Chou. Parallel differentiable reachability for learning and planning with certified neural dynamics and controllers. In *Robotics: Science and Systems XXII*, 2026. URL https://trustworthyrobotics.github.io/diffreach_site/.

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness verification for transformers. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxwPJHFwS.

Zhouxing Shi, Qirui Jin, Zico Kolter, Suman Jana, Cho-Jui Hsieh, and Huan Zhang. Neural network verification with branch-and-bound for general nonlinearities. In *Tools and Algorithms for the Construction and Analysis of Systems*, pp. 315–335, 2025. doi: 10.1007/978-3-031-90643-5_17.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1129–1141, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0cbc5671ae26f67871cb914d81ef8fc1-Abstract.html.

Anton Xue, Lars Lindemann, and Rajeev Alur. Chordal sparsity for SDP-based neural network verification. *Automatica*, 161:111487, 2024. doi: 10.1016/j.automatica.2023.111487.

Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Skxuk1rFwB.

Nikola Zubić and Davide Scaramuzza. Regularity and stability properties of selective SSMs with discontinuous gating. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=7Vav53cDeN.

A POLYNOMIAL ENCLOSURES AND DIRECT CERTIFICATION

The rules below complete the construction in Section 3: they enclose omitted terms, establish readout containment and cost bounds, and give the RMS and direct-certificate derivations.

A.1 ENCLOSURE OPERATIONS AND SHARED REMAINDERS

A.1.1 SHARED REMAINDER FUNCTIONS

The representation in Section 3.2 stores each remainder's function identifier, value enclosure, and fixed linear maps. Linear operations compose these maps; addition combines maps with the same identifier before bounding their image. Distinct remainder functions are enclosed separately. Nonlinear operations create new remainders enclosing inherited contributions, discarded products, and approximation error.

For remainder functions $r_o(\xi) \in R_o$ with downstream maps $\mathsf{L}_{o,j}$, using one value of r_o at all its occurrences gives a subset of the sums obtained by choosing each occurrence independently. For centered boxes $R_o = \{r : |r| \le \rho_o\}$, the corresponding exact support radii in direction a are

$$B_{\mathrm{shared}}(a) = \sum_o \left\| \mathrm{diag}(\rho_o) \left(\sum_j \mathsf{L}_{o,j} \right)^T a \right\|_1,$$
$$B_{\mathrm{split}}(a) = \sum_{o,j} \left\| \mathrm{diag}(\rho_o) \mathsf{L}_{o,j}^T a \right\|_1.$$

The box support formula and triangle inequality give $B_{\mathrm{shared}}(a) \le B_{\mathrm{split}}(a)$. A remainder's contribution vanishes when its combined map vanishes in direction a, regardless of its box width.

A.1.2 PRODUCTS AND NONLINEARITIES

For polynomial enclosures $x = x_0 + x_1 + x_2 + r_x$ and $y = y_0 + y_1 + y_2 + r_y$, the retained product is

$$[xy]_{\le 2} = x_0 y_0 + (x_0 y_1 + x_1 y_0) + (x_0 y_2 + x_2 y_0 + x_1 y_1). \tag{7}$$

Subtracting this expression from the full product gives

$$r_{xy} = x_0 r_y + r_x y_0 + x_1(y_2 + r_y) + (x_2 + r_x) y_1 + (x_2 + r_x)(y_2 + r_y). \tag{8}$$

The last term includes products between inherited remainders. Bounding every term on the full perturbation box gives the product enclosure used in injection, gating, and matrix multiplication, with factor order preserved.

Unary nonlinearities. For $y = f(u)$, expand about u_0 and retain terms through degree two in ξ. The smooth rule requires $f \in C^3$ on a neighborhood of an interval I containing every segment from u_0 to $u(\xi)$. A bound $M_3 \ge \sup_I |f'''|$ bounds the analytic remainder by $M_3|u - u_0|^3/6$ in magnitude; inherited remainders and discarded products are enclosed as well. If I crosses a nonsmooth threshold, use a polynomial approximation with a validated error over all of I instead. Each operation must also satisfy its domain conditions, such as positivity for inverse square roots.

A polynomial enclosure for an operand w, possibly matrix-valued after vectorization, is $w = w_0 + w_1 + w_2 + r_w$ with $r_w(\xi) \in \mathcal{R}_w$, as in Eq.(2).

Lemma 2 (Contracted second-order polynomial and remainder enclosure)**.** *Let f be C^3 on an open set containing every segment $w_0 + \lambda(w(\xi) - w_0)$, $\lambda \in [0,1]$, $\xi \in [-1,1]^s$. With derivatives at w_0, set*

$$P_f = f(w_0) + Df[w_1] + Df[w_2] + \tfrac{1}{2} D^2 f[w_1, w_1], \qquad \delta_{\mathrm{hi}} = w_2 + r_w, \quad \delta = w_1 + \delta_{\mathrm{hi}}. \tag{9}$$

Then

$$f(w) - P_f = Df[r_w] + D^2 f[w_1, \delta_{\mathrm{hi}}] + \tfrac{1}{2} D^2 f[\delta_{\mathrm{hi}}, \delta_{\mathrm{hi}}] + \mathcal{T}_3,$$
$$\mathcal{T}_3 = \tfrac{1}{2} \int_0^1 (1-\lambda)^2 D^3 f(w_0 + \lambda\delta)[\delta, \delta, \delta]\, d\lambda. \tag{10}$$

A bound covering every term on the right gives a valid output polynomial enclosure; in particular, an interval enclosure of the directional third derivative over the entire segment, multiplied by $1/6$*, encloses* $\mathcal{T}_3$*.*

Proof. Apply the integral Taylor formula to $f(w_0+\delta)$ through degree two, substitute $\delta = w_1 + w_2 + r_w$, and subtract the polynomial P_f in Eq.(9). Symmetry of the second derivative combines the two mixed terms into $D^2 f[w_1, \delta_{\mathrm{hi}}]$. □

A.1.3 Refinement by intersection

Two enclosures of the same graph quantity may be intersected. When refinement is enabled, the implementation does so whenever an independent value enclosure $\mathcal{G}$ is available. For a polynomial enclosure $v = c + v_1 + v_2 + r_v$ with $\mathcal{G} \ni v(\xi)$, write $r_v(\xi) \in \mathcal{R}_v$ for its current value enclosure. Then

$$\mathcal{R}_v \leftarrow \mathcal{R}_v \cap \big(\mathcal{G} - c - v_1([-1,1]^s) - v_2([-1,1]^s)\big). \tag{11}$$

The function $r_v(\xi)$ is unchanged; the operation updates its value enclosure $\mathcal{R}_v$. Available ranges include the decay bound from the enclosed exponential weight and the RMS projection bound in Appendix A.4. The same rule applies at the layer-input interface of Eq.(23): intersecting refines the domain for original-graph propagation, retaining affine dependence on ξ with the quadratic part and remainder enclosed together as a set.

A.2 Readout construction and soundness

A.2.1 Compact transition products

For a recurrent head as in Section 3.1, use the recurrence stated there on the affected suffix. In the matrix-state notation, $h_t \in \mathbb{R}^{n\times d_v}$, $B_{t,a} \in \mathbb{R}^{n\times r_{t,a}}$, $U_{t,a} \in \mathbb{R}^{r_{t,a}\times d_v}$, and $C_t \in \mathbb{R}^{n\times d_o}$ may all depend on the same perturbation vector ξ. Vector states are the case $d_v = 1$. An injection indexed by t may use earlier inputs; its index specifies where it enters. Induction gives Eq.(4) without any commutation assumption.

Definition 1 (Compact state-transition-product parameterization)**.** A compact parameterization represents $\Phi_{r:t} = \Psi(\theta_{r:t})$ using a fixed matrix-valued map Ψ on the perturbation box, including prefix and empty products, with $d_\theta \le c_0 + c_1 n$ independent of H. It supplies a sound parameter accumulation rule with per-update time and storage independent of interval length; contracted evaluators for $C^T\Psi B$, $C^T\partial_a\Psi B$, and $C^T\partial_{ab}\Psi B$; and enclosures of the derivative contractions needed for omitted terms, including third-order directional terms. The evaluators have declared resource bounds and do not propagate individual hidden-state-coordinate polynomials in ξ.

Here ∂_a differentiates a transition parameter; operand enclosures carry its dependence on ξ.

A.2.2 Contracted polynomial and containment

Lemma 2 requires smoothness of $C^T\Psi(\theta)B$ with respect to its operands. Piecewise selectors may still supply valid polynomial enclosures for those operands. For $f = \mathcal{K}$ and $w = (\operatorname{vec} C, \operatorname{vec} B, \theta)$, let $E = \Psi(\theta_0)$, $E_a = \partial_a\Psi(\theta_0)$, and $E_{ab} = \partial_a\partial_b\Psi(\theta_0)$. The retained pieces are

$$\begin{aligned} K_0 &= C_0^T E B_0, \qquad K_1 = C_1^T E B_0 + C_0^T E B_1 + \textstyle\sum_a \theta_{1,a} C_0^T E_a B_0, \\ K_2 &= C_2^T E B_0 + C_0^T E B_2 + C_1^T E B_1 + \textstyle\sum_a \theta_{2,a} C_0^T E_a B_0 \\ &\quad + \textstyle\sum_a \theta_{1,a}(C_1^T E_a B_0 + C_0^T E_a B_1) + \tfrac{1}{2}\sum_{a,b}\theta_{1,a}\theta_{1,b} C_0^T E_{ab} B_0. \end{aligned} \tag{12}$$

Products are contracted over the state index before storing output coefficients. Lemma 2 encloses their tail using full-segment derivative bounds, with coefficient-arithmetic error added. Componentwise derivative bounds followed by contraction are also valid, but can lose cancellation.

Theorem 3 (Structured history/readout contraction)**.** *Let a head use the transition families and recurrence of Section 3.1, so that Eq.(4) holds on the perturbation box.*

(i) Accumulated structure and contracted evaluation. *Each transition form in that section generates its history products from additively accumulated parameters:* $\Phi_{r:t} = \Psi(\theta_{r:t})$ *with* $\theta_{r:t} =$

$\sum_{v=r+1}^{t} \gamma_v$, *where* γ_v *collects the generator coefficients of step* v *and* $d_\theta = \dim \theta$ *is independent of* H *and at most linear in* n*, namely* $d_\theta = 1$ *for a diagonal or scalar-identity generator and* $d_\theta = m_{\mathrm{rot}} + 1 \le n/2 + 1$ *for a scalar decay with* m_{rot} *fixed rotation planes. For operands* $C \in \mathbb{R}^{n \times r_C}$ *and* $B \in \mathbb{R}^{n \times r_B}$*, the contracted map* $\mathcal{K}(C, B, \theta) = C^T \Psi(\theta) B$*, all of its potentially nonzero first and second parameter derivatives, and interval enclosures of its third-order directional terms are evaluated in* $O(n r_C r_B)$ *arithmetic operations and* $O(n)$ *elementary-function evaluations, with* $O(n(r_C + r_B) + r_C r_B + d_\theta)$ *working storage. None of these evaluations forms or propagates a polynomial in the input perturbation variables for an individual hidden-state coordinate.*

(ii) Containment. *Assume valid polynomial enclosures in the common input variables for the prefix, readouts, injection operands and signals, direct terms, and accumulated parameters, together with derivative enclosures valid for the map on each full operand segment. Expand each contracted map through degree two. Enclose its Taylor remainder, inherited operand remainders, discarded higher-degree products, and coefficient-arithmetic errors. Multiply each resulting polynomial enclosure by the injection-signal enclosure, then sum all terms in Eq.(4). The result is a polynomial enclosure containing* o_t *on the perturbation box, and successive valid gate, RMS, affine, and scan operations preserve containment through the network.*

(iii) Taylor coefficients. *If the operands are twice differentiable at* $\xi = 0$ *and their Taylor coefficients through degree two in* ξ *are exact, exact arithmetic yields the degree-two Taylor polynomial of* o_t*. For general polynomial enclosures, including threshold-crossing envelopes, the conclusion is containment.*

The same statement holds for any recurrence that admits a compact parameterization of its state-transition products (Definition 1), with its declared parameter and evaluation bounds in place of (i).

Proof. Induction on the recurrence of Section 3.1 gives Eq.(4), including the prefix term with $B = h_{-1}$. Replacing each transition product by $\Psi(\theta_{r:t})$ gives the contracted representation. For the supported families, the parameter accumulation, derivative contractions, and resource bounds are established in Appendix A.3, proving (i).

Lemma 2 gives containment for each contracted map. Multiplication by its injection-signal enclosure uses Eq.(8), with matrix factor order preserved. Summing these enclosures, including coefficient-arithmetic error, proves (ii) and Theorem 1. Induction through valid gate, RMS, affine, residual, and scan operations gives network-wide containment.

Eq.(12) forms polynomial products after the state contractions. Under the smoothness and exactness assumptions in (iii), the chain and product rules identify the retained terms with the degree-two Taylor polynomial of o_t. □

A.3 Transition families and computational cost

Fixed commuting generators. For fixed commuting matrices Λ_a and $F_t = \exp(\sum_a \gamma_{t,a} \Lambda_a)$, the identity $\exp(A)\exp(B) = \exp(A + B)$ gives $\Phi_{r:t} = \Psi(\theta_{r:t})$ with $\theta_{r:t,a} = \sum_{v=r+1}^{t} \gamma_{v,a}$ and $\Psi(\theta) = \exp(\sum_a \theta_a \Lambda_a)$, including the prefix sum from $v = 0$ and zero parameters for the empty product. The derivatives are $\partial_a \Psi = \Psi \Lambda_a$, $\partial_{ab} \Psi = \Psi \Lambda_a \Lambda_b$, and analogously through order three. The following instances provide the evaluators and resource bounds required by Definition 1.

A.3.1 Diagonal transitions

For a Mamba-1 channel j, let $\delta_{r:t,j} = \sum_{v=r+1}^{t} \Delta_{v,j}$, with empty sums zero and $\delta_{-1:t,j} = \sum_{v=0}^{t} \Delta_{v,j}$. The injection amplitude is $b^{\mathrm{inj}}_{r,j} = \Delta_{r,j} u_{r,j}$, and the contracted scalar map is $\kappa_j(v, \delta) = \sum_{\nu=1}^{n} [v]_\nu e^{a_{j,\nu} \delta}$. Here $\Psi(\delta) = \mathrm{diag}(e^{a_j \delta})$, and Eq.(4) becomes

$$o_{t,j} = D_j u_{t,j} + \kappa_j(C_t \odot h_{-1,j}, \delta_{-1:t,j}) + \sum_{r=0}^{t} b^{\mathrm{inj}}_{r,j} \kappa_j(B_r \odot C_t, \delta_{r:t,j}). \tag{13}$$

For a scalar-identity transition, the contraction reduces to $e^{a_h \delta} C_t^T B_r$.

To enclose each contracted map, first form an enclosure of the componentwise product $v = C_t \odot B_r$, or $v = C_t \odot h_{-1,j}$ for the prefix term. Write $v = v_0 + v_1 + v_2 + r_v$ and $\delta = \delta_0 + \delta_1 + \delta_2 + r_\delta$.

For the fixed channel j, define the signed state sums $\mathcal{S}_\ell(z) = \sum_\nu a_{j,\nu}^\ell e^{a_{j,\nu}\delta_0}[z]_\nu$ for $\ell = 0, 1, 2$. Applying Eq.(3) jointly to v and δ gives

$$\begin{aligned}[\kappa_j(v,\delta)]_{\le 2} = \mathcal{S}_0(v_0) + \mathcal{S}_0(v_1) + \mathcal{S}_1(v_0)\delta_1 \\ + \mathcal{S}_0(v_2) + \mathcal{S}_1(v_0)\delta_2 + \delta_1\mathcal{S}_1(v_1) + \tfrac{1}{2}\delta_1^2\mathcal{S}_2(v_0).\end{aligned} \tag{14}$$

For the remainder derivation, set $v_{\mathrm{hi}} = v_2 + r_v$, $\delta_{\mathrm{hi}} = \delta_2 + r_\delta$, and $b = \delta - \delta_0$. The following specializes Lemma 2 modewise.

Expansion. Apply Taylor's formula to each exponential about δ_0 through second order. After multiplying by the full modal operand and summing, the third-order term uses the third-derivative contraction

$$\mathcal{D}_3(v,\delta) = \sum_\nu a_{j,\nu}^3 e^{a_{j,\nu}(\delta_0+\vartheta_\nu b)}[v]_\nu, \qquad \vartheta_\nu \in [0,1]. \tag{15}$$

A different ϑ_ν is permitted in each modal remainder.

Subtracting the retained polynomial. Subtracting the retained polynomial in Eq.(14) from the modal Taylor expansion yields the exact remainder identity

$$\begin{aligned}r_\kappa = \mathcal{S}_0(r_v) + \mathcal{S}_1(v_0)r_\delta + \delta_1\mathcal{S}_1(v_{\mathrm{hi}}) + \delta_{\mathrm{hi}}\mathcal{S}_1(v_1) + \delta_{\mathrm{hi}}\mathcal{S}_1(v_{\mathrm{hi}}) \\ + \mathcal{S}_2(v_0)(\delta_1\delta_{\mathrm{hi}} + \tfrac{1}{2}\delta_{\mathrm{hi}}^2) + \tfrac{1}{2}b^2\mathcal{S}_2(v - v_0) + \tfrac{1}{6}b^3\mathcal{D}_3(v,\delta).\end{aligned} \tag{16}$$

Enclosing the remainder. The exponential derivative enclosure covers the full accumulated-step segment and full operand range. The signed maps $\mathcal{S}_0$ and $\mathcal{S}_1$ act on retained coefficients before evaluating support bounds. For $\mathcal{S}_2(v - v_0)$, a contracted coefficient enclosure and a direct range enclosure can both be formed and intersected. The remaining products and modal derivative ranges are enclosed by valid interval operations.

Cancellation before nonlinear enclosure. The following proposition formalizes the two-step cancellation example in Section 3.3 and its effect over a longer history.

Proposition 4 (Separation under marginal Taylor remainders). *Let $H = 2m \ge 2$, $a > \varepsilon > 0$, $\kappa > 0$, and $\xi \in [-1, 1]$. Consider the scalar recurrence $h_0 = 1$ and*

$$h_t = e^{-\Delta_t(\xi)}h_{t-1}, \qquad \Delta_{2j-1}(\xi) = a + \varepsilon\xi, \quad \Delta_{2j}(\xi) = a - \varepsilon\xi \quad (1 \le j \le m).$$

The margin $\phi_H(\xi) = (1+\kappa)e^{-Ha} - h_H(\xi)$ satisfies $\min_{\xi\in[-1,1]}\phi_H(\xi) = \kappa e^{-Ha} > 0$. Define the marginal Taylor remainders by

$$P_\pm(\xi) = 1 \pm \varepsilon\xi + \tfrac{1}{2}\varepsilon^2\xi^2, \quad r_\pm(\xi) = e^{\pm\varepsilon\xi} - P_\pm(\xi), \quad \mathcal{R}_\pm = \{r_\pm(z) : z \in [-1,1]\}.$$

Retain the common ξ in $P_\pm$, but replace $r_\pm(\xi)$ by variables $u_\pm \in \mathcal{R}_\pm$ independent of ξ and each other, reusing the same $u_\pm$ at every occurrence of its sign. The relaxed state is

$$\widetilde{h}_H(\xi, u_-, u_+) = e^{-Ha}\big[(P_-(\xi) + u_-)(P_+(\xi) + u_+)\big]^m.$$

With $\rho_{\exp} = e^\varepsilon - 1 - \varepsilon - \varepsilon^2/2 > 0$, its margin obeys

$$\min_{\xi\in[-1,1],\, u_\pm\in\mathcal{R}_\pm}\big[(1+\kappa)e^{-Ha} - \widetilde{h}_H(\xi, u_-, u_+)\big] \le e^{-Ha}\big[1 + \kappa - (1+\rho_{\exp})^H\big].$$

Consequently, this relaxation admits no positive margin lower bound whenever $H\log(1+\rho_{\exp}) \ge \log(1+\kappa)$.

