How Long Is a Recurrent Circuit Faithful? Retained Memory and Fading Error

Anonymous authors
Paper under double-blind review

Abstract

A recurrent circuit must carry its own state, not receive a corrected state before each short test. Certifying persistent execution is difficult when retained units hold information indefinitely: the usual contraction test on the model–circuit comparison can fail even when their output discrepancy is small. We separate retained circuit memory from fading error. An exact finite memory transducer, combined with protocol-constrained error contraction, gives convergent all-time risk bounds and failure histories that can be replayed without requiring the retained memory to contract. We quantify the error that remains when memory summaries are coarse and give a complementary reset–stutter certificate for token-conditioned affine recurrences. The method builds on invariant-set and differential-verification principles rather than replacing them. A coupled write/hold example exercises the new premise. Six fully trained parity recurrences produce smaller circuits certified at tolerance 0.1 under a reset-bounded token grammar; five of six refreshed selections at that tolerance fail exact persistent evaluation. Separate lower-calibrated scaling tests compare adaptive suffix refinement against block-aware invariant bounds, including difficult unresolved cases. Finally, tolerance reanalysis of fixed dense nonlinear networks show that local success followed by later failure depends strongly on the chosen metric and operating point.

1 Introduction

A recurrent explanation is an executable claim about how information moves through a model. Deleting a unit once and observing little output change does not establish that a circuit can operate without that unit for the rest of a sequence. The distinction is particularly sharp when a local evaluation periodically replaces the circuit's state with a projection of the original model's state. That procedure tests short-term response from a supplied state, not the circuit's ability to preserve the information that produced it.

Causal abstraction and intervention-based circuit evaluation make this distinction meaningful (Geiger et al., 2025; Conmy et al., 2023; Hanna et al., 2024). It is also related to the classical difference between prediction error and free-run simulation error in system identification (Ljung, 1999; Ribeiro et al., 2020). We therefore ask a narrower question than whether one selection rule is generally better: *under what declared input protocol can a fixed persistent circuit be certified for every execution length?*

A direct approach constructs a differential system for the original network and its masked version, then bounds its reachable outputs. Differential verification already preserves correlations between related networks (Mohammadinejad et al., 2020; Paulsen et al., 2020). Invariant-set and switched-gain methods already combine a finite reachable sum with a stable tail (Raković et al., 2005; Athanasopoulos et al., 2017; Naghnaeian & Voulgaris, 2016). However, a contraction test on the entire comparison can exclude precisely the memory we wish to explain. A retained latch may have eigenvalue one and continuously drive omitted units. Its value is bounded and known from the input history, but it does not fade.

Our approach keeps that persistent information explicitly in the verification state and requires forgetting only in the remaining error dynamics. This extends the scope of the certificate while using established invariant-set methods. It yields three concrete contributions. **First**, an exact memory-transducer compilation and a convergent verification procedure with feasible witnesses cover coupled

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

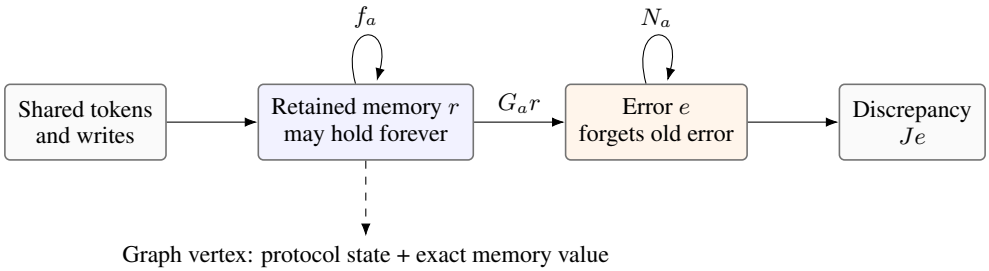


Figure 1: The memory that must not fade is kept in a finite control state. Only the output-relevant error must contract along legal paths. In the coupled control, writes set r to ± 1, both HOLD and RELAX preserve it, and $G_a r$ is nonzero. This differs from deleting a decoupled common memory before verification.

retained memory that defeats whole-comparison contraction. **Second**, a memory-resolution bound identifies what is lost when that memory is coarsened; a reset–stutter result gives a different finite certificate when informative computation is protocol-bounded but execution time is not. **Third**, experiments separate certification of learned circuits from numerical stress tests: fully trained token recurrences, generic lower-calibrated switched systems with equal-strength stability bounds, and complete tolerance sweeps of dense nonlinear controls.

2 What is being certified?

The original model and persistent circuit obey

$$\begin{aligned} x_{t+1} &= A_{a_t} x_t + B_{a_t} u_t + c_{a_t}, \qquad y_t = C x_t, \\ z_{t+1} &= P(A_{a_t} z_t + B_{a_t} u_t + c_{a_t}) + (I - P)\mu, \quad \widehat{y}_t = C z_t, \end{aligned} \tag{1}$$

where P selects retained coordinates and the ablation value μ is fixed. Both executions receive the same fresh input $u_t \in [-1, 1]^b$ and the same token or mode a_t. Initial states are $x_0 = x_c + G_0 \xi$, $\xi \in [-1, 1]^s$, and $z_0 = P x_0 + (I - P)\mu$. An input protocol is a finite directed graph; allowed token sequences are paths from its designated start. Every prefix is evaluated. The circuit receives no subsequent original-model state.

For the fixed mask, define

$$\mathcal{R}_H(P) = \sup_{\text{legal inputs}} \max_{0 \le t \le H} \|y_t - \widehat{y}_t\|_\infty, \qquad \mathcal{R}_\infty(P) = \sup_{H < \infty} \mathcal{R}_H(P). \tag{2}$$

The validity horizon at tolerance ε is $H_\varepsilon = \sup\{H : \mathcal{R}_H \le \varepsilon\}$. A bound on $\mathcal{R}_\infty$ below ε certifies $H_\varepsilon = \infty$. A failure history identifies a finite upper bound on H_ε; exact enumeration can identify the first failure. This is agreement with the original computation, not task correctness. The two can differ substantially in our learned models.

Writing $d = x - z$ gives an affine comparison. With E the retained coordinate embedding and $Q = I - P$, its compact state is (d, r) with $r = E^\top z$. In particular,

$$d^+ = A_a d + Q A_a E r + Q B_a u + Q(A_a Q \mu + c_a - \mu). \tag{3}$$

Appendix A.1 gives the complete $n + k$-state recurrence. It folds $Qz = Q\mu$ exactly. A fixed mean is kept constant rather than treated as a freely varying input. For each fixed token word, output discrepancies are affine functions of the actual initial and later inputs. Row-wise sign choices therefore give exact maxima and feasible shared-input witnesses.

Relation to reachability and gain analysis. With zero initial state and zero offsets, (2) is the worst-case peak output of a switched system under unit box inputs, hence its induced infinity gain. Minimal robust positively invariant (mRPI) approximations and constrained invariant multisets already provide convergent reachable-set bounds (Kolmanovsky & Gilbert, 1998; Raković et al., 2005; Blanchini & Miani, 2008; Athanasopoulos et al., 2017); switched-gain work directly studies this performance

Table 1: Assumptions and guarantees relative to the closest control literature. Differences in initial sets or offsets do not by themselves establish algorithmic superiority.

Method	Relevant setting	Relationship to this work
mRPI approximation (Raković et al., 2005)	Stable linear time-invariant (LTI) dynamics; bounded additive polytopic disturbances	The exact-sum plus stable-tail principle is inherited; our token graph and retained-memory compilation are extra structure.
Invariant multisets (Athanasopoulos et al., 2017)	Constrained stable switching; specified disturbance-set and strongly connected graph assumptions	After exact memory compilation, the error system is a constrained affine system. We keep actual initial transients, degenerate inputs, fixed offsets, and reachable seeds.
Switched gains (Naghnaeian & Voulgaris, 2016; Naghnaeian et al., 2018)	Induced gains; linear programming (LP) characterizations for specified classes and stable-system approximations	The centered zero-start risk is the same performance quantity. No claim that suffix search supersedes these algorithms.
This memory certificate	Finite retained-memory transducer; only legal error products contract	Noncontracting memory can drive error. Upper and lower bounds refer to the same feasible token/input histories.

quantity (Naghnaeian & Voulgaris, 2016; Naghnaeian et al., 2018). A whole-comparison finite-history bound is an output-directed specialization of these ideas. The question here is how to retain a circuit's noncontracting memory, preserve feasible histories, and state exactly which input grammar the resulting claim covers. Table 1 makes the comparison explicit.

3 KEEP PERSISTENT MEMORY; FORGET ONLY ERROR

Consider a comparison with cascade form

$$r_{t+1} = f_{a_t}(r_t), \qquad e_{t+1} = N_{a_t} e_t + G_{a_t} r_t + c_{a_t} + D_{a_t} u_t, \qquad y_t - \widehat{y}_t = J e_t. \tag{4}$$

The variable r represents retained circuit memory, not necessarily the original model's retained coordinates. The error block may be obtained by structural or exact observable reduction of (3). A general mask always gives a cascade with the circuit state upstream; the condition that its error block contracts is a meaningful restriction. The following first result only needs bounded memory.

Proposition 1 (Bounded-memory upper certificate)**.** *Suppose $r_t \in Z_{q_t}$ follows verified compact invariant sets and the legal error products are uniformly contractive in blocks. Replace $G_a r_t + c_a + D_a u_t$ by its image over $Z_{q_t} \times [-1,1]^b$. Any sound invariant-set output bound of this enlarged error system bounds $\mathcal{R}_\infty(P)$. No contraction of r is required.*

This sound outer approximation can remain loose because it loses consistency between memory values at different times. The next result retains that consistency exactly.

Lemma 2 (Legal block bound)**.** *On a finite graph, let $\beta_j = \max_{|w|=j,\ w\ legal} \|N_w\|_\infty$, $\beta_0 = 1$. If $\lambda = \beta_p < 1$, forcing is bounded by f, and the initial error norm is at most b_0, then every reachable error has norm at most*

$$R = \max_{0 \le r < p} \left(\beta_r B + f \sum_{j<r} \beta_j \right), \qquad B = \max\left\{ b_0, \frac{f \sum_{j<p} \beta_j}{1-\lambda} \right\}. \tag{5}$$

Moreover $\beta_{kp+r} \le \lambda^k \beta_r$.

The products include paths from every reachable graph vertex, but exclude forbidden words. This elementary block test is a sufficient constrained-stability certificate; multinorm and path-complete criteria can be stronger (Philippe et al., 2016a; Ahmadi et al., 2014).

Theorem 3 (Memory-conditioned all-time certificate)**.** *Suppose the reachable pairs (q_t, r_t) in (4) form a finite exact transducer and the error products satisfy Lemma 2. There is an equivalent driven error system whose graph vertices are these pairs. It has nested computable bounds*

$$\underline{r}_m \le \mathcal{R}_\infty(P) \le \overline{r}_m, \qquad \overline{r}_m - \underline{r}_m \le 2R\|J\|_\infty \beta_m, \tag{6}$$

Table 2: Memory-conditioned certificate procedure. Resource limits never imply faithfulness.

1	Compile the persistent shared-input comparison; fold fixed coordinates and check an exact observable quotient.
2	If a finite retained-memory transducer is available, move it to the protocol graph. Otherwise use the full comparison or a sound bounded-memory outer approximation.
3	Verify legal block products and a boundary set for the remaining error. Give the same products to the block-ball and block-template baselines.
4	Initialize one backward root per reachable graph vertex and one feasible seed history per vertex.
5	Compute each node's signed affine support upper and its feasible-seed lower value. Evaluate original-start histories exactly at the start vertex.
6	If a lower value exceeds ε, replay its optimizing full history and return FALSIFIED. Prune any class whose upper is at most ε.
7	Expand a remaining high-upper node through legal incoming edges, intersecting child bounds with its parent. Return CERTIFIED when none remain, otherwise UNRESOLVED at the budget.

using complete suffix depth m, exact shorter original-start histories, and one feasible error seed at each graph vertex. Every lower bound is realized or improved by a replayable original token/input history. A tolerance strictly above or below the true risk is eventually decided; a boundary equality need not be.

Why it works. At the start of a suffix we know the exact retained memory. Two histories meeting that same graph vertex can differ in their error state by at most $2R$. Their subsequent forcing is identical when the same suffix inputs are used. Only this error difference must shrink. By contrast, discarding the retained memory leaves a fresh forcing uncertainty at every future step; longer unrolling does not remove it. Proposition 5 quantifies this distinction: memory mismatch δ contributes up to $\|J\| g\delta \sum_{j<m} \beta_j$, a generally nonvanishing term, in addition to the fading boundary error.

A coupled latch. Let $r \in \{0, \pm 1\}$ be a retained write/hold memory. The omitted scalar obeys $e^+ = \alpha_a e + gr$, with $\alpha_{\mathrm{HOLD}} = 9/8$ and $\alpha_a = 1/4$ for RELAX and the two WRITE tokens. HOLD and RELAX preserve r; writes set it to ± 1. The protocol forbids consecutive HOLDs. The model reads $r + e$ and the circuit reads r. The retained memory has unit eigenvalue and drives the discrepancy, so it cannot be removed as an unobservable common state. The whole comparison fails contraction, as does the error under unrestricted switching. Yet the legal two-step error bound is $\beta_2 = 9/32 < 1$. The exact all-time risk is

$$\mathcal{R}_\infty = \frac{68}{23} g. \tag{7}$$

Alternating RELAX and HOLD at $r = 1$ approaches this value. A two-protocol-state invariant interval proves the matching upper bound. For $g = 1/32$, the memory-conditioned verification certifies tolerance 0.1 after 113 nodes and gives a failure witness at tolerance 0.05 after 37 nodes. The memory may be held indefinitely by legal RELAX/HOLD sequences.