Proof. Each consecutive pair has sum $\Delta_{2j-1}(\xi) + \Delta_{2j}(\xi) = 2a$, so $h_H(\xi) = e^{-Ha}$. Choosing $u_\pm = r_\pm(\xi)$ recovers this true state, making the stated set a relaxation. Since $r_+(1) = r_-(-1) = \rho_{\exp}$, both marginal ranges contain $\rho_{\exp}$. At $\xi = 0$, $P_+(0) = P_-(0) = 1$, so the allowed choice $u_+ = u_- = \rho_{\exp}$ yields $\widetilde{h}_H = e^{-Ha}(1+\rho_{\exp})^H$ and the displayed bound. The choice is not realizable by the original recurrence, where $r_+(0) = r_-(0) = 0$. □

The separation applies to marginal remainder ranges without joint constraints; a stepwise method that retains the relation among r_+, r_-, and ξ can also recover the cancellation. Here $\sum_t \Delta_t = Ha$ exposes that relation before enclosing the exponential.

A.3.2 Rotation, decay, and delayed injection

For the joint recurrence in Corollary 1, each injection v_r contributes through its current step and the next step. The identity $\Phi_{r+1:t}F_{r+1} = \Phi_{r:t}$ gives the paired history

$$\begin{aligned} h_t &= \Phi_{-1:t}h_{-1} + (1-\lambda_0)\Delta_0\Phi_{-1:t}v_{-1} + \sum_{r=0}^{t} w_{r,t}\Phi_{r:t}v_r, \\ w_{r,t} &= \lambda_r\Delta_r + \mathbf{1}_{r<t}(1-\lambda_{r+1})\Delta_{r+1}. \end{aligned} \tag{17}$$

The delayed v_{-1} term depends on perturbed step-zero parameters and is enclosed with the suffix. Fixed-plane phase increments and input-dependent decay receive polynomial enclosures in the shared ξ; sine and cosine are enclosed on the full phase ranges. Finite-rank injection and readout are contracted as $C_t^T\Psi(\theta_{r:t})B_rU_r$, retaining all rank indices and cross terms through Eq.(12) and Eq.(8).

A.3.3 Evaluation cost and combined recurrence

For m_{rot} disjoint fixed planes and scalar decay, $d_\theta = m_{\mathrm{rot}} + 1 \le n/2 + 1$. Streaming the scalar modes and two-dimensional blocks evaluates the contracted value and all potentially nonzero first and second parameter derivatives in $O(nr_Cr_B)$ operations, without forming a dense transition; the working storage is $O(n(r_C+r_B)+r_Cr_B+d_\theta)$. Mixed derivatives of distinct phases vanish because the planes are disjoint. Degree-two substitution in s shared perturbation variables still retains their cross terms and costs $O(nr_Cr_B(s+1)^2)$ for the contracted arithmetic. Full-segment derivative remainders use interval bounds on the operand segments, exponential, sine, and cosine, evaluated blockwise at fixed derivative order.

Corollary 1 (Joint rotation–decay, trapezoidal, finite-rank recurrence)**.** *For each head, let*

$$\begin{aligned} &F_t = e^{\Delta_t a_t}\operatorname{diag}_j \operatorname{Rot}(\Delta_t\omega_{t,j}), \qquad v_t = B_tU_t, \quad B_t \in \mathbb{R}^{n\times r_B},\ U_t \in \mathbb{R}^{r_B\times d_v}, \\ &h_t = F_th_{t-1} + (1-\lambda_t)\Delta_tF_tv_{t-1} + \lambda_t\Delta_tv_t, \qquad o_t = d_t + C_t^Th_t, \quad C_t \in \mathbb{R}^{n\times r_C}, \end{aligned} \tag{18}$$

where $a_t, \omega_t, \Delta_t, \lambda_t, B_t, U_t, C_t$ may all depend on the same perturbation vector ξ, the rotation planes are fixed, and r_B, r_C are finite. Given valid polynomial enclosures for these parameters, the prefix state, and the delayed injection, together with full-segment enclosures, the recurrence satisfies Theorem 3 with

$$\theta_{r:t} = \Big(\sum_{v=r+1}^{t}\Delta_va_v,\ \big(\textstyle\sum_{v=r+1}^{t}\Delta_v\omega_{v,j}\big)_j\Big) \tag{19}$$

and the two-path weights of Eq.(17).

Proof. Input-dependent scalar decay and fixed-plane rotations share fixed commuting generators, which gives Eq.(19), and the blockwise evaluators above satisfy the complexity conditions. The identity $\Phi_{r+1:t}F_{r+1} = \Phi_{r:t}$ holds for the full rotation–decay matrices, and the scalar weights commute with them, so Eq.(17) holds with all mechanisms present, including the delayed-prefix term. Substituting $v_r = B_rU_r$ and applying C_t^T gives $w_{r,t}(C_t^T\Psi(\theta_{r:t})B_r)U_r$ and the prefix terms, and the matrix-valued polynomial-enclosure rule retains all rank indices and all cross terms between decay, phase, weights, selectors, and injection signals in the same perturbation variables. Theorem 3 gives polynomial-enclosure containment. □

A.4 RMS enclosure and cancellation

A.4.1 Polynomial and complete remainder

Write $F(x) = Wxf(\upsilon)$, where W includes the fixed learned weights and projection, $\upsilon = d_{\mathrm{rms}}^{-1}\|x\|^2 + \eta$, $f(\upsilon) = \upsilon^{-1/2}$, and $\eta > 0$. For $x = x_0 + x_1 + x_2 + r$, define $\upsilon_0 = d_{\mathrm{rms}}^{-1}\|x_0\|^2 + \eta$, $\upsilon_1 = 2x_0^Tx_1/d_{\mathrm{rms}}$, and $\upsilon_2 = (2x_0^Tx_2 + \|x_1\|^2)/d_{\mathrm{rms}}$. Let $f_0 = f(\upsilon_0)$, $f_0' = f'(\upsilon_0)$, and $f_0'' = f''(\upsilon_0)$. The retained polynomial has degree components

$$\begin{aligned} F_0 &= Wx_0f_0, \qquad F_1 = Wx_1f_0 + Wx_0f_0'\upsilon_1, \\ F_2 &= Wx_2f_0 + Wx_1f_0'\upsilon_1 + Wx_0(f_0'\upsilon_2 + \tfrac{1}{2}f_0''\upsilon_1^2). \end{aligned} \tag{20}$$

The linear contribution of the inherited remainder is given by Eq.(6). Combining the numerator and denominator maps before bounding their image of r retains cancellation in this linear contribution. The remainder identity in Lemma 2, with $w = x$, also includes $D^2F(x_0)[x_1, x_2 + r]$, $\frac{1}{2}D^2F(x_0)[x_2 + r, x_2 + r]$, and the integral third-order Taylor remainder. The mixed terms are enclosed on $\mathcal{B}$, and the third-derivative bounds cover every segment $x_0 + \lambda(x(\xi) - x_0)$, $0 \le \lambda \le 1$, $\xi \in \mathcal{B}$. The positive stabilizer ensures smoothness along these segments.

A.4.2 Radial and tangent perturbations in RMS

For the map in Eq.(6), write $b = d_{\text{rms}}^{-1}\|x_0\|_2^2$. Suppose one scalar shared remainder function produces a perturbation in the direction of the expansion point x_0, $r = \alpha_{\text{rad}}x_0$, $|\alpha_{\text{rad}}| \le \bar{\alpha}_{\text{rad}}$; this is the radial direction. For an output direction a, let B_{shared} be the maximum absolute projection of the summed linear contributions when both use the same $\alpha_{\text{rad}} \in [-\bar{\alpha}_{\text{rad}}, \bar{\alpha}_{\text{rad}}]$. Let B_{split} be the sum of the separate maximum absolute projections, allowing the two contributions to choose their values of α_{rad} independently. The two radii equal $\bar{\alpha}_{\text{rad}}|a^T W x_0|/(b+\eta)^{3/2}$ times η and $2b + \eta$, respectively. Thus, whenever $B_{\text{split}} > 0$,

$$\frac{B_{\text{shared}}}{B_{\text{split}}} = \frac{\eta}{2b + \eta}. \tag{21}$$

If $B_{\text{split}} = 0$, both directional contributions vanish. For a remainder perpendicular to x_0, called tangent, $x_0^T r = 0$: the denominator contribution vanishes, leaving only the numerator contribution.

Independent value bound. Cauchy–Schwarz gives

$$\left| \frac{(Wx)_i}{\sqrt{d_{\text{rms}}^{-1}\|x\|_2^2 + \eta}} \right| \le \sqrt{d_{\text{rms}}}\, \|W_{i:}\|_2 \qquad \text{for every } x.$$

This bound can refine the remainder by Eq.(11), independently of the polynomial expansion.

A.5 Direct certificate and relaxation error

With classifier weights W_y, W_k and biases b_y, b_k, we combine the classification head's signed contributions from the residual branch and gated readout before bounding the final residual block. With x the residual-branch input and $g \odot o$ the gated readout,

$$\begin{aligned} \phi &= (W_y - W_k)^T x + (W_y - W_k)^T W_{\text{out}}(g \odot o) + b_y - b_k \\ &= c_\phi + a_\phi^T \xi + \sum_{i \le j} q_{\phi,ij}\xi_i\xi_j + r_\phi(\xi). \end{aligned} \tag{22}$$

Combining the signed coefficients first preserves cancellations between these paths. We then bound the constant, linear, quadratic, and remainder terms separately.

Quadratic lower bound. For coefficients $q_{\phi,ij}$ with total coefficient-error bound δ_q on the box, the coefficient-sign lower bound is

$$\sum_i \min(q_{\phi,ii}, 0) - \sum_{i<j} |q_{\phi,ij}| - \delta_q.$$

Indeed, $\xi_i^2 \in [0, 1]$ and $|\xi_i\xi_j| \le 1$ bound the diagonal and off-diagonal terms, respectively. The implementation takes the larger of this bound and the lower endpoint of an independently propagated enclosure of the same quadratic part.

Relaxation error. Intersecting with an independently propagated quadratic range can correct looseness of the coefficient-sign bound, but direct certification still bounds the linear and quadratic parts separately. For example, on $[-1, 1]$ this gives -1 for $\xi + \xi^2$, whose true minimum is $-1/4$.

B RESPONSE AND ENERGY IDENTITIES FOR FIXED QUADRATIC CONSTRAINTS

Polynomial enclosures supply ranges for local envelopes, which define a linear response and a quadratic energy in shared residual coordinates. The constructions below preserve both under invertible coordinate changes, using temporal and RMS structure to evaluate the resulting operators. Appendices D and E turn this fixed representation into a checked bound for the original network.

B.1 LOCAL CONSTRAINTS AND SHARED RESIDUALS

A quadratic constraint (QC) bounds a nonlinear residual using its inputs. For products, we use the normalized residual and two parent constraints specified in Section 4.1. Both inequalities constrain the same residual, which is reused at every later occurrence of the product. A zero input radius requires no normalized product residual.

The polynomial enclosures supply the ranges needed to construct such local inequalities. At a layer input $x = c_{\mathrm{dom}} + A_{\mathrm{dom}}\xi + x_2 + r_x$, Full-QC uses

$$x(\xi) - c_{\mathrm{dom}} - A_{\mathrm{dom}}\xi \in \mathcal{R}_{\mathrm{dom}}, \qquad \mathcal{R}_{\mathrm{dom}} \supseteq \{x_2(\xi) + r_x(\xi) : \xi \in [-1,1]^s\}. \tag{23}$$

Imported affine bounds tighten the node ranges used to construct local QCs; the QC response is compiled from the original graph.

Local envelopes and common residual assignments. Each active nonlinear operation contributes a linear response row and one shared error coordinate. Consider a finite causal graph with one scalar row per vector component on the input box with deviations $d_a(\xi)$ from fixed node centers and input rows $d_{\mathrm{in}} = P\xi + b_{\mathrm{in}}(\xi)$.

Proposition 5 (Compilation of compatible local envelopes)**.** *For each active nonlinear node a, suppose a valid local envelope on the established enclosure of that operation's inputs supplies, for every original execution, one scalar ϵ_a and one $\delta_a \in \mathcal{A}_a$ with*

$$d_a = \sum_{b\in \mathrm{pa}(a)} \alpha_{ab} d_b + \sigma_a \epsilon_a + \delta_a, \qquad \epsilon_a^2 \le s_{a,r}^2, \quad s_{a,r} = \sum_b \kappa_{ar,b} d_b, \quad r \in \mathcal{I}_a, \tag{24}$$

Here $\mathrm{pa}(a)$ indexes predecessors and $\mathcal{I}_a$ indexes constraints on the same residual ϵ_a. The coefficients $P, \alpha, \sigma, \kappa$ are fixed, σ_a is the residual amplitude, and constraint rows use available predecessor signals, with any auxiliary signal included in the graph. Inactive or affine nodes have no error coordinate and keep their residual in δ_a. Choose $w_a > 0$ and $\lambda_{a,r} > 0$, put $e_a = w_a\epsilon_a$, and compile

$$\begin{aligned} &\widehat{d}_{\mathrm{in}} = P\xi, && \widehat{d}_a = \textstyle\sum_b \alpha_{ab}\widehat{d}_b + (\sigma_a/w_a)e_a, \quad b_a = \textstyle\sum_b \alpha_{ab} b_b + \delta_a, \\ &p_{a,r} = \sqrt{\lambda_{a,r}} \textstyle\sum_b \kappa_{ar,b}\widehat{d}_b, && (b_p)_{a,r} = \sqrt{\lambda_{a,r}} \textstyle\sum_b \kappa_{ar,b} b_b, \end{aligned} \tag{25}$$

omitting the error term at inactive nodes. Then:

1. *For every ξ, choosing the local witnesses once gives a common witness $e(\xi)$ with $d_a = \widehat{d}_a(\xi, e(\xi)) + b_a(\xi)$ at every node and $p_{a,r} + (b_p)_{a,r} = \sqrt{\lambda_{a,r}} s_{a,r}$ for all rows simultaneously.*
2. *For any partition whose block dependency graph is acyclic, including dependencies of response and constraint rows, assign each active error and its constraint rows to one block and export every response coordinate used by another block or the property. Then $J = \|p\|^2 - \|e\|^2 = \sum_i J_i$ with disjoint local error sets and fixed linear wiring, and repeated uses of a signal refer to the same compiled coordinate.*
3. *Each block has fixed maps $p_i = P_{\mathrm{in},i} q_i + P_{\mathrm{err},i} e_i$ and $y_i = Y_{\mathrm{in},i} q_i + Y_{\mathrm{err},i} e_i$; if $I - P_{\mathrm{err},i}^T P_{\mathrm{err},i} \succ 0$, completing the square gives a local representation that preserves the response and energy and retains every error direction. Equivalent temporal or low-rank constructions are given below; Appendix D bounds the composed energy.*
4. *If $\max_a w_a^2 / \sum_r \lambda_{a,r} \le \tau < 1$, with $\tau = 0$ allowed when no error is active, the common witness satisfies $\|e(\xi)\|^2 \le \tau \|p + b_p(\xi)\|^2$.*

Proof. At each node, the envelope supplies one witness valid for all its rows; choose it once. Starting from the input decomposition, subtracting the compiled response from Eq.(24) gives the discrepancy recursion, hence $d_a = \widehat{d}_a + b_a$ at every node, and applying the fixed parent maps gives the simultaneous row identity. Compiled rows are linear in input perturbations and errors, so substitution within a block gives fixed maps; an acyclic grouping that exports all boundary variables preserves all cross-block dependencies, and assigning each squared parent row and each negative error norm once gives the energy split. Square completion gives the local realization. Finally, $e_a^2 = w_a^2\epsilon_a^2 \le \frac{w_a^2}{\sum_r \lambda_{a,r}} \sum_r \lambda_{a,r} s_{a,r}^2 \le \tau \sum_r \big(p_{a,r} + (b_p)_{a,r}\big)^2$; summing over the disjoint active errors proves the last claim. □

Grouping the first and last nodes of a chain $u \to v \to w$ creates a two-way dependency with the block containing v. An activation consumed both by shared selectors and by a recurrence must therefore be exported by an upstream block, with its error coordinate assigned once (Appendix C). Singleton blocks in topological order are always causal.

Two elementary envelopes instantiate Eq.(24). For a unary operation on $|d| \le \bar{d}$, $\bar{d} > 0$, choose a slope ℓ and a valid bound $|f(c+d) - f(c) - \ell d| \le L|d|$; for $L > 0$, set $\sigma = L\bar{d}$, $\epsilon_f = (f(c+d) - f(c) - \ell d)/(L\bar{d})$, and $s_{\text{in}} = d/\bar{d}$, so that $d_y = \ell d + \sigma\epsilon_f$ and $\epsilon_f^2 \le s_{\text{in}}^2$. A derivative enclosure $|f' - \ell| \le L$ suffices for smooth f, and absolute continuity with this bound almost everywhere covers continuous clipping. For a product, the residual $\epsilon_{\text{prod}} = d_u d_v/(\bar{d}_u \bar{d}_v)$ with $\sigma = \bar{d}_u \bar{d}_v$ satisfies both parent constraints from Section 4.1 with one witness. A zero radius introduces no normalized error, and vector normalizations keep their shared denominator as explicit nodes.

Stacking the weighted parent responses and scaled residuals gives the fixed quadratic energy $J = \|p\|^2 - \|e\|^2 = \sum_i J_i$. It is evaluated on the compiled linear responses; Appendix E.2 relates it to the original network through the propagated discrepancies.

Compiled rows and coordinate units. For the running instance, the local basis contains the incoming state, external inputs, and assigned scaled residuals. Let r_b be the row mapping this basis to the homogeneous response of parent b. For a constraint using one parent, write $\kappa_{a,b}$ for its normalization and $\lambda_{a,b}$ for its weight; combinations of parents can be represented by auxiliary nodes. Eq.(25) then gives

$$r_a = \sum_b \alpha_{ab} r_b + (\sigma_a/w_a)\mathbf{u}_a^T, \qquad r_{a,b}^p = \sqrt{\lambda_{a,b}}\,\kappa_{a,b} r_b, \tag{26}$$

where $\mathbf{u}_a$ selects the single assigned error coordinate e_a. A product has two parent rows but one shared error column. Input rows retain their coordinate scales; affine maps and repeated uses of a signal combine the same rows. Stored-center and coefficient discrepancies enter δ_a; arithmetic in the compiled rows is enclosed as specified in Appendix E.