4 VERIFYING PROTOCOL-BOUNDED COMPUTATION

The verification first constant-folds fixed ablations and removes exact unobservable comparison directions. This is classical realization reduction (Moore, 1981; Antoulas, 2005; Philippe et al., 2016b; Kiefer, 2020); it does not change the intervention basis or the mask. It then uses either exact finite memory or, when it contracts, the full comparison. Each backward suffix stores signed output coefficients, accumulated offsets, and input support. It is refined only when its upper bound exceeds the requested tolerance. An exact corner calculation supplies lower witnesses. Algorithm 2 states the decision procedure.

A different protocol mechanism can provide an all-time certificate without contraction at all. It is useful for learned token recurrences whose hidden states do not form an exact finite set.

Proposition 4 (Reset–stutter certificate)**.** *Suppose a reset restores both declared initial states and the protocol start. Suppose stutter tokens leave both states and readouts unchanged. If the protocol permits at most L informative tokens between resets, then the all-time risk equals the maximum discrepancy over informative prefixes of length at most L.*

Table 3: Fully trained parity models at tolerance 0.1. P: persistent selection; R: refreshed selection. All six P masks certify; five R masks pass their local test but fail exact persistent evaluation. Accuracy is one for both methods.

n	Seed	k_P	Risk P	k_R	Local R	Risk R	R cert.?
16	0	14	0.01	13	0.08	0.19	no
16	1	14	0.07	14	0.06	0.07	yes
16	2	14	0.02	13	0.10	0.19	no
32	0	23	0.09	22	0.09	0.18	no
32	1	25	0.06	24	0.09	0.19	no
32	2	24	0.05	23	0.08	0.13	no

This follows by deleting stutters and splitting at resets. Arbitrarily many episodes and arbitrarily long waits are covered. The informative-length cap remains part of the claim. Our trained example uses twelve binary tokens per episode; it does not claim faithfulness beyond that cap.

Scope. A finite-vocabulary token-conditioned affine recurrence fits (1) exactly: the token selects its matrix and fixed offset. General state-dependent continuous gates in Mamba or gated linear attention require extra abstraction, not an automatic application of this theorem (Gu & Dao, 2023; Yang et al., 2023; Orvieto et al., 2023). A fixed feature basis may be conjugated before masking, but exact support and sparsity are basis-dependent. Nonlinear feature maps or resample ablations need their own coupling contract (Geiger et al., 2023; Massidda et al., 2023; Zhang & Nanda, 2023). Worst-case output discrepancy is stronger than task accuracy or distributional average fidelity. The nonlinear study below reports both rather than treating them as interchangeable.

5 Experiments

All certificates use exact dyadic deployments and independent witness replay. Numerical LPs propose invariant boxes; rational residual checks verify them. Floating-point training and nonlinear evaluation are separate. Node caps and all failed or unresolved requests are retained. Reported runtimes describe these evaluations and are not complexity guarantees.

5.1 A trained token recurrence with discovered circuits

We train six dense token-conditioned affine models on running binary parity: dimensions 16 and 32, three seeds each. Both token matrices, offsets, initial state, and readout are optimized end to end; a small learned gate penalty encourages removable coordinates. No hidden-state targets, planted block, faithful mask, or target cardinality are supplied. Greedy deletion uses the same 64 trajectories and twelve positions for both selectors, differing only in state refresh every four steps versus persistent execution. At each tolerance chosen in advance, we select the smallest mask along the respective deletion path that passes its validation criterion.

The deployed coefficients are quantized to 16 fractional bits. We enumerate all 8191 informative prefixes for each distinct mask, independently replay the worst word, and use Proposition 4 to certify any number of episodes and QUERY stutters. RESET and QUERY are fixed interface operations, not learned hidden dynamics. Both retained-to-omitted and omitted-to-retained matrix couplings are nonzero in these trained models. At tolerance 0.1, all six persistently selected masks certify, retaining 14 of 16 or 23–25 of 32 units. Five of six refreshed selections pass locally but fail exact persistent evaluation. Across the four tolerance levels the counts are 20/24 and 5/24, respectively. The procedure also rejects four persistent selections at other tolerances: validation alone is not a certificate. Every selected circuit preserves parity sign accuracy over the enumerated lengths, including some with larger output discrepancy. At the same retained cardinalities selected by the persistent rule at 0.1, both deletion paths produce identical masks in all six models. The observed difference is therefore a stopping-calibration effect, not a demonstrated ranking advantage. Thus the certificate checks a stronger, explicitly chosen notion of faithfulness than task success.

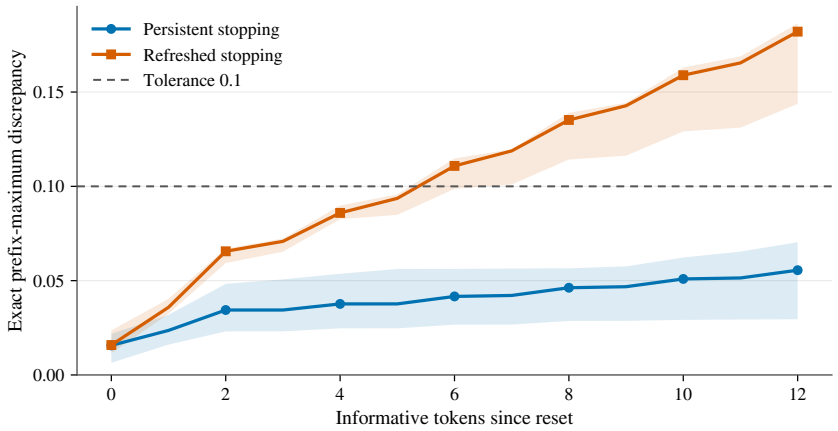


Figure 2: Exact risk versus the number of informative tokens for the six learned models at the 0.1 selection tolerance. Curves are model medians; bands are interquartile ranges. Both methods use the same trajectories and positions for discovery. The masks have different cardinalities because each selector stops at its own validation threshold; at matched cardinality their masks coincide. The protocol allows arbitrarily many QUERY stutters and RESET episodes, but at most twelve informative tokens per episode.

Table 4: Adaptive results with a 6000-node cap. Nodes and total seconds are medians over certifications only; the final column is the median upper/lower ratio over unresolved requests. Total includes feasible lower-bank construction, rank screening, stability, box construction when used, and verification. The block-only and fixed-depth comparators certify zero requests.

Switching	Margin	Tail	Cert.	Nodes	Total s	Unres. ratio
Comparable	1%	ball	0/24	–	–	7.00
Comparable	1%	box	0/24	–	–	3.71
Comparable	10%	ball	0/24	–	–	7.00
Comparable	10%	box	0/24	–	–	3.71
Slow + fast	1%	ball	7/24	2,119	2.24	1.31
Slow + fast	1%	box	16/24	1,720	2.24	1.18
Slow + fast	10%	ball	11/24	1,853	1.24	1.46
Slow + fast	10%	box	18/24	855.00	1.07	1.22

5.2 Lower-calibrated, non-involutive scaling

The numerical stress test has 48 dense signed systems with $n \in \{16, 32, 64, 128\}$, two or four modes, and three slow-rate parameters .8, .9, .95. Random nonsymmetric perturbations and an upper shift produce non-normal, non-involutive matrices; their observed spectral radii, legal block length p, β_p, and radius R are evaluated per model. There are two declared switching regimes: one slow mode with faster recovery modes, or multiple comparably slow modes. These are generated comparison-system tests, not another claimed population of learned circuits.

A fixed bank of realizable length-160 periodic input words gives a lower bound ℓ_{bank}, independently replayed. Before running any upper verification, set $\varepsilon = (1+\gamma)\ell_{\mathrm{bank}}$ for $\gamma \in \{.01, .10\}$. Success is *not* built in. When it occurs, $\ell_{\mathrm{bank}} \le \mathcal{R}_\infty \le (1+\gamma)\ell_{\mathrm{bank}}$ establishes an actual risk bracket of relative width at most γ.

Every method receives the identical legal signed products. Comparators are a block-ball bound, a rationally verified block-box invariant LP, and a fixed-depth exact-sum plus boundary approximation using the same 6000-node budget as adaptive refinement. The latter uses the mRPI principle and receives the same stability certificate. Both ball and box boundary geometries are used in the fixed and adaptive methods. Table 4 separates outcomes by regime; no quick falsifications contribute to the reported certification count. The box-refined verification certifies 16/24 slow–fast systems at the one-percent bracket and 18/24 at ten percent. The ball-refined counts are 7/24 and 11/24. All

96 comparable-slow requests remain unresolved; none is removed from the analysis. The largest successful one-percent case has $n = 128$, 2 modes, 5505 nodes and depth 176 (verification 54.80 s, full cost 55.96 s). The feasible lower values range from 0.15 to 2.46. All 48 systems have certified full observable rank: this is not scaling a hidden two-dimensional linear quotient. The comparable-slow regime is a necessary negative result: a short periodic lower witness need not locate the worst switching sequence, and branching can defeat suffix refinement. We report the remaining bracket and label these requests unresolved. The data support a conditional use case, not general superiority over invariant-set methods or exact support LPs.

5.3 Memory coupling and cancellation preprocessing

The latch of (7) produces twelve fixed-tolerance requests: nine certify and three return feasible counterexamples. Its six-state memory/protocol graph has one continuous error coordinate. A protocol-unaware product test and a whole-comparison contraction test both fail; the legal two-step test succeeds. This demonstrates coupled retained memory without whole-comparison contraction.

The delayed-cancellation controls use an exact observable quotient before checking. Two cancelling chains of total dimension $2(L + 1)$ reduce to a difference chain of dimension $L + 1$, for delays $L = 2, 4, 8, 12, 24, 48$. A quotient plus verified invariant-box geometry certifies all twelve cases; a scalar ball on the same quotient can still fail at the node cap. At the longest delay the box verification uses 91 or 549 nodes, rather than one, because finite-grid outward padding is appreciable relative to the tiny tolerance. An exact direction-closure LP also solves these controls. Failure of a scalar boundary approximation therefore does not establish inherent difficulty.

5.4 Dense nonlinear models: operating point matters

We re-evaluate 240 fixed masks from thirty trained 16-unit recurrent neural networks (RNNs) with hyperbolic tangent (tanh) activations on delayed query, reset memory, and two-register tasks. The masks retain twelve or fourteen units and were selected by initial-prefix greedy, refreshed-window greedy, persistent greedy, or integrated-gradient gating. This tolerance reanalysis reuses the same models and masks without retraining or reselection; cells sharing a trained model are correlated. At tolerance 0.1, none of the refreshed or persistent selections passes its own local test. At tolerance 0.5, 40/60 refreshed masks pass locally and 26 of those fail by step 40; persistent selection has 29/60 local passes and twelve later failures. The analysis examines all ten tolerance levels rather than selecting only a favorable threshold.

Using a different criterion—at least 95% query accuracy locally—50/60 refreshed masks pass and fifteen drop below that accuracy later. Persistent selection has 37 local passes and three later task failures. Empty-query prefix windows are excluded, not counted as perfect performance. These counts do not show that persistent selection always performs better: the paired same-trajectory ablation wins 30/60 cells, ties in eighteen, and loses in twelve. Integrated gradients (IG) gating beats a size-matched random seed median in only 30/60 cells.

6 Discussion and limitations

The main lesson is structural: a certificate should preserve memory that remains relevant. Whole-comparison contraction is sufficient but can exclude a useful recurrent explanation; exact memory conditioning or reset semantics can replace it in restricted regimes. The stable-tail calculation itself comes from control theory. Exact finite memory may be expensive or unavailable, coarse memory can leave a nonvanishing error floor, and bounded-reset certification does not extrapolate beyond the declared informative-length limit. General nonlinear RNNs are evaluated empirically, not universally certified.

Lower-calibrated tolerances avoid the predetermined success of upper calibration (Proposition 6), and all methods receive the same stability products. Cancellation controls show why reduction and boundary geometry matter before attributing failure to temporal complexity. The certificate returns feasible failure histories or UNRESOLVED when its memory representation or computation is insufficient.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and polish writing. Additionally, we used generative AI tools for editing code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Amir Ali Ahmadi, Raphaël M. Jungers, Pablo A. Parrilo, and Mardavij Roozbehani. Joint spectral radius and path-complete graph lyapunov functions. *SIAM Journal on Control and Optimization*, 52(1):687–717, 2014. doi: 10.1137/110855272.

Athanasios C. Antoulas. Approximation of large-scale dynamical systems. SIAM, 2005. URL https://epubs.siam.org/doi/book/10.1137/1.9780898718713.

Nikolaos Athanasopoulos, Konstantinos Smpoukis, and Raphaël M. Jungers. Invariance in constrained switching. *arXiv preprint arXiv:1702.00598*, 2017. URL https://arxiv.org/abs/1702.00598.

Borja Balle, Prakash Panangaden, and Doina Precup. A canonical form for weighted automata and applications to approximate minimization. In *ACM/IEEE Symposium on Logic in Computer Science*, pp. 701–712, 2015. doi: 10.1109/LICS.2015.70.

Borja Balle, Prakash Panangaden, and Doina Precup. Singular value automata and approximate minimization. *Mathematical Structures in Computer Science*, 29(9):1444–1478, 2019. doi: 10.1017/S0960129519000094.