In the running model of Appendix C, a temporal cell contains the error coordinates for its unary activations, modal exponentials, carry products, Δu injection, multiplication by shared B, readout with shared C, and final gate. A fully active cell has $4n + 5$ errors and $7n + 7$ parent rows. For normalized width d_{rms}, an RMS block contains at most $2d_{\text{rms}} + 1$ errors and $3d_{\text{rms}} + 1$ parent rows. These sets partition the instantiated coordinates; affine maps create no nonlinear errors.

B.2 Block representation and square completion

For block input q_b and residual coordinates e, write the homogeneous parent and output responses as $p = P_{\text{in}} q_b + P_{\text{err}} e$ and $y = Y_{\text{in}} q_b + Y_{\text{err}} e$; in the temporal case q_b determines the incoming state and external signal sequence, and e stacks the e_t. For $K = I - P_{\text{err}}^T P_{\text{err}} \succ 0$ and $F = P_{\text{err}}^T P_{\text{in}}$, set $\zeta = K^{1/2} e - K^{-1/2} F q_b$, so that

$$\|p\|^2 - \|e\|^2 = q_b^T Q q_b - \|\zeta\|^2, \qquad y = G q_b + L\zeta, \qquad \Sigma = LL^T, \tag{27}$$

with $Q = P_{\text{in}}^T P_{\text{in}} + F^T K^{-1} F$, $G = Y_{\text{in}} + Y_{\text{err}} K^{-1} F$, and $L = Y_{\text{err}} K^{-1/2}$. For fixed q_b, the change from e to ζ is invertible and preserves every residual direction. Expanding the square proves the energy identity; substitution of $e = K^{-1/2}\zeta + K^{-1} F q_b$ gives the output maps. We require

actions of Q, G, G^T, and Σ, with $Q \succeq 0$. An auxiliary output vector ψ lets us recover these four actions from one quadratic maximization:

$$\sup_e \{\|p\|^2 - \|e\|^2 + 2\psi^T y\} = q_b^T Q q_b + 2\psi^T G q_b + \psi^T \Sigma \psi. \tag{28}$$

The temporal and RMS constructions evaluate this same functional.

B.3 Temporal square completion

Finite-horizon IQC and Riccati methods provide related analyses of temporal energy and induced gains (Seiler et al., 2019; Buch et al., 2021). Here the discrete-time construction realizes the selected quadratic constraints and retains all error directions for subsequent composition.

Proposition 6 (Temporal quadratic realization)**.** *For the finite-dimensional state-space response over a finite horizon in Eq.(29), with fixed matrices, boundary maps $s_0 = S_b q_b$ and $q_t = \mathsf{U}_t q_b$, and disjoint assigned error coordinates e_t, suppose every K_t in Eq.(30) is positive definite. Then a local representation with boundary q_b and $\sum_t \dim(e_t)$ transformed errors preserves the response and energy and retains every error direction. Its Q, G, G^T, and Σ actions are evaluated by temporal recurrences in $O(H)$ local work at fixed local dimensions plus the boundary maps, while avoiding the stacked error-response matrix. The construction accepts general A_t, including noncommuting matrices.*

Temporal model. The temporal block has block input q_b, state s_t, parent-row output p_t, and block output y_t. With boundary inputs $q_t = \mathsf{U}_t q_b$ and assigned error coordinates e_t, its fixed linear responses are

$$\begin{aligned} s_{t+1} &= \mathsf{A}_t s_t + \mathsf{B}_t q_t + \mathsf{E}_t e_t, \\ p_t &= \mathsf{C}_t s_t + \mathsf{D}_t q_t + \mathsf{F}_t e_t, \\ y_t &= \mathsf{L}_t s_t + \mathsf{M}_t q_t + \mathsf{N}_t e_t. \end{aligned} \tag{29}$$

Backward square completion starts with terminal tail coefficient $P_H = 0$. Its quadratic state coefficients obey

$$\begin{aligned} K_t &= I - \mathsf{F}_t^T \mathsf{F}_t - \mathsf{E}_t^T P_{t+1} \mathsf{E}_t \succ 0, \\ X_t &= \mathsf{F}_t^T \mathsf{C}_t + \mathsf{E}_t^T P_{t+1} \mathsf{A}_t, \\ P_t &= \mathsf{C}_t^T \mathsf{C}_t + \mathsf{A}_t^T P_{t+1} \mathsf{A}_t + X_t^T K_t^{-1} X_t. \end{aligned} \tag{30}$$

General probe and square completion. The condition $K_t \succ 0$ is checked at every time step. To recover boundary and output actions, fix q_b and introduce output probes ψ_t in the suffix value

$$V_t(s) = \sup_{e_t,\dots,e_{H-1}} \sum_{j=t}^{H-1} \left(\|p_j\|^2 - \|e_j\|^2 + 2\psi_j^T y_j\right) = s^T P_t s + 2\ell_t^T s + c_t^{\mathrm{pr}}. \tag{31}$$

Initialize $\ell_H = 0$ and $c_H^{\mathrm{pr}} = 0$. Alongside Eq.(30), set $d_t = \mathsf{B}_t q_t$, $f_t = \mathsf{D}_t q_t$, and update backward:

$$\begin{aligned} \nu_t &= \mathsf{F}_t^T f_t + \mathsf{E}_t^T (P_{t+1} d_t + \ell_{t+1}) + \mathsf{N}_t^T \psi_t, \\ \ell_t &= \mathsf{C}_t^T f_t + \mathsf{A}_t^T (P_{t+1} d_t + \ell_{t+1}) + \mathsf{L}_t^T \psi_t + X_t^T K_t^{-1} \nu_t, \\ c_t^{\mathrm{pr}} &= c_{t+1}^{\mathrm{pr}} + \|f_t\|^2 + d_t^T P_{t+1} d_t + 2\ell_{t+1}^T d_t + 2\psi_t^T \mathsf{M}_t q_t + \nu_t^T K_t^{-1} \nu_t. \end{aligned} \tag{32}$$

At time t, the terms involving the current error are $-e_t^T K_t e_t + 2e_t^T (X_t s_t + \nu_t)$. With $\psi = 0$, set

$$\zeta_t = K_t^{1/2} e_t - K_t^{-1/2} (X_t s_t + \nu_t). \tag{33}$$

Square completion gives $\|p_t\|^2 - \|e_t\|^2 + V_{t+1}(s_{t+1}) = V_t(s_t) - \|\zeta_t\|^2$, and summation telescopes to $\sum_t (\|p_t\|^2 - \|e_t\|^2) = V_0(S_b q_b) - \|\zeta\|^2$. The ν_t are linear in q_b, and s_t is linear in q_b and earlier errors, so Eq.(33) is jointly linear and block triangular with invertible diagonal blocks $K_t^{1/2}$; substitution into the outputs gives fixed G, L. The quadratic $V_0(S_b q_b) = q_b^T Q q_b$ satisfies $Q \succeq 0$, since $e = 0$ is feasible with nonnegative energy. This proves Proposition 6's realization.

Forward and adjoint actions. Completing this square gives the updates above and the forward recovery

$$e_t^* = K_t^{-1}(X_t s_t + \nu_t), \qquad s_{t+1} = \mathsf{A}_t s_t + d_t + \mathsf{E}_t e_t^*, \qquad y_t^* = \mathsf{L}_t s_t + \mathsf{M}_t q_t + \mathsf{N}_t e_t^*. \tag{34}$$

Write $\psi = \mathrm{stack}(\psi_t)$ and $y^* = \mathrm{stack}(y_t^*)$. Then the local probe identity in Eq.(28) gives $y^* = Gq_b + \Sigma\psi$. Thus a zero probe gives Gq_b, and zero boundary gives $\Sigma\psi$. To evaluate the boundary actions, use the recovered p_t, s_t, e_t^* and run one adjoint pass with $\lambda_H^{\mathrm{adj}} = 0$:

$$\begin{aligned} \lambda_t^{\mathrm{adj}} &= \mathsf{A}_t^T \lambda_{t+1}^{\mathrm{adj}} + \mathsf{C}_t^T p_t + \mathsf{L}_t^T \psi_t, \\ k_t &= \mathsf{B}_t^T \lambda_{t+1}^{\mathrm{adj}} + \mathsf{D}_t^T p_t + \mathsf{M}_t^T \psi_t, \\ S_b^T \lambda_0^{\mathrm{adj}} + \sum_{t=0}^{H-1} \mathsf{U}_t^T k_t &= Qq_b + G^T\psi. \end{aligned} \tag{35}$$

These expressions are one half of the gradients of the maximized functional; stationarity removes derivatives of the maximizing errors. Setting $\psi = 0$ or $q_b = 0$ yields Qq_b or $G^T\psi$, respectively. Shared input maps therefore pull back and sum channel contributions in the same boundary coordinates.

B.4 RMS REALIZATION

B.4.1 LOW-DIMENSIONAL ERROR COUPLING

Proposition 7 (A low-rank coupling realization)**.** *Let $e \in \mathbb{R}^{m_e}$, let $V_{\mathrm{sub}} \in \mathbb{R}^{m_e \times k}$ have orthonormal columns, and let $J = q^T Q_b q + \|Aq + BV_{\mathrm{sub}}^T e\|^2 - \|e\|^2$ and $y = Cq + De$ with $Q_b \succeq 0$. If $K = I_k - B^T B \succ 0$ and $F = B^T A$, then*

$$\begin{gathered} \zeta_c = K^{1/2} V_{\mathrm{sub}}^T e - K^{-1/2} F q, \qquad \zeta_\perp = V_\perp^T e, \\ Q = Q_b + A^T A + F^T K^{-1} F, \quad G = C + DV_{\mathrm{sub}} K^{-1} F, \; \Sigma = D\big[I + V_{\mathrm{sub}}(K^{-1} - I_k)V_{\mathrm{sub}}^T\big] D^T \end{gathered} \tag{36}$$

preserve the local response and energy and retain every error direction, for any orthonormal complement $V_\perp$.

Proof. Write $e = V_{\mathrm{sub}}\alpha + V_\perp \beta$, so $\|e\|^2 = \|\alpha\|^2 + \|\beta\|^2$. Square completion in α retains $-\|\beta\|^2$ and gives the invertible change $e = V_{\mathrm{sub}} K^{-1} F q + V_{\mathrm{sub}} K^{-1/2} \zeta_c + V_\perp \zeta_\perp$. Substitution gives G and the output factor $L = [DV_{\mathrm{sub}} K^{-1/2} \;\; DV_\perp]$, and $V_\perp V_\perp^T = I - V_{\mathrm{sub}} V_{\mathrm{sub}}^T$ gives Σ without forming $V_\perp$. □

The inversion has dimension k, while all m_e error coordinates remain and complementary errors may reach the output through $DV_\perp$.

B.4.2 RMS COEFFICIENTS AND COUPLED DIRECTIONS

For one token, the RMS operations are $s_i = x_i^2$, $v = d_{\mathrm{rms}}^{-1} \sum_i s_i + \epsilon_{\mathrm{rms}}$, $\varrho_{\mathrm{rms}} = v^{-1/2}$, $z_i = x_i \varrho_{\mathrm{rms}}$, and $\nu_i = w_i^{\mathrm{rms}} z_i$. The error coordinates e_1, e_0, and e_2 correspond respectively to the squares, inverse square root, and final products. Write $\chi = \widehat{v}$ and $\varrho = \widehat{\varrho}$ for the homogeneous responses of the averaged square and its inverse square root. Their coupling is

$$\chi = a^T q_b + g^T e_1, \qquad \varrho = \alpha\chi + \sigma e_0, \qquad y = L_d q_b + b\varrho + D_{\mathrm{prod}} e_2, \tag{37}$$

with energy

$$q_b^T Q_d q_b + c_{\mathrm{inv}}^2 \chi^2 + m_{\mathrm{inv}}^2 \varrho^2 - \|e_1\|^2 - e_0^2 - \|e_2\|^2, \tag{38}$$

where $Q_d, L_d, D_{\mathrm{prod}}$ are diagonal and $Q_d \succeq 0$. The coefficient map in Eq.(39) defines a, g and all remaining coefficients from the local error rows. For $g \neq 0$, Proposition 7 applies with V_{sub} selecting the direction $g/\|g\|$ in e_1 and the coordinate e_0. For $g = 0$, e_1 does not enter χ; only the scalar e_0 coupling remains.

The coefficients can be recovered directly from the local nonlinear rows. Let $u_i > 0$ and $v_i^{\mathrm{out}} > 0$ be the input and output coordinate units, so the homogeneous input is $\widehat{x}_i = u_i (q_b)_i$ and

$y_i = w_i^{\mathrm{rms}} \widehat{z}_i / v_i^{\mathrm{out}}$. Compile these operations by Eq.(26). Here $\widehat{\mu}$ is the fixed averaging coefficient representing $1/d_{\mathrm{rms}}$; ℓ_i, α are the square and inverse-square-root envelope slopes, and $\beta_i^{(x)}, \beta_i^{(\varrho)}$ are the product's fixed parent slopes. Let $t_i^s, t^{\mathrm{inv}}, t_i^p$ denote the residual coefficients of $(e_1)_i, e_0, (e_2)_i$, respectively, using σ_a/w_a for each active node in Eq.(26). Define the weighted parent coefficients from $\sqrt{\lambda_{a,b}}\kappa_{a,b}$: k_i^s for the square's input row, k^{inv} for the inverse-square-root row, and $k_i^{p,x}, k_i^{p,\varrho}$ for the product's two parent rows. Substitution gives

$$\begin{aligned} a_i &= \widehat{\mu}\ell_i u_i, & g_i &= \widehat{\mu} t_i^s, & \sigma &= t^{\mathrm{inv}}, \\ b_i &= w_i^{\mathrm{rms}} \beta_i^{(\varrho)} / v_i^{\mathrm{out}}, & (L_d)_{ii} &= w_i^{\mathrm{rms}} \beta_i^{(x)} u_i / v_i^{\mathrm{out}}, & (D_{\mathrm{prod}})_{ii} &= w_i^{\mathrm{rms}} t_i^p / v_i^{\mathrm{out}}, \\ (Q_d)_{ii} &= u_i^2\big((k_i^s)^2 + (k_i^{p,x})^2\big), & c_{\mathrm{inv}} &= k^{\mathrm{inv}}, & m_{\mathrm{inv}}^2 &= \sum_i (k_i^{p,\varrho})^2. \end{aligned} \tag{39}$$

Take $m_{\mathrm{inv}} \geq 0$. The weighted parent rows are $k_i^s u_i (q_b)_i$, $c_{\mathrm{inv}}\chi$, $k_i^{p,x} u_i (q_b)_i$, and $k_i^{p,\varrho}\varrho$; summing their squares gives the positive terms of Eq.(38).

At inactive sites, set the corresponding residual and weighted parent coefficients to zero. The full-index notation above uses zeros at those sites; the actual error vectors retain only instantiated coordinates, with g restricted to active square coordinates and D_{prod} restricted to active product columns. If the inverse-square-root error is inactive, its e_0 coordinate is absent. The stabilizer affects the inverse-square-root domain and slope but has no homogeneous response.

For $\omega = \|g\| > 0$, decompose e_1 into its component along g and its orthogonal complement. Only the former enters χ. Together with e_0 it has core

$$K_{\mathrm{rms}} = I_2 - D_{\mathrm{rms}}^T D_{\mathrm{rms}} \succ 0, \qquad D_{\mathrm{rms}} = \begin{pmatrix} c_{\mathrm{inv}}\omega & 0 \\ m_{\mathrm{inv}}\alpha\omega & m_{\mathrm{inv}}\sigma \end{pmatrix}. \tag{40}$$

Square completion produces

$$Q = Q_d + q_c a a^T, \qquad G = L_d + g_c b a^T, \qquad \Sigma = D_{\mathrm{prod}} D_{\mathrm{prod}}^T + s_c b b^T. \tag{41}$$

Set $c_{\mathrm{rms}} = (c_{\mathrm{inv}}, m_{\mathrm{inv}}\alpha)^T$, $f = D_{\mathrm{rms}}^T c_{\mathrm{rms}}$, and $\ell = (\alpha\omega, \sigma)^T$. Then

$$q_c = c_{\mathrm{rms}}^T c_{\mathrm{rms}} + f^T K_{\mathrm{rms}}^{-1} f, \quad g_c = \alpha + \ell^T K_{\mathrm{rms}}^{-1} f, \quad s_c = \ell^T K_{\mathrm{rms}}^{-1} \ell. \tag{42}$$

The $g = 0$ case has zero coupling to the square-error vector and uses the corresponding uncoupled form.

B.5 Global composition and matrix-free evaluation

Theorem 8 (Representation of a fixed QC system)**.** *Fix the local envelopes, coefficients, weights, error scales, and active residuals. For block i, let q_i collect incoming signals and write its weighted parent responses as $p_i = P_{\mathrm{in},i} q_i + P_{\mathrm{err},i} e_i$. Suppose $I - P_{\mathrm{err},i}^T P_{\mathrm{err},i} \succ 0$ for every block. Suppose also that the block graph is acyclic, each error and its constraint rows belong to one block, and every signal used by another block or the output is retained as the same variable. The temporal and RMS square-completion constructions then give an invertible change from e to ζ with*

$$z = \mathcal{A}\xi + \mathcal{N}\zeta, \qquad J = \xi^T K_{\mathrm{src}} \xi + z^T Q_0 z - \|\zeta\|^2. \tag{43}$$

Here z collects the block responses, including input coordinates needed by the output, and $K_{\mathrm{src}}, Q_0 \succeq 0$ collect the quadratic costs of the incoming signals. The transformation retains every error direction and preserves both the response and J of the fixed QC system. Forward and adjoint recursions evaluate the required linear maps and their Gram matrices without forming a dense global response matrix.

Proof. Causally ordered blocks obey

$$q_i = S_i \xi + \sum_{j<i} W_{ij}^{\mathrm{wire}} y_j, \tag{44}$$

where every use of an exported signal refers to the same coordinate. Shared selectors and residual signals use one common coordinate across blocks. Stack the local maps as $G = \mathrm{diag}_i\, G_i$, $L =$

$\operatorname{diag}_i L_i$, $S = \operatorname{col}_i S_i$, and $W^{\mathrm{wire}} = [W_{ij}^{\mathrm{wire}}]$, which is strictly block lower triangular. Given ξ and earlier errors, q_i is determined before e_i, so the local changes $\zeta_i = T_i^e e_i + T_i^q q_i$ form a block-triangular map with invertible diagonal blocks T_i^e. Recovering $e_i = (T_i^e)^{-1}(\zeta_i - T_i^q q_i)$ and then y_i in topological order proves bijectivity and $\dim \zeta = \dim e$. Stacking the outputs gives $y = GS\xi + GW^{\mathrm{wire}} y + L\zeta$; since GW^{wire} is nilpotent,

$$y = \mathcal{Z}GS\xi + \mathcal{Z}L\zeta, \qquad \mathcal{Z} = (I - GW^{\mathrm{wire}})^{-1}. \tag{45}$$

With $Q_{\mathrm{loc}} = \operatorname{diag}_i Q_i$ and $q = S\xi + W^{\mathrm{wire}} y$, summing the local identities under energy additivity gives $J = q^T Q_{\mathrm{loc}} q - \|\zeta\|^2$, retaining all cross terms of the wiring. Collect costs whose boundaries depend only on ξ into K_{src}. Express the remaining costs in the collected responses z. When a boundary depends on both ξ and z, append a deterministic copy of ξ to z. This gives Eq.(43); both cost matrices are PSD because they are pullbacks of PSD local costs. With $\Sigma = \operatorname{diag}_i \Sigma_i$, the error-response Gram is $\Theta_y = \mathcal{Z}\Sigma\mathcal{Z}^T$. For an output right-hand side f, the recurrences

$$v_i = f_i + \sum_{k>i} (W_{ki}^{\mathrm{wire}})^T G_k^T v_k \text{ (reverse)}, \qquad u_i = \Sigma_i v_i + G_i \sum_{j<i} W_{ij}^{\mathrm{wire}} u_j \text{ (forward)} \tag{46}$$

solve $(I - (W^{\mathrm{wire}})^T G^T)v = f$ and $(I - GW^{\mathrm{wire}})u = \Sigma v$, so $u = \Theta_y f$. The source response is a zero-error forward execution and its adjoint is $S^T G^T v$; no factor L_i or dense response matrix is needed.

□

The global energy condition adds a restriction beyond local positivity. Let block 1 have $p_1 = 0$ and $y_1 = e_1$, and block 2 have $q_2 = y_1$, $p_2 = 2q_2$, and $y_2 = e_2$. Both local error cores are the identity, but the total energy is $3e_1^2 - e_2^2$, whose supremum over errors is unbounded. Shared signals also require their cross terms: if two exported signals equal the same scalar error, their joint Gram is $\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$.

B.5.1 Layer wiring and operator actions

For the running stack, let ν_i denote the normalized signal at layer i. The block boundaries are

$$\nu_i = U_i x_{i-1} + L_i^{\mathrm{rms}} \zeta_i^{\mathrm{rms}}, \qquad x_i = S_i^{\mathrm{skip}} x_{i-1} + G_i \nu_i + L_i^p \zeta_i^p, \qquad x_{-1} = S_0 \xi. \tag{47}$$

Let $R_i^{\mathrm{rms}} \succeq 0$ and $C_i \succeq 0$ be the input costs of RMS and the temporal block. For the interleaved state $z = (x_0, \nu_0, \ldots, x_{L-1}, \nu_{L-1})$, the first RMS input belongs to the source, while the other local input costs belong to collected states:

$$K_{\mathrm{src}} = S_0^T R_0^{\mathrm{rms}} S_0, \qquad z^T Q_0 z = \sum_{i=0}^{L-1} \nu_i^T C_i \nu_i + \sum_{i=1}^{L-1} x_{i-1}^T R_i^{\mathrm{rms}} x_{i-1}. \tag{48}$$

Equivalently, $Q_0 = \operatorname{diag}(R_1^{\mathrm{rms}}, C_0, \ldots, R_{L-1}^{\mathrm{rms}}, C_{L-2}, 0, C_{L-1})$; for $L = 1$ it is $\operatorname{diag}(0, C_0)$. The source-only and error-only executions of Eq.(47) define $\mathcal{A}$ and $\mathcal{N}$.

Residual skip maps act through the state recurrence, retaining terms such as $2x^T (S^{\mathrm{skip}})^T W G \bar{x}$ under a downstream cost W.

For a general local boundary $q_b = T_b \xi + V_b z$, its cost includes $2\xi^T T_b^T Q_b V_b z$. Appending a deterministic copy of ξ to the state represents this cost by the PSD pullback $[T_b\ V_b]^T Q_b [T_b\ V_b]$. The appended coordinates have zero error response. The same augmentation handles a direct source term in the terminal output.

Forward and adjoint actions. Apply Eq.(46) to Eq.(47) in causal block order $(\nu_0, x_0, \nu_1, x_1, \ldots)$, then permute to the stored order of z. The RMS block maps x_{i-1} to ν_i through U_i. The temporal block maps (x_{i-1}, ν_i) to x_i through $[S_i^{\mathrm{skip}}, G_i]$, with input cost $\operatorname{diag}(0, C_i)$. These choices include the skip connection in the general wiring.

For $\Theta = \mathcal{N}\mathcal{N}^T$, start the forward pass at zero source response and set adjoint variables beyond the terminal layer to zero. Denote the reverse-pass variables by (ψ_i^x, ψ_i^ν). The quadratic action needs only that pass: $f^T \Theta f = \sum_i [(\psi_i^x)^T \Sigma_i^p \psi_i^x + (\psi_i^\nu)^T \Sigma_i^{\mathrm{rms}} \psi_i^\nu]$. The source action $\mathcal{A}\xi$ is the zero-error forward recurrence.

C MAMBA-STYLE INSTANTIATIONS

For each model, we specify how to parameterize accumulated state transitions (Theorem 1) and enclose the remaining network operations (Proposition 5). We then group the QC rows in causal order (Theorem 8).

C.1 MAMBA-1 RUNNING INSTANCE

We first show the objects on a Mamba-1 block. Its residual-stream input is $x_t \in \mathbb{R}^d$. The block applies root mean square normalization (RMS normalization), fixed projections, a causal convolution, a selective scan, a gate, and an output projection. With learned diagonal weight ω and positive stabilizer ϵ_{rms},

$$\mathrm{RMS}(x) = \omega \odot x\big(d^{-1}\|x\|^2 + \epsilon_{\mathrm{rms}}\big)^{-1/2}.$$

Let $v_t \in \mathbb{R}^m$ be the convolution output and $u_t = \mathrm{SiLU}(v_t)$. For channel j, the state $h_{t,j} \in \mathbb{R}^n$ and readout satisfy

$$\begin{aligned} h_{t,j} &= \mathrm{diag}(e^{a_j \Delta_{t,j}})h_{t-1,j} + \Delta_{t,j} u_{t,j} B_t, \\ o_{t,j} &= C_t^T h_{t,j} + D_j u_{t,j}, \qquad x'_t = x_t + W_{\mathrm{out}}(g_t \odot o_t), \end{aligned} \tag{49}$$

where a_j is a fixed negative vector, $B_t, C_t \in \mathbb{R}^n$ are input-dependent selectors shared across channels, and $g_t = \mathrm{SiLU}(W_z\,\mathrm{RMS}(x_t))$. In the primary pre-SiLU selection variant,

$$(r_t, B_t, C_t) = W_{\mathrm{sel}} v_t, \qquad \Delta_t = \mathrm{softplus}(W_\Delta r_t + b_\Delta). \tag{50}$$

The post-SiLU variant replaces v_t by u_t in the same equation, so the activation then feeds both selection and injection.

C.2 QC GROUPING FOR PRE- AND POST-SILU SELECTION

With pre-SiLU selection in Appendix C.1, the activation enters only the channel injections. Its activation errors can therefore be assigned to the channel temporal blocks. Its state-transition map is the diagonal exponential of Appendix A.3.1.

With post-SiLU selection, the activation feeds both the shared selectors and every channel injection. The history identity is unchanged; the operand polynomial enclosures for B_t, C_t, and Δ_t are formed from the same SiLU polynomial enclosure. To make the shared activation available before the channel recurrences, the QC construction assigns its SiLU errors to an upstream QC block containing the convolution and SiLU operation, which exports the activation u using a fixed affine response and one assigned error coordinate per active scalar SiLU operation. Its error core is the identity. The selector maps, step nonlinearities, and channel temporal blocks all import the exported u, and the residual block keeps $(x, \bar{x}, u)$ together at its boundary.

C.3 MAMBA-2

A Mamba-2 block projects its normalized input to (z, x, B, C, r), applies a causal convolution and SiLU to x, B, and C, and computes $\Delta_{t,h} = \mathrm{softplus}(r_{t,h} + b_h)$ in a separate branch. Each head has a scalar rate $a_h < 0$ and a matrix state updated by

$$h_{t,h} = e^{a_h \Delta_{t,h}} h_{t-1,h} + \Delta_{t,h} B_t x_{t,h}^T, \qquad y_{t,h} = C_t^T h_{t,h} + D_h x_{t,h},$$

with B_t, C_t shared within a group of heads. The output is gated by $\mathrm{SiLU}(z_t)$ before an RMS normalization over all inner coordinates and the output projection. Its state-transition-product parameterization is the scalar-identity instance of Appendix A.3, with payload $U_t = \Delta_{t,h} x_{t,h}^T$.

For the QC route, the convolution, SiLU, and step errors are assigned to one QC block; it exports x, B, C, and Δ to the head blocks. The decay, injection, and readout errors are assigned to head temporal blocks, which import the exported selectors and reuse the same error coordinates for grouped B_t, C_t across heads. The RMS after gating couples all heads and forms a separate block, an instance of Proposition 7; the gate forms the block between them.

C.4 MAMBA-3-STYLE RECURRENCE

The Mamba-3-style block follows the recurrence of Lahoti et al. (2026) with rank one and without output normalization. It projects its normalized input to $(z, x, B, C, r, \hat{a}, \hat{\lambda}, \hat{\varphi})$ without a convolution. It normalizes B and C by RMS over state coordinates and adds learned biases, sets $a_t = -\max(\mathrm{ht}(\hat{a}_t), a_{\min})$ with $a_{\min} > 0$ and

$$\mathrm{ht}(u) = \begin{cases} 1 + u, & u \geq 0, \\ (1-u)^{-1}, & u < 0, \end{cases} \qquad u \in \mathbb{R}. \tag{51}$$

Thus $\mathrm{ht}(u) > 0$ and the floor keeps the decay magnitude positive. The function ht is continuously differentiable but not twice differentiable at zero. Enclosures cover the full parent range; ranges crossing zero or the decay-floor threshold use the threshold-crossing rules of Appendix A. With $\Delta_t = \mathrm{softplus}(r_t + b)$ and $\lambda_t = \mathrm{sigmoid}(\hat{\lambda}_t)$, it then accumulates a phase $\vartheta_t = \vartheta_{t-1} + \pi\Delta_t \tanh(\hat{\varphi}_t)$. With rotations $\mathrm{Rot}(\vartheta_t)$ on fixed coordinate planes, the implemented recurrence uses the rotated selectors $\widetilde{B}_t = \mathrm{Rot}(\vartheta_t)(\mathrm{RMS}(B_t) + b_B)$ and $\widetilde{C}_t = \mathrm{Rot}(\vartheta_t)(\mathrm{RMS}(C_t) + b_C)$ and a stored state

$$S_t = e^{\Delta_t a_t} S_{t-1} + (1-\lambda_t)\Delta_t e^{\Delta_t a_t} \widetilde{B}_{t-1} x_{t-1}^T + \lambda_t \Delta_t \widetilde{B}_t x_t^T, \qquad y_t = \widetilde{C}_t^T S_t + D x_t,$$

followed by a SiLU gate and the output projection; rank $r_{\mathrm{mimo}} > 1$ replaces the payloads and readouts by rank-r_{mimo} blocks.

The state $h_t = \mathrm{Rot}(-\vartheta_t) S_t$ satisfies the joint recurrence of Corollary 1 with unrotated selectors $\mathrm{RMS}(B_t) + b_B$ and $\mathrm{RMS}(C_t) + b_C$, transition $F_t = e^{\Delta_t a_t}\,\mathrm{Rot}(-(\vartheta_t - \vartheta_{t-1}))$, and phase rate $\omega_t = -\pi \tanh(\hat{\varphi}_t)$ on each rotated plane; unrotated coordinates share the decay. These are equivalent parameterizations: one rotates the selectors and the other uses rotating state coordinates. The prefix phase and delayed injection belong to the enclosed prefix context of the affected suffix. The parameterization has one shared decay parameter and one phase parameter for each rotated plane.

For the QC route, the selector RMS normalizations, heavy-tail floor, and softplus, sigmoid, tanh, sine, and cosine envelope errors are assigned to one QC block containing those selector operations. The decay, two-path injection, and readout errors are assigned to time blocks; their separator carries both the state and the previous injection, so the delayed path is represented once. A normalization of the output before the gate, when present, is another instance of Proposition 7.

D CERTIFICATION CHECKS AND SOURCE BOUNDS

This appendix gives the sufficient bounds used after realization of the fixed QC system. The state z collects homogeneous responses; original-node centers and additive discrepancies are handled separately in Appendix E.1. The full-space check controls the parent-energy operator, the terminal check controls the selected output direction, and the source check bounds the remaining quadratic form in the input perturbation variables ξ. Residual-based matrix bounds are established tools in verified numerical analysis (Rump, 2011). Theorem 8 supplies $z = \mathcal{A}\xi + \mathcal{N}\zeta$ and its quadratic energy J. Let $\ell_\phi = a_{\mathrm{out}}^T z$ be the homogeneous output response, $\Theta = \mathcal{N}\mathcal{N}^T$, $b_{\mathrm{out}} = \mathcal{N}^T a_{\mathrm{out}}$, and $E_0 = I - \mathcal{N}^T Q_0 \mathcal{N}$. Writing $Q_{\mathrm{out}} = Q_0 + a_{\mathrm{out}} a_{\mathrm{out}}^T$, the full-space and terminal checks establish

$$\mathcal{N}^T Q_0 \mathcal{N} \preceq \beta I, \qquad 0 < \beta < 1, \tag{52}$$

$$b_{\mathrm{out}}^T E_0^{-1} b_{\mathrm{out}} \leq \bar{\chi}_{\mathrm{term}} < 1, \qquad E = I - \mathcal{N}^T Q_{\mathrm{out}} \mathcal{N} \succeq (1 - \bar{\chi}_{\mathrm{term}})(1-\beta) I. \tag{53}$$

The first inequality controls all error directions; the terminal bound keeps their quadratic coefficient negative after the squared output is added. Completing the square in free ζ gives $\ell_\phi^2 + J \leq \xi^T M \xi$, where

$$M = K_{\mathrm{src}} + \mathcal{A}^T Q_{\mathrm{out}} \mathcal{A} + F_{\mathrm{src}}^T E^{-1} F_{\mathrm{src}}, \qquad F_{\mathrm{src}} = \mathcal{N}^T Q_{\mathrm{out}} \mathcal{A}. \tag{54}$$

The matrices Q_0, Q_{out} act on collected responses, while K_{src}, M act on the input perturbation variables. A verified majorant $M \preceq \mathrm{diag}(\rho^{\mathrm{src}})$ with $\rho^{\mathrm{src}} \geq 0$ gives $U_{\mathrm{src}} = \sum_i \rho_i^{\mathrm{src}}$ and $\ell_\phi^2 + J \leq U_{\mathrm{src}}$ on $[-1, 1]^s$ for all free errors. The following checks construct these bounds from operator actions and enclosed residuals of numerical solves.

D.1 Eliminating a block's errors and bounding the orthogonal complement

At zero source input, $z = \mathcal{N}\zeta$, so Eq.(52) is equivalent to $z^T(Q_0/\beta)z - \|\zeta\|^2 \le 0$ for every ζ. To bound this energy, eliminate blocks in reverse order of Eq.(47): each layer's temporal errors, then its upstream RMS errors. Each elimination passes a quadratic cost to the remaining boundary, where it is added to the local costs. The two bounds below implement these steps while covering all error directions.

Let e_{priv} denote the error vector for the block being eliminated, let y be the boundary response before adding this error, let L map the error to the same boundary coordinates, and let $W \succeq 0$ be the accumulated boundary quadratic weight. The direct positivity premise is $I - L^T W L \succ 0$ (equivalently, $I - W^{1/2} L L^T W^{1/2} \succ 0$). Completing the square then gives

$$\sup_{e_{\text{priv}}} \left\{ (y + L e_{\text{priv}})^T W (y + L e_{\text{priv}}) - \|e_{\text{priv}}\|^2 \right\} = y^T W^{1/2} (I - W^{1/2} L L^T W^{1/2})^{-1} W^{1/2} y. \quad (55)$$

Write $\Sigma = LL^T$ and $R_{\text{fb}} = W^{1/2} \Sigma W^{1/2}$. The bound on R_{fb} must cover every direction, including those outside the span of the retained numerical eigenvectors. For any thin matrix F, set

$$\begin{gathered} Z = W^{1/2} F, \qquad R = F^T W F, \qquad T_F = (WF)^T (\Sigma W) F, \\ \|R_{\text{fb}} - Z Z^T\|_F^2 = \operatorname{tr}\big((\Sigma W)^2\big) - 2 \operatorname{tr} T_F + \operatorname{tr}(R^2). \end{gathered} \quad (56)$$

The three traces are enclosed independently — the two squared traces upward and $\operatorname{tr} T_F$ downward, each with its own directed-rounding bound. Let δ_{priv} be that enclosed residual bound, which also bounds $\|R_{\text{fb}} - Z Z^T\|$, and suppose $\delta_{\text{priv}} < 1$. Then $R_{\text{fb}} \preceq Z Z^T + \delta_{\text{priv}} I$, and with $a = 1 - \delta_{\text{priv}}$ and $aI - R \succ 0$,

$$W^{1/2} (I - R_{\text{fb}})^{-1} W^{1/2} \preceq \big(W + WF(aI - R)^{-1} F^T W \big)/a. \quad (57)$$

The checker factorizes the small core $aI - R$ and evaluates its boundary action. With no retained direction, F is empty, $\delta_{\text{priv}} = \|R_{\text{fb}}\|_F$ from the full trace alone, and Eq.(57) reduces to the scalar cap $W/(1 - \delta_{\text{priv}})$.

At an RMS boundary, let V_{sub} have full column rank, with Gram $G_{\text{sub}} = V_{\text{sub}}^T V_{\text{sub}} \succ 0$ and exact orthogonal projector $P = V_{\text{sub}} G_{\text{sub}}^{-1} V_{\text{sub}}^T$, and suppose

$$\begin{pmatrix} A_{\text{rms}} & C_\times \\ C_\times^T & R_{\text{fb}} \end{pmatrix} \succeq 0, \qquad \|(I - P) R_{\text{fb}} (I - P)\| \le \delta < 1. \quad (58)$$

Let $K_{\text{cmp}} = (1 - \delta) G_{\text{sub}} - V_{\text{sub}}^T R_{\text{fb}} V_{\text{sub}} \succ 0$. Then

$$\sup_{e_{\text{rms}}} \left\{ x^T A_{\text{rms}} x + 2 x^T C_\times e_{\text{rms}} + e_{\text{rms}}^T R_{\text{fb}} e_{\text{rms}} - \|e_{\text{rms}}\|^2 \right\} \le \frac{x^T (A_{\text{rms}} + C_\times V_{\text{sub}} K_{\text{cmp}}^{-1} V_{\text{sub}}^T C_\times^T) x}{1 - \delta}. \quad (59)$$

Here A_{rms} denotes the local boundary-cost block.