Aishwarya Balwani. Mechinterp for recurrent computation: Time-resolved circuit discovery in RNNs. Mechanistic Interpretability Workshop, virtual poster listing, 2026. URL https://mechinterpworkshop.com/posters/virtual/.

Franco Blanchini and Stefano Miani. *Set-Theoretic Methods in Control*. Birkhäuser, 2008. doi: 10.1007/978-0-8176-4606-6.

Edmund Clarke, Orna Grumberg, Somesh Jha, Yuan Lu, and Helmut Veith. Counterexample-guided abstraction refinement. Computer Aided Verification, LNCS 1855, pp. 154–169, 2000. URL https://link.springer.com/chapter/10.1007/10722167_15.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. Advances in Neural Information Processing Systems, 2023. URL https://arxiv.org/abs/2304.14997.

Patrick Cousot and Radhia Cousot. Abstract interpretation: A unified lattice model for static analysis of programs by construction or approximation of fixpoints. Proceedings of the ACM Symposium on Principles of Programming Languages, pp. 238–252, 1977. URL https://www.di.ens.fr/~cousot/COUSOTpapers/POPL77.shtml.

Róbert Csordás, Christopher Potts, Christopher D. Manning, and Atticus Geiger. Recurrent neural networks learn to store and generate sequences using non-linear representations. *arXiv preprint arXiv:2408.10920*, 2024. URL https://arxiv.org/abs/2408.10920.

Munther A. Dahleh and J. Boyd Pearson. ℓ_1-optimal feedback controllers for MIMO discrete-time systems. *IEEE Transactions on Automatic Control*, 32(4):314–322, 1987. doi: 10.1109/TAC.1987.1104603. URL https://ieeexplore.ieee.org/document/1104603/.

Jean-Michel Dion, Christian Commault, and Jacob van der Woude. Generic properties and control of linear structured systems: a survey. *Automatica*, 39(7):1125–1144, 2003. doi: 10.1016/S0005-1098(03)00104-3.

Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy Mann, and Pushmeet Kohli. A dual approach to scalable verification of deep networks. arXiv preprint arXiv:1803.06567, 2018. URL https://arxiv.org/abs/1803.06567.

Jeffrey L. Elman. Finding structure in time. Cognitive Science, 14(2):179–211, 1990. URL https://doi.org/10.1207/s15516709cog1402_1.

Danielle Ensign and Adrià Garriga-Alonso. Investigating the indirect object identification circuit in Mamba. *arXiv preprint arXiv:2407.14008*, 2024. URL https://arxiv.org/abs/2407.14008.

Lorenzo Farina and Sergio Rinaldi. *Positive Linear Systems: Theory and Applications*. Wiley, 2000. doi: 10.1002/9781118033029.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. *arXiv preprint arXiv:2303.02536*, 2023. URL https://arxiv.org/abs/2303.02536.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

Antoine Girard and George J. Pappas. Approximate bisimulation: A bridge between computer science and control theory. *European Journal of Control*, 17(5–6):568–578, 2011. doi: 10.3166/EJC.17.568-578.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. URL https://arxiv.org/abs/2312.00752.

Nicola Guglielmi and Vladimir Yu. Protasov. Exact computation of joint spectral characteristics of linear operators. *Foundations of Computational Mathematics*, 13(1):37–97, 2013. doi: 10.1007/s10208-012-9121-0.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. Conference on Language Modeling, 2024. URL https://arxiv.org/abs/2403.17806.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, 1997. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Richard M. Karp. Reducibility among combinatorial problems. Complexity of Computer Computations, pp. 85–103, 1972. URL https://link.springer.com/chapter/10.1007/978-1-4684-2001-2_9.

Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. Computer Aided Verification, 2017. URL https://arxiv.org/abs/1702.01135.

Stefan Kiefer. Notes on equivalence and minimization of weighted automata. arXiv preprint arXiv:2009.01217, 2020. URL https://arxiv.org/abs/2009.01217.

Ilya Kolmanovsky and Elmer G. Gilbert. Theory and computation of disturbance invariant sets for discrete-time linear systems. *Mathematical Problems in Engineering*, 4(4):317–367, 1998. doi: 10.1155/S1024123X98000866.

Ching-Tai Lin. Structural controllability. *IEEE Transactions on Automatic Control*, 19(3):201–208, 1974. doi: 10.1109/TAC.1974.1100557.

Zexiang Liu, Tzanis Anevlavis, Necmiye Ozay, and Paulo Tabuada. Automaton-based implicit controlled invariant set computation for discrete-time linear systems. *arXiv preprint arXiv:2109.12378*, 2021. URL https://arxiv.org/abs/2109.12378.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, second edition, 1999. URL https://www.control.isy.liu.se/books/sysid/.

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. Proceedings of Machine Learning Research, volume 213, pp. 68–87, 2023. URL https://proceedings.mlr.press/v213/massidda23a.html.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer circuit faithfulness metrics are not robust. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2407.08734.

Sara Mohammadinejad, Brandon Paulsen, Chao Wang, and Jyotirmoy V. Deshmukh. DiffRNN: Differential verification of recurrent neural networks. *arXiv preprint arXiv:2007.10135*, 2020. URL https://arxiv.org/abs/2007.10135.

Bruce C. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. IEEE Transactions on Automatic Control, 26(1):17–32, 1981. URL https://doi.org/10.1109/TAC.1981.1102568.

Mohammad Naghnaeian and Petros G. Voulgaris. Characterization and optimization of ℓ_∞ gains of linear switched systems. *IEEE Transactions on Automatic Control*, 61(8):2203–2218, 2016. doi: 10.1109/TAC.2015.2494369.

Mohammad Naghnaeian, Petros G. Voulgaris, and Geir E. Dullerud. ℓ_p analysis and synthesis of linear switched systems: A unified input-output and state-space approach. *SIAM Journal on Control and Optimization*, 56(2):1181–1205, 2018. doi: 10.1137/16M1104299.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. *arXiv preprint arXiv:2303.06349*, 2023. URL https://arxiv.org/abs/2303.06349.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. Proceedings of Machine Learning Research, volume 28(3), pp. 1310–1318, 2013. URL https://proceedings.mlr.press/v28/pascanu13.html.

Gonçalo Paulo, Thomas Marshall, and Nora Belrose. Does transformer interpretability transfer to RNNs? *arXiv preprint arXiv:2404.05971*, 2024. URL https://arxiv.org/abs/2404.05971.

Brandon Paulsen, Jingbo Wang, and Chao Wang. ReluDiff: Differential verification of deep neural networks. In *International Conference on Software Engineering*, pp. 714–726, 2020. doi: 10.1145/3377811.3380337.

Matthew Philippe, Ray Essick, Geir E. Dullerud, and Raphaël M. Jungers. Stability of discrete-time switching systems with constrained switching sequences. *Automatica*, 72:242–250, 2016a. doi: 10.1016/j.automatica.2016.05.015. URL https://arxiv.org/abs/1503.06984.