To see why the complement remains covered, factor the joint PSD block as $[B\ D]^T [B\ D]$ and decompose $e_{\text{rms}} = V_{\text{sub}} \alpha + w$, $w \perp \text{range}(V_{\text{sub}})$. The squared norm of D restricted to the complementary space is at most δ. For fixed α, eliminating w therefore bounds the energy by $\|Bx + D V_{\text{sub}} \alpha\|^2/(1 - \delta) - \alpha^T G_{\text{sub}} \alpha$. Completing the remaining square in α gives Eq.(59). For a numerically computed V_{sub}, the checker encloses $\eta \ge \|I - G_{\text{sub}}\|$, bounds the projector defect by $\|P - V_{\text{sub}} V_{\text{sub}}^T\| \le \eta$, and takes for δ the enclosed norm of the computed residual $(I - V_{\text{sub}} V_{\text{sub}}^T) R_{\text{fb}} (I - V_{\text{sub}} V_{\text{sub}}^T)$ — the smaller of its Frobenius and spectral bounds, plus the interval radius of its entries — inflated by $\eta(2 + \eta)\|R_{\text{fb}}\|$ for that defect. Positivity of K_{cmp} is then established by a rounding-aware Cholesky factorization of the enclosed small core. When all elimination checks succeed, the remaining cost is zero at the zero source boundary, establishing Eq.(52).

D.2 Terminal bounds

Write $L_0 = \mathcal{N}^T Q_0 \mathcal{N} \preceq \beta I$, so $E_0 = I - L_0 \succeq (1 - \beta) I$. If $\chi_{\text{term}} = b_{\text{out}}^T E_0^{-1} b_{\text{out}} \le \bar{\chi}_{\text{term}} < 1$, then

$$E = E_0^{1/2} \left(I - E_0^{-1/2} b_{\text{out}} b_{\text{out}}^T E_0^{-1/2} \right) E_0^{1/2} \succeq (1 - \bar{\chi}_{\text{term}})(1 - \beta) I.$$

This establishes the terminal gap used in the source bound.

To bound χ_{term} using residuals of the local equations, choose a fixed trial state $v \in \mathrm{range}(\mathcal{N})$ and set $u_0 = \mathcal{N}^+ v$. Completing the square gives

$$\chi_{\mathrm{term}} = 2a_{\mathrm{out}}^T v + v^T Q_0 v - \|u_0\|^2 + (b_{\mathrm{out}} - E_0 u_0)^T E_0^{-1}(b_{\mathrm{out}} - E_0 u_0). \tag{60}$$

For any output right-hand side f, split the residual as

$$b_{\mathrm{out}} - E_0 u_0 = \mathcal{N}^T(a_{\mathrm{out}} + Q_0 v - f) + (\mathcal{N}^T f - u_0).$$

If local forward/adjoint equation defects give checked nonnegative squared-norm bounds $\varepsilon_{\mathrm{adj}}, \varepsilon_{\mathrm{loc}}$ for these components, and U_{term} bounds the first three terms in Eq.(60), then

$$\chi_{\mathrm{term}} \le U_{\mathrm{term}} + \frac{(\sqrt{\varepsilon_{\mathrm{adj}}} + \sqrt{\varepsilon_{\mathrm{loc}}})^2}{1-\beta}. \tag{61}$$

The reported runs use this bound. The stored trial, reconstructed right-hand side, and residual bounds are defined below. Throughout, i indexes layers of the wiring in Eq.(47), whose maps are S_i^{skip}, G_i, and U_i, and Σ_i^p and Σ_i^{rms} denote the local output Grams of the temporal and RMS blocks of layer i.

Trial state. The trial is one fixed stored column v in the collected state coordinates, split per layer into its temporal part v_i^x and its RMS part v_i^ν. Running the wiring backwards on v gives the local equation residuals it implies,

$$r_i^x = v_i^x - G_i v_i^\nu - S_i^{\mathrm{skip}} v_{i-1}^x, \qquad r_i^\nu = v_i^\nu - U_i v_{i-1}^x, \tag{62}$$

with terms before the first layer set to zero.

Adjoint variables. Let $f_0 = a_{\mathrm{out}} + Q_0 v$ be the target right-hand side, split into $(f_{0,i}^x, f_{0,i}^\nu)$. Applying the adjoint pass of Eq.(46), with the layer mapping in Appendix B.5.1, to f_0 returns numerical adjoint variables (ψ_i^x, ψ_i^ν). Interpreted as fixed stored vectors, they solve the same recursion for the reconstructed right-hand side $\widehat{f}$ with components

$$\widehat{f}_i^x = \psi_i^x - (S_{i+1}^{\mathrm{skip}})^T \psi_{i+1}^x - U_{i+1}^T \psi_{i+1}^\nu, \qquad \widehat{f}_i^\nu = \psi_i^\nu - G_i^T \psi_i^x, \tag{63}$$

so they represent $\mathcal{N}^T \widehat{f}$ exactly, without forming $\mathcal{N}$. The residual split above is used with $\widehat{f}$ in place of f.

Trial value. U_{term} encloses $2a_{\mathrm{out}}^T v + v^T Q_0 v$ minus the adjoint lower bound $2\sum_i((\psi_i^x)^T r_i^x + (\psi_i^\nu)^T r_i^\nu) - \sum_i((\psi_i^x)^T \Sigma_i^p \psi_i^x + (\psi_i^\nu)^T \Sigma_i^{\mathrm{rms}} \psi_i^\nu)$ on $\|u_0\|^2$.

Adjoint defect $\varepsilon_{\mathrm{adj}}$. The first residual component is then $\mathcal{N}^T(f_0 - \widehat{f})$, and $\varepsilon_{\mathrm{adj}}$ encloses $(f_0 - \widehat{f})^T \Theta (f_0 - \widehat{f})$, evaluated by the backward-only actions of Appendix B.5.1. Adjoint variables that solved the adjoint equations exactly would give $\widehat{f} = f_0$ and $\varepsilon_{\mathrm{adj}} = 0$; finite precision is what makes this term positive.

Local-equation defects $\varepsilon_{\mathrm{loc}}$. For each block, apply its Gram matrix to the adjoint variables and subtract the trial's local equation residual. To bound the resulting contribution to $\varepsilon_{\mathrm{loc}}$, let $\Sigma_i^p \succeq \mathrm{blockdiag}_t(d_{i,t} I)$ and $\Sigma_i^{\mathrm{rms}} \succeq \mathrm{blockdiag}_t(s_{i,t} I)$ be checked floors over the token blocks. For positive floors,

$$\varepsilon_{\mathrm{loc}} = \sum_{i,t} \frac{\| [\Sigma_i^p \psi_i^x - r_i^x]_t \|^2}{d_{i,t}} + \sum_{i,t} \frac{\| [\Sigma_i^{\mathrm{rms}} \psi_i^\nu - r_i^\nu]_t \|^2}{s_{i,t}}, \tag{64}$$

with upward rounding in every term. To obtain the temporal floors $d_{i,t} > 0$, use three disjoint output-only error families: the final gated product, the gate, and the readout product. Include their slopes and map them through W_{out} and the output units. Enclose their Gram matrix and apply a rounding-aware Cholesky factorization. A positive temporal core has $(I - P_{\mathrm{err}}^T P_{\mathrm{err}})^{-1} = K^{-1} \succeq I$ (Appendix B.2), so Σ_i^p dominates this Gram matrix. The RMS floors $s_{i,t} \ge 0$ are the checked lower endpoints, over the block, of the diagonal $D_{\mathrm{prod}} D_{\mathrm{prod}}^T$ of Eq.(41); since $\Sigma^{\mathrm{rms}} = D_{\mathrm{prod}} D_{\mathrm{prod}}^T + s_c b b^T$ with $s_c = \ell^T K_{\mathrm{rms}}^{-1} \ell \ge 0$ under a positive core, that diagonal is a Loewner lower bound.

Range conditions. The floors provide a sufficient range test for this split. Block i contributes $L_i \zeta_i$ to the collected state, so its local equation residual in Eq.(62) is attainable exactly when it lies in

$\mathrm{range}(L_i)$; with the causal wiring in Eq.(44), a v whose layerwise residuals all lie in those ranges is produced by some ζ, that is, $v \in \mathrm{range}(\mathcal{N})$. A positive floor certifies $\Sigma_i = L_i L_i^T$ positive definite on the block, hence that range blockwise. A block with $s_{i,t} = 0$ is admitted only at the zero initial boundary, where $r_0^\nu = v_0^\nu$: its trial block is set to zero, which keeps v feasible, and its quadratic contribution is $[\psi_i^\nu]_t^T [\Sigma_i^{\mathrm{rms}} \psi_i^\nu]_t$. At later layers, the sufficient range test accepts a candidate when the floor is positive and records an unresolved range condition when the floor is zero.

This positive-floor test is stronger than the range condition itself. For example, a block with $\Sigma_i = 0$ and zero local equation residual satisfies the range condition but fails the positive-floor test; the recorded status is an unresolved range condition.

Acceptance. Equation defects, inverse energies, and scalar arithmetic all receive enclosing bounds. The terminal candidate is accepted only when v lies in the checked feasible range, the outward-rounded $\varepsilon_{\mathrm{adj}}, \varepsilon_{\mathrm{loc}}$ are finite, and Eq.(61) gives $\bar{\chi}_{\mathrm{term}} < 1$. It then supplies the gap $E \succeq (1 - \bar{\chi}_{\mathrm{term}})(1 - \beta) I$ required by Theorem 10.

D.3 Full-source completion and residual actions

The source check bounds the same matrix M either by correcting a numerical trial or by using direct spectral estimates; every candidate must pass the diagonal-cap check. Theorem 8 gives, for free transformed errors,

$$\begin{aligned} \ell_\phi^2 + \|p\|^2 - \|e\|^2 &= \xi^T K_{\mathrm{src}} \xi + (\mathcal{A}\xi + \mathcal{N}\zeta)^T Q_{\mathrm{out}} (\mathcal{A}\xi + \mathcal{N}\zeta) - \|\zeta\|^2 \\ &= \xi^T M \xi - (\zeta - E^{-1} F_{\mathrm{src}} \xi)^T E (\zeta - E^{-1} F_{\mathrm{src}} \xi). \end{aligned} \tag{65}$$

The positive gap on E makes this a finite upper bound on the full error space. For any numerical trial Y, define

$$M_Y = K_{\mathrm{src}} + \mathcal{A}^T Q_{\mathrm{out}} \mathcal{A} + F_{\mathrm{src}}^T Y + Y^T F_{\mathrm{src}} - Y^T E Y. \tag{66}$$

Let $\mu > 0$ satisfy $E \succeq \mu I$; the terminal check supplies $\mu = (1 - \bar{\chi}_{\mathrm{term}})(1 - \beta)$. With $\mathcal{E}_Y = F_{\mathrm{src}} - EY$, expanding the completion gives the residual correction

$$M - M_Y = \mathcal{E}_Y^T E^{-1} \mathcal{E}_Y \preceq \mu^{-1} \mathcal{E}_Y^T \mathcal{E}_Y. \tag{67}$$

Thus M is the exact source matrix, M_Y is the matrix produced by a stored trial, and $\mathcal{E}_Y^T E^{-1} \mathcal{E}_Y$ is the nonnegative correction that the checker must enclose. A finite trial is therefore admissible even when its residual is large; the residual correction then enlarges the source cap.

Two full-space candidates used by the source check are defined directly from the same complete source response. Let $\mathcal{A}$ contain all s input perturbation columns, let K_{src} be the source cost, and let $\Theta = \mathcal{N}\mathcal{N}^T$. Under the checked gap $E \succeq \mu I$ with $0 < \mu \leq 1$, the zeroth-order spectral candidate is

$$M_{\mathrm{spec}} = K_{\mathrm{src}} + \mu^{-1} \mathcal{A}^T Q_{\mathrm{out}} \mathcal{A}. \tag{68}$$

The first-moment candidate keeps the deterministic energy and bounds the residual correction by $\mu^{-1} F_{\mathrm{src}}^T F_{\mathrm{src}}$:

$$M_{\mathrm{1mom}} = K_{\mathrm{src}} + \mathcal{A}^T Q_{\mathrm{out}} \mathcal{A} + \mu^{-1} (Q_{\mathrm{out}} \mathcal{A})^T \Theta (Q_{\mathrm{out}} \mathcal{A}). \tag{69}$$

Both satisfy $M \preceq M_{\mathrm{spec}}$ and $M \preceq M_{\mathrm{1mom}}$. For the first bound, put $S = Q_{\mathrm{out}}^{1/2} \mathcal{N}$. The gap implies $I - SS^T \succeq \mu I$, so the source energy after eliminating errors is bounded by $\mu^{-1} \mathcal{A}^T Q_{\mathrm{out}} \mathcal{A}$. The second follows by setting $Y = 0$ in Eq.(67), since $F_{\mathrm{src}}^T F_{\mathrm{src}} = (Q_{\mathrm{out}} \mathcal{A})^T \Theta (Q_{\mathrm{out}} \mathcal{A})$. Their right-hand sides are evaluated as outward-rounded matrix enclosures. The checker encloses each matrix-radius norm before applying the same diagonal-cap test $\widehat{M} \preceq \mathrm{diag}(\rho^{\mathrm{src}})$.

The trial and residual expressions can also be evaluated without constructing $\mathcal{N}$. For a matrix of source right-hand sides V_{src}, set

$$Z_{\mathrm{src}} = \mathcal{A} + \Theta V_{\mathrm{src}}, \qquad D_{\mathrm{src}} = Q_{\mathrm{out}} Z_{\mathrm{src}} - V_{\mathrm{src}}, \qquad Y = \mathcal{N}^T V_{\mathrm{src}}. \tag{70}$$

Then the exact trial and residual energies are

$$\begin{aligned} M_Y &= K_{\mathrm{src}} + Z_{\mathrm{src}}^T Q_{\mathrm{out}} Z_{\mathrm{src}} - V_{\mathrm{src}}^T \Theta V_{\mathrm{src}}, \\ \mathcal{E}_Y^T \mathcal{E}_Y &= D_{\mathrm{src}}^T \Theta D_{\mathrm{src}}. \end{aligned} \tag{71}$$

Indeed, $\mathcal{E}_Y = \mathcal{N}^T D_{\mathrm{src}}$, and expansion of the first line gives the source trial in Eq.(66). Forward/adjoint actions implement the Θ contractions. When columns are processed in batches, the calculation accumulates every cross-batch contraction into the same full Gram matrix.

Outward coefficient and action bounds, together with the residual energy, produce a Loewner majorant $\widehat{M} \succeq M$. The final check establishes $\widehat{M} \preceq \mathrm{diag}(\rho^{\mathrm{src}})$ with $\rho^{\mathrm{src}} \geq 0$. Consequently $\xi^T M \xi \leq \sum_i \rho_i^{\mathrm{src}} \xi_i^2 \leq \sum_i \rho_i^{\mathrm{src}}$ on $[-1,1]^s$.

D.4 The recentered source bound

An alternative upper bound on the same source matrix M separates the deterministic source response from its enclosure uncertainty: $\mathcal{A} = A_0 + \Delta_A$, where A_0 is the stored midpoint and $|\Delta_A| \leq \Delta_A^{\mathrm{rad}}$. Here Δ_A encloses uncertainty about the fixed QC system's source-response matrix, while the original-graph residuals remain in b_v.

Proposition 9 (Recentered full-source bound)**.** *Under Theorem 8, let $Q_{\mathrm{out}} \succeq 0$, $E \succeq \mu I$ with $0 < \mu \leq 1$, and $\Theta = \mathcal{N}\mathcal{N}^T$ on the full error space. Define $H_\mu(A) = A^T Q_{\mathrm{out}} A + \mu^{-1}(Q_{\mathrm{out}}A)^T \Theta (Q_{\mathrm{out}}A)$. For every $\eta_{\mathrm{rec}} > 0$,*

$$M \preceq K_{\mathrm{src}} + (1+\eta_{\mathrm{rec}}) H_\mu(A_0) + (1+\eta_{\mathrm{rec}}^{-1}) \mu^{-1} \Delta_A^T Q_{\mathrm{out}} \Delta_A. \tag{72}$$

Proof. Set $Q_{\mathrm{eff}} = Q_{\mathrm{out}} + Q_{\mathrm{out}} \mathcal{N} E^{-1} \mathcal{N}^T Q_{\mathrm{out}}$, so $M = K_{\mathrm{src}} + \mathcal{A}^T Q_{\mathrm{eff}} \mathcal{A}$. The gap gives

$$Q_{\mathrm{out}} \preceq Q_{\mathrm{eff}} \preceq \mu^{-1} Q_{\mathrm{out}}, \qquad Q_{\mathrm{eff}} \preceq Q_{\mathrm{out}} + \mu^{-1} Q_{\mathrm{out}} \Theta Q_{\mathrm{out}}. \tag{73}$$

For the first upper bound, $S = Q_{\mathrm{out}}^{1/2} \mathcal{N}$ has $I - SS^T \succeq \mu I$ because its nonzero Gram spectrum agrees with that of $S^T S$ and $\mu \leq 1$. The identity $Q_{\mathrm{eff}} = Q_{\mathrm{out}}^{1/2} (I - SS^T)^{-1} Q_{\mathrm{out}}^{1/2}$ then applies even for singular Q_{out}. The second upper bound follows from $E^{-1} \preceq \mu^{-1} I$. Apply the weighted Young inequality in the PSD form Q_{eff} to $\mathcal{A} = A_0 + \Delta_A$, assigning factors $1 + \eta_{\mathrm{rec}}$ and $1 + \eta_{\mathrm{rec}}^{-1}$ to the center and error energies. Bound the center by the second inequality in Eq.(73) and the error by the first to obtain Eq.(72). □

The computation fixes A_0, encloses $|\Delta_A| \leq \Delta_A^{\mathrm{rad}}$, and includes coefficient and action errors in both energies. The smallest checked scalar cap is used in Theorem 10.

E Original-graph discrepancies and numerical assurance

The fixed QC response need not equal the centered original network. Additive discrepancies account for this difference, and their bounds combine with the matrix checks of Appendix D to certify the original margin.

E.1 Local discrepancies and propagation

The local envelopes of Eq.(24) separate a node's centered value into its fixed linear response, assigned residual, and local discrepancy $\delta_a(\xi)$. On the common original-execution assignment of Proposition 5, the propagated discrepancy obeys $b_a = \sum_b \alpha_{ab} b_b + \delta_a$, with $\delta_a(\xi) \in \mathcal{A}_a$. Here $\mathcal{A}_a$ encloses the local discrepancy; b_a also includes discrepancies inherited through the fixed parent coefficients.

The local discrepancy enclosure includes the following contributions:

- Original biases and nominal-operation mismatch relative to the stored center, to the extent not already accounted for by that center.
- Prefix-state and convolution-context uncertainty.
- Differences between the original physical input radii under the stated real-arithmetic model and stored binary64 perturbation scales, applied over the full perturbation box.
- Differences between original interval coefficients and the fixed homogeneous QC coefficients, applied to the full original parent ranges.

- Nonlinear residuals not represented by active assigned error coordinates, including the full residuals of inactive or below-threshold nonlinearities.

The propagated discrepancy b_a is the difference between the original node and its fixed input-and-error response. The polynomial remainder is the difference between that node and its retained polynomial; its enclosure supplies the value ranges used to construct the local constraints.

Amplitude deficits. An assigned error coordinate represents its declared residual amplitude, and the additive discrepancy receives any excess. Let a unary node compute f with fixed slope σ, declared amplitude σ_a, coefficient κ, center c and parent range $\mathcal{D} \ni c$, and let $\bar{d}$ bound $|d|$ over $\{d : c + d \in \mathcal{D}\}$. What the coordinate has to cover is the largest deviation of a secant slope from σ on that same range:

$$\sup_{d\neq 0,\ c+d\in\mathcal{D}} \left| \frac{f(c+d)-f(c)}{d} - \sigma \right| \le \max\big(|\sigma^- - \sigma|,\ |\sigma^+ - \sigma|\big) = \text{needed}, \tag{74}$$

where $[\sigma^-, \sigma^+]$ is a checked enclosure of those secant slopes and *needed* is the resulting upper bound. For example, for $f(x) = x^2$ at $c = 0$ on $\mathcal{D} = [-1, 1]$ with $\sigma = 1$, the secant slopes lie in $[-1, 1]$. The deviation reaches 2 at $d = -1$, whereas using only the upper slope endpoint would give zero. Then

$$\text{deficit} = \max(0,\ \text{needed} - \sigma_a\kappa), \qquad \mathcal{A}_a \supseteq [-\text{deficit} \cdot \bar{d},\ \text{deficit} \cdot \bar{d}], \tag{75}$$

all endpoints enclosed outward. If the node has no assigned error coordinate, the same formula applies with $\sigma_a\kappa = 0$, and an enclosure of the full residual enters the additive discrepancy. For a product with parent radii $\bar{d}_u, \bar{d}_v$ and parent coefficients κ_u, κ_v, set $\varsigma = \max(1, \bar{d}_v/(\sigma_a\kappa_u), \bar{d}_u/(\sigma_a\kappa_v))$. The assigned error coordinate then represents the product residual up to the factor $1/\varsigma$, and the discrepancy receives the rest, $\pm\bar{d}_u\bar{d}_v(1 - 1/\varsigma)$, which vanishes exactly when the declared amplitude covers both parent rows. Coefficient and center deviations between the original graph and the fixed QC system enter the same local discrepancy enclosure. A deficit therefore weakens the bound through B_p and $\underline{b}_\phi$.

E.2 Output bound for the original network

Let $\ell_\phi = a_{\text{out}}^T z$ be the linear output response. On an original execution, $\phi = c + \ell_\phi + b_\phi(\xi)$, where c is the stored constant and b_ϕ is the propagated property discrepancy. Any mismatch between c and the exact nominal margin is included in b_ϕ.

For constraint row (a, r), apply its fixed parent coefficients to the propagated discrepancies:

$$(b_p)_{a,r} = \sqrt{\lambda_{a,r}} \sum_b \kappa_{ar,b} b_b, \qquad B_p \ge \sup_\xi \|b_p(\xi)\|^2, \qquad \max_a \frac{w_a^2}{\sum_r \lambda_{a,r}} \le \tau < 1. \tag{76}$$

Bound the discrepancy at the property node from below to obtain $\underline{b}_\phi$, including any outward correction required when subtracting the original multiclass logits.

If no residual is active, take $\tau = 0$.

For the common witness, local parent constraints give

$$\|e(\xi)\|^2 \le \tau\|p + b_p(\xi)\|^2, \qquad \|e(\xi)\|^2 - \|p\|^2 \le \frac{\tau}{1-\tau} B_p. \tag{77}$$

The second inequality follows from the first by completing the square in p.

Combining this original-execution bound with the free-error bound from Appendix D gives

$$\ell_\phi^2 + J \le U_{\text{src}}, \qquad -J \le \frac{\tau}{1-\tau} B_p. \tag{78}$$

The first bound holds for free errors; the second uses the errors of an original execution. Their sum bounds ℓ_ϕ^2 on that execution. Section 4.2 abbreviates these bounds as $U = U_{\text{src}}$ and $V = \tau B_p/(1 - \tau)$.

Theorem 10 (Output bound for the original network)**.** *Suppose the local envelopes admit the shared original-execution assignment, Theorem 8 applies, and the checks above establish Eq.(78). If $b_\phi(\xi) \geq \underline{b}_\phi$ on the input box, then*

$$\underline{m}_{\mathrm{QC}} = c + \underline{b}_\phi - \sqrt{U_{\mathrm{src}} + \frac{\tau}{1-\tau} B_p} \leq \inf_{X \in \mathcal{X}} \phi(X). \tag{79}$$

Proof. On each original execution, adding the inequalities in Eq.(78) bounds ℓ_ϕ^2. Hence $\ell_\phi \geq -\sqrt{U_{\mathrm{src}} + \tau B_p/(1-\tau)}$; combining this with $\phi = c + \ell_\phi + b_\phi$ and the uniform lower bound on b_ϕ proves Eq.(79). □

Candidate generation is described in Appendix G.

E.3 Sources of numerical enclosure error

Numerical error enters the certificate through three components:

1. *Domain construction.* Rounding bounds enclose polynomial coefficients in ξ, remainders, and full node/parent ranges. These bounds support reuse of a compatible polynomial enclosure, propagation, and local-envelope construction.
2. *Original graph versus fixed homogeneous QC system.* The local discrepancies in Appendix E.1 propagate through b_v, then through B_p and $\underline{b}_\phi$.
3. *Realization and checks for the fixed QC system.* Matrix compilation and operator actions use rounding bounds. Approximate solves enter the spectral, terminal, and source bounds through enclosed residuals (Appendix D).

The implementation assumes IEEE binary64 arithmetic, conservative reduction and matrix-product error bounds, and explicit underflow allowances. Nonlinear endpoints use directed MPFR evaluation followed by directed binary64 conversion. Selected operations use higher-precision ball arithmetic before outward conversion; Appendix F.1.3 specifies the libraries and precisions.

Direct certification validates the final polynomial, remainder, and support bounds. Full-QC also validates local positivity, the full-space and terminal conditions, the source majorant, and the original-graph discrepancies. To compute the final QC endpoint, round the square root upward and the final subtraction downward.

A second graph traversal, using the same directed-interval and envelope routines, rechecks the original-graph discrepancies and the correspondence between compiled blocks and the original graph. The supplementary code documents these checks and comparisons against sparse and dense operator realizations.

Representative certificate. For a Yelp task with $H = 16$ and $\varepsilon_{\mathrm{in}} = .1$, the checked Full-QC margin lower bound is approximately .2674, with $\beta = .9$ and $\bar{\chi}_{\mathrm{term}} \approx .7932$. The parent-discrepancy correction $\tau B_p/(1-\tau) \approx .001079$ is included in Eq.(79).

F Experimental protocol and additional results

The experiments below separate certified-radius coverage, enclosure construction, reuse of intermediate bounds, and joint certification. Additional evaluations cover multiple output margins, perturbation supports, model structures, and computational cost.

F.1 Experimental setup and reproducibility

F.1.1 Model training and tokenization

Yelp Polarity models use 100,000 training, 20,000 validation, and 38,000 test reviews; the AG News model uses 100,000, 10,000, and 7,600 inputs of its four classes. Training and validation rows are

drawn from the training split by a stratified split with a fixed seed, and the test rows come from the dataset's own test split.

The tokenizer lowercases text and separates words, numbers, and punctuation. The vocabulary contains the 20,000 most frequent training tokens plus padding, unknown, and classification tokens, giving 20,003 embeddings. Each input keeps its first 127 content tokens, is padded on the left to $T = 128$, and ends with the classification token. Each checkpoint learns its embedding table; the supplementary code specifies the tokenization and index mapping. The architecture comparison uses the same vocabulary and tokenized inputs (Appendix F.5.2).

Optimization uses AdamW at learning rate 5×10^{-5} with weight decay 0.005, gradient clipping 0.5, and cross-entropy loss. The learning rate warms up linearly over the first 2% of updates and then follows a cosine decay to a tenth of its peak. Runs last 4 epochs at effective batch size 32 (128 for AG News) with seed 0, use mixed-precision forward and loss on GPU with single-precision parameters, and select the checkpoint with the best validation accuracy, breaking ties by validation loss, with early stopping after 2 epochs without improvement. Test accuracy is measured after that selection. Every checkpoint in Table 4 uses a causal convolution of width 4 and state dimension $n = 16$. The architecture checkpoints of Appendix F.5.2 share that state dimension; the Mamba-3-style block has no convolution (Appendix C.4).

Table 4: Evaluated checkpoints trained with the recipe above. Parameters are in millions; accuracies are percentages.

Checkpoint	Task	L	d	m	Params	Val	Test
Main	Yelp Polarity	4	384	768	12.57	90.30	89.38
AG News	AG News, four classes	4	384	768	12.57	87.28	86.82
$d = 192$	Yelp Polarity	4	192	384	5.11	87.35	87.04
$d = 768$	Yelp Polarity	4	768	1536	34.58	91.54	91.63
$L = 1$	Yelp Polarity	1	384	768	8.90	89.68	88.80
$L = 2$	Yelp Polarity	2	384	768	10.13	90.29	89.23
$L = 8$	Yelp Polarity	8	384	768	17.46	89.50	89.44

F.1.2 Embedding-box scale

To interpret the absolute perturbation radii across checkpoints, we compare the box size with embedding norms and vocabulary distances. For a token embedding $e_t \in \mathbb{R}^d$, the coordinate box has corner ℓ_2 radius $R_2 = \varepsilon_{\text{in}}\sqrt{d}$. We report $E = 100R_2/\|e_t\|_2$, $N_2 = 100R_2/\delta_2(t)$, and $N_\infty = 100\varepsilon_{\text{in}}/\delta_\infty(t)$, where $\delta_p(t)$ is the exact distance to the nearest distinct non-padding, non-CLS vocabulary embedding in norm p. Ratios are computed per distinct experimental token center before taking the median. At $\varepsilon_{\text{in}} = .01$, medians across checkpoints range from 0.99–1.02% for E, 0.81–0.86% for N_2, and 0.32–0.36% for N_∞. At $\varepsilon_{\text{in}} = .1$, the corresponding ranges are 9.87–10.17%, 8.11–8.56%, and 3.16–3.56%. For an individual token, $N_\infty < 100\%$ means that its perturbation box contains no other token embedding.

F.1.3 Statistical and numerical conventions

Coverage includes all tasks, including unresolved and resource-limited queries. Bound distributions use available endpoints; each table specifies the samples included in timing statistics. Quartiles use linear interpolation. Exact aggregation conventions are given with the supplementary code.

The reported runs use CPython 3.10 with NumPy and SciPy on binary64, directed MPFR through `gmpy2` at 80 bits for nonlinear endpoints, and 192-bit Arb balls through `python-flint` for deep-model prefix computation and selected point-value linear/sum operations (Appendix E.3). Thread counts are fixed per run because reduction order can change numerical proposals.

F.1.4 Attack generation and validation

Attack candidates come from projected signed-gradient descent on the original model, in the normalized coordinates of the declared support: from a start point, $x \leftarrow \text{clip}(x - 0.1 \text{ sign } \nabla_x \phi, -1, 1)$.

The first start is the nominal input, or, after the first radius of a sweep, the previous radius's best point rescaled into the new box; the remaining starts are uniform on $[-1, 1]$. We use two starts with at most 40 updates each, subject to the time budgets in Appendix F.2.

The best floating-point margin selects a candidate for validation. The candidate is then mapped back to a physical input and validated independently: its distance to the nominal input is computed exactly in rational arithmetic on the declared support, and the original margin is enclosed with the interval backend. A strictly negative upper enclosure of that margin accepts the candidate as a witness, whose exact distance is $r_{\mathrm{atk},i}$.

F.2 Certified robustness radii

Benchmark selection. The benchmark draws 64 correctly classified Yelp Polarity reviews by deterministic stratified sampling over labels (positive/negative) and three content-length strata (short: 8–32, medium: 33–64, long: 65–127 retained content tokens), using a fixed seed. No verification outcome is consulted during selection. The resulting sample contains 32 positive and 32 negative reviews: 25 short, 24 medium, and 15 long. Each review contributes four positions, giving 256 tasks. With N content tokens in the $T = 128$ input, the four zero-based positions and affected horizons are

$$p_j = 127 - N + \lceil jN/4 \rceil - 1, \qquad H_j = 128 - p_j = N + 2 - \lceil jN/4 \rceil, \qquad j = 1, \ldots, 4.$$

Left-padding and CLS embeddings remain fixed. All 384 coordinates of the selected content row vary, and the original prefix supplies the state and convolution context.

The comparison includes IBP, CROWN-IBP, SSM-Affine$_0$, SSM-Affine$_1$, direct certification, and Full-QC on the same 256 tasks. IBP propagates coordinate intervals; CROWN-IBP combines forward interval bounds with backward affine propagation (Zhang et al., 2020). SSM-Affine$_0$ and SSM-Affine$_1$ are the separate graph-specific affine references defined in Appendix H.

The methods use three radius-search schedules chosen for their certification scales. SSM-Affine$_1$, direct certification, and Full-QC share a schedule that starts at .1, followed by .15 after a positive bound or .05 otherwise. Further queries refine or expand the tested radii according to the outcomes. SSM-Affine$_0$ uses an initial coarse-search range of 10^{-4}–10^{-2}, followed by bracket refinement, with at most eight scheduled queries and one geometric refinement. IBP and CROWN-IBP share an initial coarse-search range of 10^{-8}–10^{-6}, followed by bracket refinement. An attained radius is the largest tested radius yielding a positive margin lower bound.

The protocol allows four hours per radius and twelve hours per method and task, with eight and twenty-four hours, respectively, for four-position queries. Reported wall times are the measured costs of the endpoint computations. Full-QC parameter search and retries within each fixed-radius query are specified in Appendix G.

Attack budgets. The single-position experiment uses radius-searching attacks generated and validated as in Appendix F.1.4. A sweep anchors at the best existing positive radius, or the declared $\varepsilon_{\mathrm{in}}$ when none exists, and probes coarse multipliers $1, 2, 4, 8$ with at most eight probes and two interior refinements. Fixed-radius experiments spend 600 seconds per radius and 1,800 seconds per physical task, shared by all methods. Accepted witnesses retain their original checkpoint and perturbation support.

Radius coverage. Table 1 reports threshold coverage and conditional radius distributions. Figure 2 shows coverage across all attained radii, with all 256 tasks in every denominator.

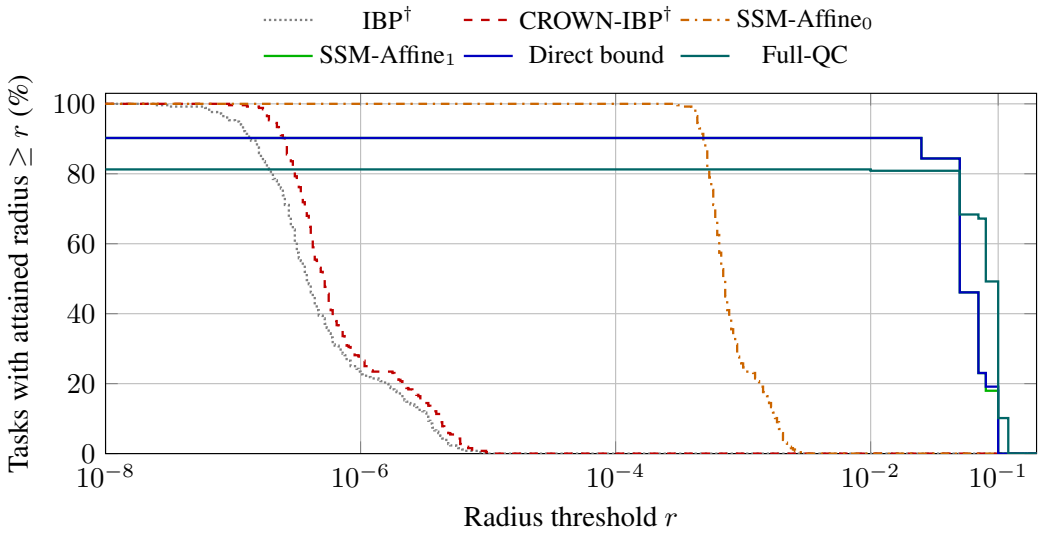

Figure 2: Attained-radius coverage over a logarithmic radius axis. At threshold r, coverage is the fraction of all 256 tasks with a recorded positive bound at radius r or above. Tasks without a positive endpoint remain in the denominator. The daggered curves are ordinary floating-point references; the others use outward rounding.

	Direct bound			Full-QC		
	.01	.05	.1	.01	.05	.1
8–32 tokens ($n = 100$)	100.0%	88.0%	12.0%	94.0%	93.0%	42.0%
33–64 tokens ($n = 96$)	89.6%	88.5%	25.0%	81.2%	81.2%	61.5%
65–127 tokens ($n = 60$)	75.0%	71.7%	21.7%	60.0%	60.0%	41.7%
first position quartile ($n = 64$)	73.4%	67.2%	18.8%	65.6%	65.6%	35.9%
second position quartile ($n = 64$)	87.5%	81.2%	10.9%	70.3%	68.8%	34.4%
third position quartile ($n = 64$)	100.0%	95.3%	12.5%	90.6%	90.6%	50.0%
fourth position quartile ($n = 64$)	100.0%	93.8%	34.4%	98.4%	98.4%	76.6%

Darker cells: greater coverage. Each column uses the full group denominator.

Figure 3: Certified-radius coverage by retained-content length and position quartile (first is earliest). Each cell is the percentage of all tasks in its group whose largest tested positive radius reaches the column threshold. Length strata and position quartiles partition the same 256 tasks.

Full-QC's higher coverage at .1 occurs in all three length strata. At .01, the direct bound and Full-QC certify more tasks at later positions (Figure 3); the affected suffix changes with position.

Attack upper radii and certified–attack intervals. For task i, let $r_i^\star$ be the distance to the closest input on the fixed support with nonpositive margin. The smallest validated attack distance gives an upper bound U_i on $r_i^\star$. We use the candidate's exact physical ℓ_∞ distance (Appendix F.1.4).

The attack search finds validated witnesses on 251 of the 256 tasks. Taking the smallest distance per task gives an attack upper-radius median of .375, interquartile range $[.259, .476]$, and range

$[.0324, 3.20]$. One task has an upper radius at most .05, and 13 have one at most .1. These witnesses are shared across the certification methods.

Let $L_{m,i}$ be method m's largest strictly certified radius. Whenever both endpoints are available,

$$L_{m,i} \le r_i^\star \le U_i.$$

For the combined result, set $L_i = \max_m L_{m,i}$ over the available endpoints from direct certification and Full-QC. The two routes jointly provide 235 lower endpoints; 231 tasks have both endpoints, four have only a lower endpoint, 20 have only an upper endpoint, and one has neither. All finite brackets satisfy $L_i < U_i$.

For the 231 finite brackets, the mean of L_i/U_i is approximately .26054. Since $L_i/r_i^\star \ge L_i/U_i$, the mean fraction of the true robustness radius attained by the best certificate is at least 26%.