Matthew Philippe, Ray Essick, Geir E. Dullerud, and Raphaël M. Jungers. Extremal storage functions and minimal realizations of discrete-time linear switching systems. *arXiv preprint arXiv:1609.05779*, 2016b. URL https://arxiv.org/abs/1609.05779.

Saša V. Raković, Eric C. Kerrigan, Konstantinos I. Kouramas, and David Q. Mayne. Invariant approximations of the minimal robust positively invariant set. *IEEE Transactions on Automatic Control*, 50(3):406–410, 2005. doi: 10.1109/TAC.2005.843854. URL https://ieeexplore.ieee.org/document/1406138/.

Kurt J. Reinschke. *Multivariable Control: A Graph-Theoretic Approach*, volume 108 of *Lecture Notes in Control and Information Sciences*. Springer, 1988. doi: 10.1007/BFb0051585.

Antônio H. Ribeiro, Koen Tiels, Jack Umenberger, Thomas B. Schön, and Luis A. Aguirre. On the smoothness of nonlinear system identification. *Automatica*, 121:109158, 2020. doi: 10.1016/j.automatica.2020.109158.

Arnab Sen Sharma, David Atkinson, and David Bau. Locating and editing factual associations in Mamba. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.03646.

Claudia Shi, Nicolas Beltran-Velez, Achille Nazaret, Carolina Zheng, Adrià Garriga-Alonso, Andrew Jesson, Maggie Makar, and David M. Blei. Hypothesis testing the circuit hypothesis in LLMs. arXiv preprint arXiv:2410.13032, 2024. URL https://arxiv.org/abs/2410.13032.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. Proceedings of Machine Learning Research, volume 70, pp. 3319–3328, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

David Sussillo and Omri Barak. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. Neural Computation, 25(3):626–649, 2013. URL https://doi.org/10.1162/NECO_a_00409.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. NeurIPS Workshop on Attributing Model Behavior at Scale, 2023. URL https://arxiv.org/abs/2310.10348.

John N. Tsitsiklis and Vincent D. Blondel. The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate. *Mathematics of Control, Signals, and Systems*, 10:31–40, 1997. doi: 10.1007/BF01219774.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023. URL https://arxiv.org/abs/2312.06635.

Majid Zamani, Alessandro Abate, and Antoine Girard. Symbolic models for stochastic switched systems: A discretization and a discretization-free approach. *Automatica*, 55:183–196, 2015. doi: 10.1016/j.automatica.2015.03.004. URL https://arxiv.org/abs/1407.2730.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. arXiv preprint arXiv:2309.16042, 2023. URL https://arxiv.org/abs/2309.16042.

A PROOFS AND EXACT IMPLEMENTATION CONTRACTS

A.1 SHARED-INPUT COMPARISON AND CONSTANT FOLDING

Let E contain the retained coordinate vectors, $P = EE^\top$, and $Q = I - P$. The model receives the same fresh input $u_t \in [-1, 1]^b$ as its persistent circuit:

$$x^+ = A_a x + B_a u + c_a,$$
$$z^+ = P(A_a z + B_a u + c_a) + Q\mu, \qquad z_0 = Px_0 + Q\mu.$$

Here μ is fixed. In particular it is not a free coordinate ranging over a box. Define $d = x - z$ and $r = E^\top z$, so $z = Er + Q\mu$. Substitution gives

$$\begin{bmatrix} d^+ \\ r^+ \end{bmatrix} = \begin{bmatrix} A_a & QA_aE \\ 0 & E^\top A_a E \end{bmatrix} \begin{bmatrix} d \\ r \end{bmatrix} + \begin{bmatrix} QB_a \\ E^\top B_a \end{bmatrix} u + \begin{bmatrix} Q(A_aQ\mu + c_a - \mu) \\ E^\top(A_aQ\mu + c_a) \end{bmatrix}. \tag{8}$$

The output difference is $[C, 0](d, r)$, the initial center is $[Q(x_c - \mu), E^\top x_c]^\top$, and the initial generator is $[QG_0, E^\top G_0]^\top$ when $x_0 = x_c + G_0\xi$. This proves equivalence of the $n + k$-dimensional compact comparison to both original executions. It also proves that mean ablation is affine and that omitted circuit coordinates may be constant-folded before any reachability computation. No copy of an input is independently optimized in the two executions.

For a fixed word w, affine unrolling gives $v_w + \sum_j T_{wj} u_j$, including the actual initial-box variables. Thus its exact maximum infinity discrepancy is

$$\max_i \left(|v_{w,i}| + \sum_j \|(T_{wj})_{i,:}\|_1 \right). \tag{9}$$

Choose the maximizing output row i, choose its sign from $v_{w,i}$ (either sign when zero), and align each independent box coordinate with the corresponding signed coefficient. This realizes the value. A finite token input has a fixed offset and contributes no box coordinate. Empty input boxes cause no problem. The implementation independently reconstructs the original and masked states using this word and corner.

A.2 PROTOCOL-AWARE BLOCK RADIUS

Proof of Lemma 2. For the compiled error system let β_j be the maximum infinity norm of a length-j product over legal paths from any reachable graph vertex, with $\beta_0 = 1$. All contiguous pieces of a legal path are legal. Hence $\beta_{i+j} \le \beta_i \beta_j$ and $\beta_{kp+r} \le \lambda^k \beta_r$ for $\lambda = \beta_p < 1$, $0 \le r < p$. Let every affine forcing vector have norm at most f and let initial errors have norm at most b_0. Over each complete block,

$$\|e_{(k+1)p}\|_\infty \le \lambda \|e_{kp}\|_\infty + f \sum_{j=0}^{p-1} \beta_j.$$

Consequently $B = \max\{b_0, f \sum_{j<p} \beta_j/(1-\lambda)\}$ bounds all block boundaries. At phase r the bound is $\beta_r B + f \sum_{j<r} \beta_j$. Their maximum is the stated R and includes every initial transient. This proof uses neither strong connectedness nor an interior point in a disturbance set. Products not permitted by the graph are never needed. □

A failed finite-p test is not a proof of instability. Weighted norms, multinorms, and path-complete Lyapunov functions can supply stronger product bounds (Philippe et al., 2016a; Ahmadi et al., 2014); the implementation reports failure of its own sufficient test. We do not call a general semidefinite programming (SDP) package or claim its performance.

A.3 A BOUNDED-MEMORY CASCADE WITHOUT WHOLE-STATE CONTRACTION

Proof of Proposition 1. Assume $r_0 \in Z_{q_0}$ and $f_a(Z_q) \subseteq Z_{q'}$ on every protocol edge. Induction places each retained circuit state in its declared set. Therefore

$$F_{q,a} = \{G_a r + c_a + D_a u : r \in Z_q,\ u \in [-1, 1]^b\}$$

contains the true forcing of the error system. Propagating $e^+ = N_a e + v$, $v \in F_{q,a}$, is an outer approximation. A forcing bound $f = \max_{q,a} \sup_{v \in F_{q,a}} \|v\|_\infty$ and Lemma 2 give an all-time radius for e, irrespective of contraction of r. Every support upper bound for this enlarged error system is sound for the original pair. □

This proposition is an input-to-state/invariant-set argument. It alone does not give tightness: replacing a persistent memory by a fresh independent disturbance can create impossible histories. For example, $r_0 = 0$, $r^+ = r$, $e_0 = 0$, and $e^+ = \alpha e + (1-\alpha)r$, $0 < \alpha < 1$, have zero true error. Supplying the coarse invariant $Z = [-1, 1]$ permits $r = 1$ at every step, producing supremum error one. No longer suffix removes that modeling gap. This distinguishes sound upper bounds from convergent two-sided certificates.

A.4 Finite-memory compilation and convergence

Proof of Theorem 3. Build the graph of reachable pairs (q, r) by exact transitions from (q_0, r_0). On its edge $((q, r), a, (q', f_a(r)))$, define the error transition

$$e^+ = N_a e + (G_a r + c_a) + D_a u.$$

Induction on the original token word proves a path correspondence in both directions: every original execution yields that graph path and error trajectory; every graph path from the initial vertex yields the same retained memory values in the original cascade. The box variables u remain independent exactly where they were originally independent. This proves equality of all-time risks, not just a simulation upper bound.

Choose one actually reachable error seed s_v at every graph vertex v, together with its witnessing prefix. For example, traverse a shortest path using center inputs and the initial center. All seeds satisfy $\|s_v\|_\infty \le R$. For a legal suffix w of length m from v, write its output as

$$T_w e + g_w(\mathbf{u}), \qquad T_w = J N_w.$$

Let $a(v, w) = \sup_{\mathbf{u}} \|T_w s_v + g_w(\mathbf{u})\|_\infty$. Its maximizing suffix inputs appended to the seed prefix form a valid full execution. Hence $a(v, w) \le \mathcal{R}_\infty$. Every reachable boundary error e at the same vertex obeys

$$\|T_w e + g_w(\mathbf{u})\|_\infty \le a(v, w) + (R + \|s_v\|_\infty)\|T_w\|_\infty \le a(v, w) + 2R\|J\|_\infty \beta_m.$$

Take the maximum over all such suffixes. Also include exact outputs for original-start words shorter than m, because these have no length-m suffix. Their maximum b_m is feasible. Define

$$\ell_m = \max\{b_m, \max_{v,w} a(v, w)\},$$
$$u_m = \max\{b_m, \max_{v,w} [a(v, w) + (R + \|s_v\|_\infty)\|T_w\|_\infty]\}.$$

Then $\ell_m \le \mathcal{R}_\infty \le u_m$ and $u_m - \ell_m \le 2R\|J\|_\infty \beta_m$. Running maxima of lower bounds and running minima of upper bounds retain these inequalities and make the reported intervals nested. Since $\beta_m \le \lambda^{\lfloor m/p \rfloor} \beta_{m \bmod p}$, the gap vanishes. Strict separation between the risk and the tolerance therefore yields a finite decision at some complete depth. Equality at the boundary need not be decided.

A direct boundary-set upper bound, $\sup_{e \in \mathcal{E}_v, \mathbf{u}} \|T_w e + g_w(\mathbf{u})\|$, is also valid whenever $\mathcal{E}_v$ contains every reachable boundary error and the seed. The scalar ball and the verified box used in the code are such sets. For a box $|e| \le r_v$, this support is evaluated row by row as $|T_w| r_v$ plus the exact affine suffix contribution. This can tighten the computed bound without changing feasibility of the lower witnesses. If the box has a larger infinity radius than R, its separate convergence constant is $(R + \|r_v\|_\infty)\|J\|_\infty \beta_m$, or one may intersect it with the original scalar bound; the original $2R$ statement is not automatically the box constant. □

The graph contains exact memory values, not rounded representatives. Its reachable orbit may be large or infinite. In that event the finite compiler returns a resource-limit result. The theorem can also use an externally supplied exact finite transducer, but the present implementation discovers a deterministic orbit from one initial retained state. Independent continuous inputs to the retained memory generally destroy finiteness; Proposition 1, or a separate verified finite abstraction, must then be used.

Depth and the tolerance margin. For a strict margin $\Delta = |\varepsilon - \mathcal{R}_\infty| > 0$, complete depth $m = kp$ decides once

$$2R\|J\|_\infty \lambda^k < \Delta.$$

When $R\|J\|_\infty > 0$, it suffices to take any nonnegative integer $k > \log(2R\|J\|_\infty/\Delta)/\log(1/\lambda)$. Thus sufficient depth grows logarithmically in inverse margin and approximately inversely in $1 - \lambda$ near one. A branching tree can still require exponentially many nodes in that depth. This rate follows from the geometric remainder. The adaptive method can prune many histories in an asymmetric slow/recovery protocol; comparably slow modes can retain a large frontier, as the scaling data show.

What approximate memory costs. The following perturbation calculation gives a quantitative version of the preceding limitation.

Proposition 5 (Memory-resolution floor)**.** *Consider two same-token cascades over a suffix of length m. Suppose their initial errors differ by at most $2R$, and their retained memories differ by at most δ at every suffix position. If $\|G_a\|_\infty \le g$ and all other input variables agree, then*

$$\|J(e_m - \widetilde{e}_m)\|_\infty \le \|J\|_\infty \left(2R\beta_m + g\delta \sum_{j=0}^{m-1} \beta_j \right). \tag{10}$$

Thus a path-consistent coarse-memory construction with feasible representatives inherits this additional error term. Contraction of the error alone cannot force it to zero. For a scalar held memory the second term is attained in the limit.

Proof. Subtract the two error recurrences. Unrolling the difference gives an initial term $N_w(e_0 - \widetilde{e}_0)$ and one term per memory difference, each multiplied by the corresponding suffix product. Apply the norm bounds and sum. Since $\sum_{j\ge 0} \beta_j \le \sum_{j<p} \beta_j/(1-\lambda)$, the limiting upper bound can remain positive. For $N = \alpha$, $G = g$, $J = 1$, equal initial errors, and constant memories separated by δ, the difference is $g\delta(1-\alpha^m)/(1-\alpha)$, exactly the forcing term of (10). □

An arbitrary rounded-state graph need not have path-consistent feasible representatives. Equation (10) does not authorize using such a graph for lower bounds. This is why the current two-sided implementation uses exact finite memories.