Table 5: Certified–attack radius brackets on the 256 tasks. Certified counts available strict lower endpoints; finite brackets require both endpoints. Width and fraction are median $[Q_{25}, Q_{75}]$ within the corresponding finite-bracket subset. The last row takes the larger available certified lower radius per task.

Lower-bound source	Certified	Finite brackets	Width $U - L$	Fraction L/U
Direct bound	231/256	227/256	0.317 [0.204, 0.419]	0.152 [0.115, 0.220]
Full-QC	208/256	204/256	0.291 [0.174, 0.397]	0.243 [0.183, 0.314]
Best of the two	235/256	231/256	0.286 [0.172, 0.377]	0.239 [0.183, 0.312]

Table 5 measures the remaining distance between a certificate and a validated counterexample, as both an absolute width and the certified fraction of the attack radius. Figure 4 shows the paired endpoints. The combined bracket width has median .286; its interquartile range is $[.172, .377]$.

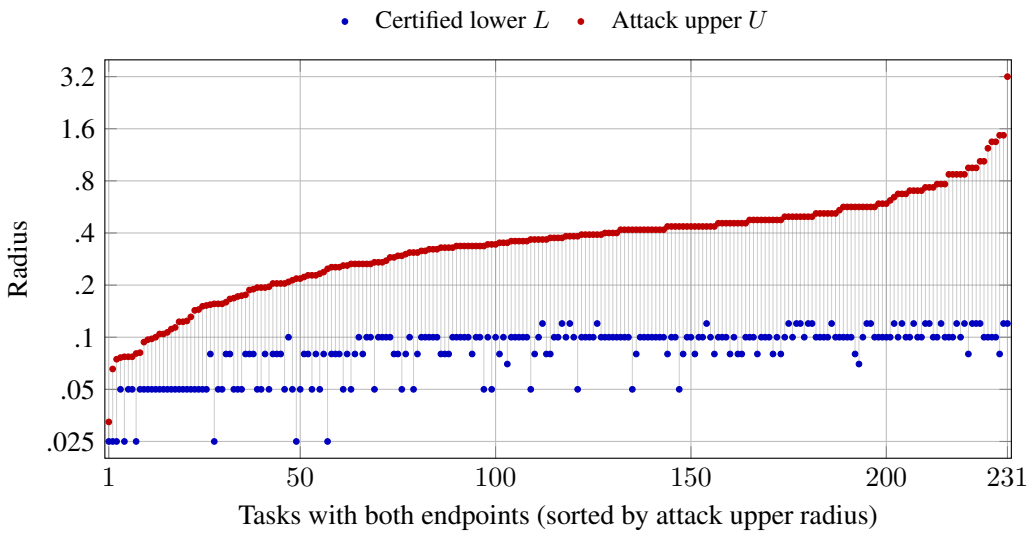

Figure 4: Certified–attack radius intervals for the 231 tasks with both endpoints, sorted by U_i. Each vertical segment joins the largest available strictly certified radius L_i to the smallest validated attack distance U_i for the same task. The vertical axis is logarithmic.

F.3 Enclosure construction

This experiment separates the effects of retaining quadratic terms, contracting histories, combining shared remainder images, and intersecting independent ranges. We use four tasks from Appendix F.4.3, two each at $H = 2$ and $H = 12$, and evaluate every configuration at .01 and .1.

Q2 is the degree-two enclosure; Q1 immediately encloses each newly formed quadratic term in the remainder. $C = 1$ accumulates transition parameters and contracts at the readout, while $C = 0$ propagates the recurrence stepwise. $R = 1$ combines the two linear contributions of each inherited remainder in projected RMS before bounding them; $R = 0$ bounds them separately. $G = 1$ applies the optional range intersections, while $G = 0$ computes the same optional bounds without applying them. The eight factorial configurations all retain degree two; Q1 uses $C = R = G = 1$.

F.3.1 Paired effects and local cancellation tests

For each of C, R, and G, a pair differs only in that switch: four tasks and four settings of the other two switches give 16 paired differences per radius and output method. The Q2/Q1 comparison uses four pairs with C, R, and G enabled. The local controls separately test whether the rules retain cancellation on objects with known ranges.

Table 2 also gives absolute endpoint medians and sign flips at .1. The direct bound remains nonpositive in all configurations. All affine flips occur on the same $H = 12$ task under different matched configurations; the three R flips correspond to three settings of C and G. None of the comparisons reverses a positive bound to nonpositive.

At .01, quadratic retention improves all four task bounds under both output methods; shared remainder images improve all 16 paired bounds, while optional range intersections change none. The largest median improvement at .01 is .0295 for the direct bound and .000750 for affine propagation, both from quadratic retention. At .1, history contraction gives a median paired improvement of .266 in the direct bound.

In a two-step decay cancellation control, accumulated history has width 3.05×10^{-16} versus 3.84×10^{-4} from separate bounding. Combining shared images also recovers the exact width .0005 of a shared cubic remainder, versus .0075 when its uses are bounded separately. For a radial inherited RMS remainder, the corresponding linear bounds are .000399 and .7992; tangent remainders show no such cancellation.

Layerwise interval widths. Figure 5 traces how replacing quadratic terms by intervals changes the maximum coordinate width across layers.

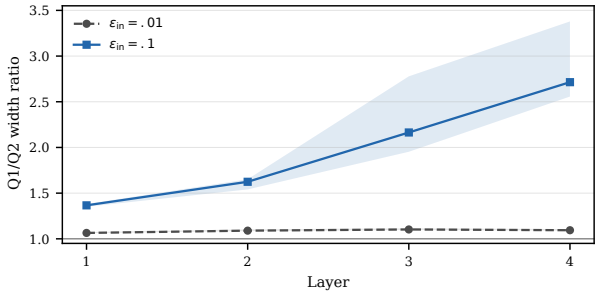

Figure 5: Effect of replacing quadratic terms by intervals across layers. Curves show the median Q1/Q2 ratio of the maximum coordinate-interval width at each layer output; bands show the range over the four tasks.

Shared remainders and range intersections. For Q2 with all three switches enabled, combining the two inherited RMS remainder images gives a smaller bound in 75% of evaluated output projections at each radius. Each comparison concerns one token's RMS enclosure and uses the maximum absolute interval endpoint across output coordinates. The maximum separate-to-combined bound ratio within each task ranges from 3.19 to 7.34 at .01 and from 3.18 to 7.35 at .1 across the four tasks. Among the intermediate domains included in these statistics, optional intersections tighten none at .01 and 2.5–4.3% at .1.

F.3.2 CONSTRUCTION COST

To assess the computational cost of history contraction, Table 6 compares construction time and peak memory with C on and off. The four tasks include two at $H = 2$ and two at $H = 12$. Ratios are computed for matching construction runs and then aggregated.

Table 6: Resources for history contraction ($C = 1$) and stepwise ($C = 0$) construction. Times are median wall-clock seconds and peak RSS is median GiB; ratios are paired $C = 0/C = 1$ medians.

H	ε	C1 time	C0 time	Ratio	C1 RSS	C0 RSS	Ratio
2	.01	183	564	3.09	4.24	15.54	3.66
2	.1	177	578	3.03	4.24	15.53	3.66
12	.01	1786	3086	1.76	15.53	18.19	1.17
12	.1	1677	3069	1.80	15.53	18.18	1.17

F.4 BOUND REUSE AND JOINT CERTIFICATION

The affine and GenBaB bound-exchange comparisons fix their output-bound procedures while changing the supplied intermediate bounds. Separate comparisons examine cross-radius behavior, end-to-end variants, and checks on fixed QC instances.

F.4.1 AFFINE BOUND REUSE ON 16 TASKS

This comparison asks how much of the output-bound improvement comes from supplying intermediate information to the same affine propagation. We fix eight reviews, positions with $H \in \{2, 12\}$, and radii .01, .05, and .1, giving 48 task–radius cells. All 384 coordinates at the selected position vary. The three affine configurations keep the graph, output-bound rules, and local-plane search procedure fixed: the intrinsic baseline, the baseline with polynomial-enclosure intervals, and the latter with compatible affine forms in ξ added. Direct certification and Full-QC use the same polynomial construction on these tasks.

Positive-bound counts use all 16 cells per radius. Paired bound differences use cells where both methods return a bound. Supplying polynomial-enclosure intervals changes the 46 completed affine cells from zero to 35 positive bounds. Adding affine bounds in ξ preserves that count and improves the endpoint by a median 0.0079 over the same 46 cells (41 higher, three equal, and two lower). Thus most of the observed change in positive-bound counts comes from the intermediate intervals. At .1, both supplied-bound affine variants certify five of the 16 cells; SSM-Affine$_0$ certifies none, direct certification succeeds on three, and Full-QC certifies four. Twelve Full-QC queries at this radius return no endpoint.

Direct certification yields 34 positive endpoints. Combining these with the affine method receiving both types of bounds covers 36 cells. Full-QC has 35 positive endpoints, all covered by that union. On those 35 returned endpoints, the median difference between Full-QC and the direct bound is -0.0169; the radius-wise medians are -0.0189, $+0.00152$, and $+0.3788$.

F.4.2 GENBAB SEARCH AND INTERMEDIATE-BOUND EXCHANGE

We select eight reviews using the same stratification by label and content length as the main panel (Appendix F.2). Each review contributes one task at $H = 2$ and one at $H = 7$–33. All 16 tasks use the main Yelp checkpoint, perturb all 384 embedding coordinates at one position, and fix $\varepsilon_{\text{in}} = .1$. Direct certification and Full-QC use outward rounding; GenBaB uses ordinary FP64, so its positive bounds are floating-point references.

Direct comparison. GenBaB computes intermediate intervals with CROWN and IBP, fixes them during root optimization, and uses shared relaxation parameters and nonlinear branching. Our methods and both GenBaB configurations use an RTX 3090 with 24 GiB GPU memory, four CPU threads, a 96 GiB host-memory limit, and four hours per task. Paired configurations run on the same host without concurrent GPU queries. Both configurations allow 300 root-optimization itera-

tions. G-bound allows 100 iterations per BaB bound optimization; G-search reduces this to 50 with earlier stopping. The supplementary code specifies their optimizer settings.

Table 7: Positive margin bounds at $\varepsilon_{\text{in}} = .1$, with eight tasks per horizon group. Daggers mark FP64 references; the other methods use outward rounding.

Method	$H = 2$	$H = 7$–33
GenBaB (G-bound)†	0/8	0/8
GenBaB (G-search)†	0/8	0/8
Direct bound	1/8	1/8
Full-QC	6/8	5/8

G-bound enters branching on 8/8 short- and 4/8 long-horizon tasks. All G-bound runs end at a time or memory limit; four long-horizon tasks exhaust GPU memory before branching. Direct certification returns six nonpositive bounds and one domain-check failure in each horizon group. Some Full-QC runs remain unresolved at the local-positivity check.

Root-only exchange. The analytic-range control supplies bounds for normalization, its projections, step sizes, and transition factors. The TaylorSSM variant imports polynomial-enclosure intervals at matching graph quantities. Both use the same CROWN root objective, shared relaxation parameters, four CPU threads, FP64, and a 300-iteration cap. Imports are intersected with existing ranges, and downstream bounds are recomputed before root optimization. Runs allow four hours and stop after the root calculation.

Table 8: Root lower bounds at $\varepsilon_{\text{in}} = .1$ (ordinary FP64). Entries give positive/completed evaluations, followed by the median lower bound; each horizon group contains eight tasks. Long-horizon exchange statistics use the same three completed tasks at $H = 7, 8, 9$. The dash denotes an unreported statistic.

Intermediate ranges	$H = 2$	$H = 7$–33
Native	0/8 (-2.03×10^{22})	—
Analytic	0/8 (-1.74×10^{7})	0/3 (-5.73×10^{7})
TaylorSSM intervals	4/8 (-0.252)	0/3 (-2.09)

TaylorSSM intervals improve all eight short-horizon root bounds over the analytic-range control, with four becoming positive. Both modes complete the same three of eight longer-horizon tasks ($H = 7, 8, 9$); TaylorSSM improves all three bounds, but both modes' bounds remain nonpositive.

The remaining five tasks per mode are excluded from completed-root statistics: three time out before returning a root bound and two exhaust GPU memory before exchange.

F.4.3 Output-bound contributions on 32 tasks

A separate 32-task Yelp panel examines output-bound contributions at radii $.01, .025, .05, .1$, with $H \in \{2, 12, 14, 16\}$. Tasks are selected by fixed-seed sampling stratified by label and content length; token and H vary across tasks. Comparing the affine variants of Appendix H.1—Affine + enclosure intervals and SSM-Affine$_1$—isolates the addition of affine bounds in the input perturbation variables, holding the output-bound rule fixed. Direct certification and Full-QC use different output computations on the same tasks and boxes. All methods use outward-rounded arithmetic. Table 3 gives coverage. "Higher/equal/lower" counts compare the first method named in each difference.

Table 9: Paired endpoint differences across radii. Entries are median $[Q_{25}, Q_{75}]$, followed by higher/equal/lower counts and $n = 32$. Differences and ties use the precision of the recorded endpoints.

Radius	Full-QC − direct bound	SSM-Affine$_1$ − Affine + enclosure intervals
0.01	-0.0253 [-0.0262, -0.0224]; 0/0/32	0.0075 [0.0052, 0.0088]; 32/0/0
0.025	-0.0319 [-0.0441, -0.0244]; 1/0/31	0.0024 [0, 0.027]; 23/9/0
0.05	-0.109 [-0.345, -0.057]; 6/0/26	0 [0, 0.0121]; 12/19/1
0.1	2.77 [2.21, 4.51]; 27/0/5	0.0355 [0.0345, 0.0625]; 30/1/1

To relate the change in coverage to the bound construction, Table 10 separates the linear, quadratic, and remainder contributions to the direct bound and the Full-QC square-root penalty.

Table 10: Median bound contributions on the 32-task comparison. The first three columns give the linear radius and the lower quadratic and remainder contributions to the direct bound. The QC column gives the Full-QC square-root penalty, excluding $c + \underline{b}_\phi$.

ε_{in}	Direct linear	Direct quadratic	Direct remainder	QC penalty
.01	.0463	−.0017	−.0014	.0689
.025	.1159	−.0104	−.0240	.1701
.05	.2317	−.0414	−.2423	.5973
.1	.4635	−.1657	−4.5771	2.1363

These component medians are consistent with the paired endpoint differences. At small radii, the median Full-QC penalty exceeds the summed magnitudes of the component medians of the direct bound; at .1, the remainder dominates that sum.

Limits of optimizing a fixed final enclosure. We test whether exact minimization of the retained polynomial could certify the margin with unchanged remainder and coefficient-error bounds. For the stored linear-coefficient intervals $[a_i^-, a_i^+]$, put $\bar{a}_i = (a_i^- + a_i^+)/2$ and $d_i = (a_i^+ - a_i^-)/2$. For the fixed quadratic coefficients $q_{\phi,ij}$ of Eq.(22), define the symmetric matrix Q_{pol} by $(Q_{\text{pol}})_{ii} = q_{\phi,ii}$ and $(Q_{\text{pol}})_{ij} = (Q_{\text{pol}})_{ji} = q_{\phi,ij}/2$ for $i < j$. Let δ_q bound their total coefficient error on $[-1, 1]^s$. With the saved constant and remainder lower bounds $\underline{c}, \underline{r}$, define

$$\begin{aligned} \kappa_{\text{pol}} &= \underline{c} + \underline{r} - \sum_i d_i - \delta_q, \\ f_{\text{pol}}(\xi) &= \kappa_{\text{pol}} + \bar{a}^T \xi + \xi^T Q_{\text{pol}} \xi, \qquad m_{\text{pol}} = \min_{\xi \in [-1,1]^s} f_{\text{pol}}(\xi). \end{aligned} \tag{80}$$

We search for feasible points in $[-1, 1]^s$ and evaluate the selected points using exact rational arithmetic on the saved binary64 coefficients, converting the result upward to binary64. Thus a returned $u \geq f_{\text{pol}}(\xi_*)$ with $u \leq 0$ proves $m_{\text{pol}} \leq 0$. It rules out a positive certificate from this fixed additive surrogate.

On the same 32 tasks at .1, 27 have such nonpositive witnesses. Their upward value bounds have median -0.9158 and range $[-8.4861, -0.1547]$.

Small-radius behavior. For an input box of radius ε, let g be the nominal margin gradient in the perturbed coordinates. With a first-order-exact frontend, the penalty in the direct bound is $\varepsilon \|g\|_1 + O(\varepsilon^2)$. In ideal arithmetic, Full-QC has the same leading term if its source matrix satisfies $M = \varepsilon^2 g g^T + o(\varepsilon^2)$, its diagonal cap is asymptotically tight, $\underline{b}_\phi = o(\varepsilon)$, and $\tau B_p/(1 - \tau) = o(\varepsilon^2)$ in Eq.(79). These conditions have not been established for the implemented pipeline: local QC relaxation, source bounds, and candidate search can each leave a first-order gap. Finite-radius comparisons alone do not determine the limit.

Check and search parameters. Increasing β relaxes the internal spectral threshold but amplifies the terminal residual correction by $1/(1-\beta)$: factors $10, 20, 40$ at $\beta = .9, .95, .975$ (Eq.(61)). The search target controls the estimated candidate norm, while the enclosed checks determine acceptance (Appendix G.3).

Discrepancy amplification. The coefficient $\tau/(1-\tau)$ is $999, 99, 9$ at $\tau = .999, .99, .9$; its effect on the margin depends on B_p relative to U_{src} inside the square root of Eq.(79).

F.4.4 End-to-end construction variants

This experiment asks whether changing the history construction or separating shared boundary responses changes end-to-end coverage. It varies two choices from Section 3 and Appendix D on 12 reviews in three horizon bands: four at $H = 2$, four at $H \in \{14, 16\}$, and four at $H \in \{53, 90, 97\}$. Stepwise-QC retains the same perturbation variables, quadratic order, and remainder-enclosure rules, together with the RMS, affine, and joint response–energy computations, but builds scan domains with a stepwise second-order recurrence. The primary history/readout construction contracts state coordinates with their readouts before forming polynomial coefficients in ξ. Separated-QC keeps the same polynomial construction and replaces the joint boundary cost in the full-space check of Appendix D by separate positive-semidefinite upper bounds. Each variant constructs its own domains and QC candidates and runs its own multiplier search. The two variants run on the eight tasks with $H \le 32$; the two primary routes run on all twelve.

Table 11: Construction comparison by affected horizon; each band has four tasks. Entries count positive certificates. Stepwise-QC and Separated-QC are evaluated on the short and medium bands.

	$H \le 8$			$9 \le H \le 32$			$H \ge 33$		
Method	.01	.05	.1	.01	.05	.1	.01	.05	.1
Full-QC	4	4	2	4	4	1	3	3	0
Direct bound	4	4	0	4	3	0	3	3	1
Stepwise-QC	4	4	2	4	4	0	—	—	—
Separated-QC	0	0	0	0	0	0	—	—	—

F.4.5 Checks on fixed QC instances

Two ablations each fix one QC instance and vary only the computation of its boundary or source bound. At the tested boundary, let $W \succeq 0$ be the effective future cost, R the diagonal residual map, G the temporal response to RMS outputs, and C_{in} the input cost. The joint boundary cost is positive semidefinite, with blocks $A = RWR$, $B = C_{\mathrm{in}} + G^T W G$, and $C = RWG$. The separated control drops the cross block C using the PSD bound

$$\begin{pmatrix} A & C \\ C^T & B \end{pmatrix} \preceq \mathrm{diag}((1+\eta_{\mathrm{sep}})A, (1+\eta_{\mathrm{sep}}^{-1})B), \qquad \eta_{\mathrm{sep}} > 0. \tag{81}$$

The implementation proposes η_{sep} from checked norm estimates and clips it to $[1/64, 64]$. A boundary with zero terminal storage needs no inflation.

The source control holds the domain, QC, and terminal gaps fixed. The reference uses the best checked source cap among the residual correction, zeroth-order spectral bound, and first-moment bound in Eq.(67), Eq.(68), and Eq.(69), respectively; the recentered evaluation uses $\eta_{\mathrm{rec}} = .001$. Before checking the diagonal cap, we add outward-rounded diagonal corrections to enclose source-matrix uncertainty.

For the fixed boundary QC at $H = 3$ and $\varepsilon_{\mathrm{in}} = .01$, separating the boundary cost leaves RMS positivity unresolved, whereas the joint check is accepted with margin 1.94466. For the fixed AG News source QC at $H = 2$ and $\varepsilon_{\mathrm{in}} = .05$, recentering changes the checked margin from -0.83762 to 0.20842.

F.5 Additional model and perturbation settings

F.5.1 Multiple margins and perturbed positions

The all-class experiment tests whether one intermediate construction can support every competing output margin. Its unit is an AG News input: a positive result requires all three margins to be positive on the same box. The 16 inputs include four per class, with four in the $H = 2$–7 band, eight in $H = 8$–23, and four in $H = 24$–32. Full-QC shares the domains and internal representation while computing terminal and source bounds for each margin.

The joint-support experiment tests dependence between perturbed positions. Its 80 tasks comprise sixteen each of single positions, contiguous pairs, and dispersed pairs, plus sixteen supports per four-position geometry. The boxes retain all cross-position products in $s = 384q$ variables, with nested supports at a fixed earliest position. Table 12 reports coverage separately by support size and geometry; enlarging a support changes both the representation size and the allowed perturbation set.

Table 12: Fixed-radius coverage of all-class decisions and joint supports. Entries count positive bounds from independent queries at each radius.

			Direct bound			Full-QC		
Setting	Tasks	s	.01	.05	.1	.01	.05	.1
All competitors	16	384	16	9	0	16	10	0
One position	16	384	16	13	0	16	15	4
Two, contiguous	16	768	14	4	0	14	4	2
Two, dispersed	16	768	15	9	0	14	7	2
Four, contiguous	16	1536	3	0	0	5	1	1
Four, dispersed	16	1536	3	0	0	5	0	1

Most failed Full-QC queries are unresolved or reach the time limit, whereas direct certification returns negative lower bounds.

F.5.2 Certification across selective recurrence structures

This experiment tests whether the enclosure and composition rules produce certificates across different selective structures. All three checkpoints use $L = 3$, $d = 192$, $m = 384$, and $n = 16$. They share a vocabulary, tokenized inputs, and a 100,000/20,000/38,000 Yelp Polarity split, and are trained separately with learning rate 3×10^{-4} and effective batch size 512. Post-SiLU Mamba-1 changes selection wiring; Mamba-2 uses 24 heads with shared selectors and gated RMS over its inner coordinates; the Mamba-3-style model uses rank-one injection, rotations on half of the state coordinates, and no output normalization.

A task is one review at one perturbed position. The 32 reviews are queried at positions with $H \in \{2, 12, 22, 32\}$. Each model has 128 tasks. The total wall-clock limit is 7,200 seconds per task. Direct certification and Full-QC use the same polynomial construction. Table 13 gives the checkpoint accuracy, perturbation radius, and coverage for each model.

Table 13: Coverage across selective structures. Acc. is test accuracy in percent; scale is $E/N_2/N_\infty$ in percent. Coverage entries count positive certificates over all recorded tasks for that model.

Model	ε_{in}	Acc.	Scale	Direct bound	Full-QC
Mamba-1 post-SiLU	.1	88.1	9.87/8.56/3.55	59/128	96/128
Mamba-2	.01	89.0	.99/.86/.36	88/128	99/128
Mamba-3 style	.01	90.3	.99/.86/.36	86/128	107/128

Table 14 separates coverage by affected horizon and reports the Full-QC query costs. The 17 Mamba-2 and 11 Mamba-3-style resource stops at $H = 32$ are timeouts. Wall-time statistics ex-

clude these runs, leaving 15 and 21 queries, respectively; every other group contributes 32 queries. Position changes with H across these groups.

Table 14: Coverage and resource use by model and horizon. Each group contains 32 tasks, all included in the coverage denominators. Full-QC timeouts are reported separately. Wall time is median [minimum, maximum] seconds over non-timeout Full-QC queries; memory is the maximum recorded peak host usage. The supplementary results distinguish nonpositive bounds from unresolved checks.

Model	H	Direct certified	Full-QC certified	Full-QC timeouts	Wall time (s)	Peak memory (GiB)
Mamba-1 post-SiLU	2	17/32	27/32	0	114 [108,151]	2.67
	12	3/32	24/32	0	871.5 [709,1210]	9.10
	22	12/32	21/32	0	2300.5 [1324,2753]	16.55
	32	27/32	24/32	0	4813 [1948,5233]	25.05
Mamba-2	2	21/32	30/32	0	121.5 [107,133]	3.13
	12	31/32	27/32	0	898.5 [776,991]	9.14
	22	28/32	27/32	0	1929 [1815,2799]	15.31
	32	8/32	15/32	17	3733 [3051,7200.5]	24.66
Mamba-3 style	2	28/32	31/32	0	114.3 [113.8,114.9]	2.88
	12	26/32	30/32	0	871.2 [808,914]	9.46
	22	21/32	27/32	0	1823 [1699.5,2511]	15.77
	32	11/32	19/32	11	2918 [2510,7200]	23.18

F.6 Computational cost

F.6.1 End-to-end cost and scaling diagnostics

The position experiment measures the cost of producing an output bound, including intermediate enclosures and the output calculation. It fixes the second Yelp input in Table 16 and the main checkpoint, and varies three positions ($H = 2, 8, 16$) and three radii ($.01, .05, .1$). Queries use an EPYC 7402 with an RTX 3090, four CPU threads, a 96 GiB host-memory limit, and exclusive use of the machine. Wall time covers the complete query, and peak memory is measured on the host.

Table 15 summarizes available timings across radii at each fixed position. Costs include every recorded finite endpoint. The Full-QC $H = 8$, radius .1 query contributes no finite endpoint to this summary, leaving two timed endpoints in that group. Each horizon corresponds to a different perturbed token embedding.

Table 15: End-to-end cost across three radii per position. Timed is the number of available endpoint computations out of three queries; positive counts use all three queries. Wall time is median [minimum, maximum] seconds, and memory is the maximum peak host usage over timed queries.

H	Method	Timed	Positive	Wall time (s)	Peak memory (GiB)
2	Direct bound	3/3	2/3	43.1 [42.6, 43.6]	1.86
2	Full-QC	3/3	3/3	220.2 [192.4, 251.8]	2.01
8	Direct bound	3/3	2/3	178.3 [177.9, 178.5]	7.01
8	Full-QC	2/3	2/3	780.6 [725.6, 835.5]	6.99
16	Direct bound	3/3	2/3	453.4 [448.7, 458.1]	15.55
16	Full-QC	3/3	3/3	1802.4 [1592.7, 2170.2]	15.66

On radius-matched queries, the median wall-time ratios of Full-QC to direct certification are 5.17 at $H = 2$ ($n = 3$), 4.38 at $H = 8$ ($n = 2$), 3.93 at $H = 16$ ($n = 3$). Peak memory is similar between routes at a given position, while the Full-QC output calculation increases elapsed time.

Table 16 extends the diagnostics to two fixed Yelp cases, varying affected suffix length, residual width, or network depth. The width and depth panels use the corresponding checkpoints in Table 4; the first $H = 2$ setting also provides their reference configuration. Resource summaries in this table

Table 16: Scaling diagnostics on two fixed Yelp inputs. Bound entries give the two inputs' margin lower bounds in a fixed order; bold positive values certify robustness. A dash means that no valid bound was returned. Wall time is the median over both inputs, and memory is their maximum peak host usage, including runs without a bound. The panels vary affected suffix length H, residual width d, and network depth L.

Setting	ε_{in}	Margin lower bounds		Wall time (s)		Peak memory (GiB)	
		Direct	Full-QC	Direct	Full-QC	Direct	Full-QC
Affected suffix length ($L = 4,\ d = 384$)							
$H = 2$	.05	**+.03675/+3.104**	**+.1401/+3.114**	42.4	228.3	1.86	2.01
	.10	$-3.505/-1.022$	—/**+1.401**	43.0	309.1	1.71	2.01
$H = 8$	.05	$-.3676$/**+2.804**	$-.04737$/**+2.878**	168.1	1568.3	7.03	7.01
	.10	$-1.055\times10^{4}/-4.182\times10^{12}$	—/—	169.3	1561.0	7.05	7.08
$H = 16$	.05	$-.5619$/**+3.038**	$-.4418$/**+2.951**	427.3	3016.4	15.73	15.68
	.10	$-483.3/-.02347$	—/**+1.362**	422.1	2505.6	15.65	15.62
$H = 32$	.05	$-.3616$/**+2.976**	$-.04688$/**+2.952**	1235.4	7149.4	32.91	32.85
	.10	—/-1.454	—/**+1.200**	1206.3	5030.6	32.83	33.03
Residual width ($L = 4,\ H = 2$)							
$d = 192$	.05	-2.755/**+3.062**	-2.760/**+3.045**	16.1	113.8	0.90	1.73
	.10	-4.468/**+2.304**	-3.676/**+2.310**	16.1	125.6	0.90	1.73
$d = 768$	.05	$-17.37/-69.95$	-15.80/—	1594.3	3762.0	29.24	29.30
	.10	—/—	—/—	833.5	2656.1	28.90	28.91
Network depth ($d = 384,\ H = 2$)							
$L = 1$	.05	$-.9310$/**+2.407**	$-.9575$/**+2.384**	10.1	60.7	0.85	1.58
	.10	-1.539/**+1.765**	-1.438/**+1.710**	10.1	81.3	0.84	1.58
$L = 2$	.05	$-.4022$/**+5.578**	—/**+5.467**	18.2	117.1	1.42	1.67
	.10	-2.330/**+3.477**	—/**+3.291**	18.4	115.0	1.42	1.66
$L = 8$	.05	—/—	—/—	113.7	865.5	2.76	2.38
	.10	—/—	—/—	91.8	779.0	2.16	2.49

include both cases, including runs that end without a bound. Missing bounds arise from unresolved certification checks or failures in domain computation; neither method returns a bound for either input at $L = 8$ at the two tested radii.

G Certificate construction and search rules

The search proposes local QC parameters and numerical trials; the enclosed checks in Appendices D and E determine whether they certify the original margin.

G.1 End-to-end certificate computation

Given the network, perturbation map, output properties, route, and budget:

1. Enclose the prefix and construct polynomial enclosures in the common input variables ξ (Appendix A).
2. For direct certification, contract the original head difference and apply the component bounds in Appendix A.5.
3. For Full-QC, instantiate local constraints, compute the structured response, and validate the local, full-space, terminal, source, and original-graph bounds. Combine accepted quantities in Eq.(79).
4. Accept a scalar property at a positive bound; accept an all-class property only when every competing margin is positive on the same box.

Domains may be shared across margins. For multiclass tasks, build one QC system covering all margin outputs. Choose coordinate units and initial weights using a designated output direction. Set the shared source budget from the smallest nominal margin. Each margin receives its own terminal,

source, and property-discrepancy checks. A combined policy may take the larger valid route bound on a task, charging both computations.

G.2 Instantiating the local constraints

One pass through the original graph assigns unary slopes and validated secant/envelope amplitudes, and the product constraints of Section 4.1, on the adopted domains. Nonlinearities whose validated amplitude exceeds the activity threshold receive one residual coordinate; all remaining residuals enter the additive discrepancy (Appendix E.1). Propagate diagonal energy backward from the property to set the coordinate units, which scale the response.

Positive constraint weights use logits $\ell_{a,r}$, and source weights have shape $\rho^{\mathrm{src}} \propto \mathrm{softmax}(\ell^{\mathrm{src}})$, with $s = qd$ source logits. The implemented weight and residual-scale parameterization is

$$\sqrt{\lambda_{a,r}} = \exp\big(\tfrac{1}{2}\,\mathrm{clip}(\ell_{a,r}, -55, 45)\big), \qquad w_a = \sqrt{0.999 \textstyle\sum_{r\in\mathcal{I}_a} \lambda_{a,r}}. \tag{82}$$

Thus the nominal τ is .999; the checker encloses the actual ratio outward. Rows tied by a shared RMS sector use one parameter. Initial weights are $\lambda_{a,r} = \max(|g_a|, 10^{-11})/|\mathcal{I}_a|$, where g_a is the unscaled output adjoint at residual a, distributed equally over its parent rows. Source logits start at zero. The automatic source budget is $\max(10^{-8}, .98c^2)$, with c the nominal margin (the smallest competitor margin for multiclass tasks).

G.3 Multiplier proposal

With $\tilde{\xi} = \mathrm{diag}(\rho^{\mathrm{src}})^{1/2}\xi$ and scaled residuals e, define $\mathcal{T}_{\mathrm{qc}} : (\tilde{\xi}, e) \mapsto (\ell_\phi, p)$. The compiled rows in Eq.(26) give its forward action and algebraic transpose. Search uses the original residual coordinates; the checks use the invertible transformed coordinates of Theorem 8. The condition $\sigma_1(\mathcal{T}_{\mathrm{qc}}) \le 1$ expresses the free-error inequality for this candidate, with source budget ρ^{src}. Search minimizes the estimated norm over $(\ell, \ell^{\mathrm{src}})$, taking the maximum over competing outputs.

Block power iteration estimates the norm, and Adam updates the logits with learning rate .06 toward a target of .81. Search returns the best iterate when it reaches the target, progress stalls, or the time budget is exhausted. Randomized initial blocks and balancing probes use fixed seeds; the exact settings are given in the supplementary code's numerical-settings guide. Wall-time budgets can change the number of updates and the returned candidate.

GMRES produces terminal and source trials using the structured operator actions. Finite incomplete iterates are admissible because the checks enclose their residuals; terminal trials must also satisfy $v \in \mathrm{range}(\mathcal{N})$ (Appendix D.2). The numerical search does not establish validity: acceptance requires the enclosed checks in Appendices D and E.

G.4 Search attempts and stopping rules

In the reported configurations, each fixed position–radius query permits at most three attempts within its wall-time budget, counting completed and failed attempts. Retries may relax the internal spectral threshold or repeat multiplier search with a stricter target; the supplementary numerical-settings guide gives the updates. Search stops at a resource limit or when it certifies the requested property. The radius experiment searches one proposal per query. It permits retries with increased β within the same three-attempt budget, but stops when a new multiplier search would be needed. Each radius initializes its own QC parameters; only attack search reuses a previous radius's best point (Appendix F.1.4).

We report the largest bound among completed proofs; a query with no completed proof is unresolved.

H The SSM-Affine reference

SSM-Affine adapts the forward/backward affine propagation of Transformer verification (Shi et al., 2020) to the selective graph. Its variants use the same graph, propagation rules, and local-plane search; they differ only in the intermediate bounds supplied to the calculation.

H.1 Variants and intermediate bounds

SSM-Affine$_0$ constructs its own interval and affine bounds, including analytic range bounds for projected RMS outputs and projections. All variants maintain intrinsic bounds $L_z^A(\xi) \le z(\xi) \le U_z^A(\xi)$ on the original box $[-1,1]^s$, recomputing prefixes from the initial state on the shared perturbation coordinates.

The interval-only variant, called *Affine + enclosure intervals* in the experiments, additionally intersects a matching node range with a valid polynomial-enclosure interval G_z. SSM-Affine$_1$ also uses the componentwise enclosure $z(\xi) \in \mathbf{c}_z + \mathbf{a}_z^T\xi + \mathbf{r}_z$, where $\mathbf{a}_z = [\underline{a}_z, \overline{a}_z]$ and $\mathbf{r}_z$ encloses the quadratic part and remainder together, as in Section 3.5. With $\bar{a}_z = (\underline{a}_z + \overline{a}_z)/2$ and $\delta_z = \frac{1}{2}\|\overline{a}_z - \underline{a}_z\|_1$, this gives

$$L_z^F(\xi) = \underline{c}_z + \underline{r}_z - \delta_z + \bar{a}_z^T\xi, \qquad U_z^F(\xi) = \overline{c}_z + \overline{r}_z + \delta_z + \bar{a}_z^T\xi.$$

SSM-Affine$_1$ intersects the intrinsic range with G_z and the interval $[\min_\xi L_z^F(\xi), \max_\xi U_z^F(\xi)]$ over $[-1,1]^s$. It selects one complete affine form per side: the intrinsic or imported lower form with the larger minimum, and the upper form with the smaller maximum. Value intervals are imported at matching graph nodes, including RMS, selectors, scan inputs, readouts, gates, and block outputs; affine forms are used where available.

H.2 Local relaxations and propagation

The resulting node ranges define the domains of the local linear relaxations. For products, lower and upper planes are convex combinations of the two corresponding McCormick facets on each parent rectangle. Products are grouped into six types (RMS, transition–state, step–input, injection, state–readout, and output–gate), each with one lower and one upper mixing weight.

Square, exponential, and inverse square root use tangent lower and secant upper planes on the intersected parent intervals. RMS uses square, mean, inverse square root, and product nodes with a denominator shared across coordinates. SiLU checks endpoint and stationary-point residual extrema for secant and nominal-derivative slopes.

A local plane $a^Tz + b$ is propagated by using each parent's lower bound for a positive coefficient and upper bound for a negative coefficient when forming a lower bound; the choices reverse for an upper bound. Exact affine maps follow the same sign rule.

Backward propagation sums coefficients from every use of a shared node before eliminating it. An affine form $b + g^T\xi$ ranges over $[b - \|g\|_1, b + \|g\|_1]$ on the input box.

The twelve mixing weights use the grid $\{0, .25, .5, .75, 1\}$. Starting at midpoints, one coordinate sweep tests the four alternatives for each weight: 49 backward evaluations in total. The reported lower bound is the larger of the forward bound and the best backward result.

H.3 Validity and numerical evaluation

Validity requires enclosed initial states and intrinsic ranges, imported bounds for the same original node and box, and local planes valid on their full parent ranges. Under these conditions, using the coefficient midpoints introduces at most δ_z error on the box. Intersection, affine-form selection, and convex combinations of McCormick facets preserve validity. Induction through the finite graph then proves signed forward and backward substitution sound, with reused-node coefficients summed before elimination.

Numerical validity requires outward enclosures for imported and intrinsic bounds, local planes, affine propagation, and final ranges, following Appendix E.3; see Appendix F.1.3 for the software environment.