A.5 Adaptive coverage and tolerance calibration

A backward node (v, w) represents all original histories that end in suffix w starting at v. Root nodes cover all possible ending vertices. A node's exact affine suffix coefficients are updated by prepending each legal incoming edge. Its histories partition into these longer suffix classes plus the original-start history when $v = v_0$; the latter is evaluated exactly before removing the node. Upper bounds of children are intersected with the parent's upper bound. This matters when a block-invariant ball is not one-step invariant.

Maintain the largest feasible seed/original-start value as the lower bound. Retain the upper bound of every safely pruned class and every live class. Their maximum bounds every finite execution. A value above the tolerance is re-evaluated by the independent original-start box oracle; its maximizing word and corner are then replayed. An empty live frontier proves an all-time certificate. A resource limit with live classes gives UNRESOLVED. These are standard branch-and-bound coverage arguments, specialized to exact shared-input witnesses (Cousot & Cousot, 1977; Clarke et al., 2000; Paulsen et al., 2020; Mohammadinejad et al., 2020).

Proposition 6 (Upper-calibrated success is predetermined)**.** *Fix a radius, seeds, and complete depth-m seed-centered upper bound u_m. If $\varepsilon > u_m$, then a backward verification using the same radius certifies after at most the complete depth-m expansion, provided the budgets allow it.*

Proof. For each output row, the triangle inequality gives $|g|+R\|T\|_1 \le |g+Ts_v|+(R+\|s_v\|_\infty)\|T\|_1$. Thus the direct node bound is at most the corresponding seed-centered bound. Every depth-m node is below ε, and every original-start value is also at most u_m. Removing all such nodes finishes the verification. Parent-bound intersections cannot invalidate this implication. □

Thus a tolerance chosen above the method's own finite-depth upper bound measures pruning and overhead rather than an independently challenging certification success rate. The generic scaling comparison in Section 5.2 fixes tolerances from feasible lower witnesses, so its success is not predetermined by this proposition.

A.6 Verified block-template comparator

For each legal length-p word let the exact composed transition be $e^+ = M_w e + c_w + G_w v$, where v stacks the actual box inputs. A centered box $|e| \le r$ contains the initial set and is invariant at block boundaries when

$$r \ge r_0, \qquad r \ge |M_w| r + |c_w| + |G_w| \mathbf{1} \quad \text{for every legal } w. \tag{11}$$

The implementation minimizes $\mathbf{1}^\top r$ using these inequalities in a numerical LP only to obtain a proposal. It then quantizes that proposal, evaluates every inequality exactly, and lets η be the largest positive residual, including $r_0 - r$. With $\lambda = \max_w \|M_w\|_\infty < 1$, add $\eta/(1-\lambda)$ (rounded upward) to every coordinate. Since $|M_w|\mathbf{1} \le \lambda \mathbf{1}$, this repairs all block inequalities; it also repairs the initial inclusion. The code checks all repaired inequalities explicitly.

Every intermediate phase is bounded by exact signed prefix propagation from this block box. The componentwise maximum of the resulting phase boxes is a global boundary box usable at any suffix cut. The direct output bound instead propagates JM_v before taking absolute values, and can be tighter than applying $|J|$ to that global box. The scalar comparator uses the identical p, β_p, R but replaces the template by an infinity ball. Both are standard set-theoretic outer approximations (Kolmanovsky & Gilbert, 1998; Raković et al., 2005; Blanchini & Miani, 2008); only the suffix refinement is ablated between the fixed-depth and adaptive runs.

A.7 Exact observability reduction

Let the row space $\mathcal{V}$ be generated by rows of JM_w over all finite words. Compute a row basis T. Since the space is right-invariant under each M_a, there are matrices $\widetilde{M}_a$ with $TM_a = \widetilde{M}_a T$, and $J = \widetilde{J}T$. Set $\widetilde{b}_a = Tb_a$, $\widetilde{D}_a = TD_a$, and transform the initial set likewise. Induction yields $\widetilde{e}_t = Te_t$ and identical outputs for every legal word and input history. Using all words rather than only protocol words can retain unnecessary dimensions but cannot delete a needed one.

This is classical observable realization reduction (Moore, 1981; Antoulas, 2005; Philippe et al., 2016b; Kiefer, 2020). It is applied to the comparison only: the candidate mask remains in its original intervention basis. Exact row elimination detects cancellation between paths which structural pruning cannot see. We check the defining identities before using a quotient. A finite-field rank screen can certify full rational rank because dyadic denominators are invertible modulo the chosen odd prime and a nonzero modular minor is nonzero over the rationals. Deficient modular rank never licenses deletion. The current quotient backend falls back to the unreduced system when the exact quotient has non-dyadic entries; no rounding is used to claim equivalence.

A.8 Reset and stutter reduction

Proof of Proposition 4. Delete every stutter token from a legal execution. Both states and their outputs at that position are unchanged, so no discrepancy value disappears except repeated copies. A reset maps both executions to their declared initial states and restores the initial protocol state. Split the remaining execution at resets. Each segment is a word of at most L informative tokens from the original start. Every output in every segment therefore occurs in the finite enumeration of informative prefixes through L. Conversely, each such prefix is a legal execution of the full protocol. The two suprema coincide. A finite violating prefix is already a feasible full witness. If all prefixes pass, arbitrarily many resets and stutters preserve the bound. □

In the learned experiment, `QUERY` is the identity on the model and on the retained circuit coordinates, and `RESET` assigns x_0 and Px_0 respectively. The protocol permits at most twelve informative binary tokens between resets, with any number of queries. This is an all-time certificate for that grammar; it is not a guarantee for thirteen or more informative tokens in one episode. No contraction, sparse topology, or exact finite orbit of the learned hidden state is assumed. The verification uses $2^{13} - 1 = 8191$ exact prefixes per distinct mask.

B Related work and structural results

Circuit semantics and bases. Intervention-based circuit discovery and evaluation includes automated pruning, edge attribution, integrated gradients, and explicit critiques of faithfulness metrics (Conmy et al., 2023; Syed et al., 2023; Sundararajan et al., 2017; Miller et al., 2024; Hanna et al., 2024; Zhang & Nanda, 2023; Shi et al., 2024). Causal abstraction and learned alignments can change the basis in which a mechanism is described (Geiger et al., 2025; 2023; Massidda et al., 2023). Our mask is fixed before verification; applying an observable quotient to its comparison is not an intervention in a newly chosen basis. This keeps the scientific claim unchanged when the verification representation changes.

Recurrent mechanisms. Time-resolved RNN circuit analyses (Balwani, 2026), interpretability transfer from transformers to RNNs (Paulo et al., 2024), Mamba tracing (Ensign & Garriga-Alonso, 2024; Sharma et al., 2024), and nonlinear RNN representations (Csordás et al., 2024) motivate studying temporal execution rather than only feedforward masks. Modern recurrent architectures include input-dependent state-space and gated linear recurrences (Gu & Dao, 2023; Orvieto et al., 2023; Yang et al., 2023); a finite-token affine subclass is a useful controlled bridge, not a theorem for those complete architectures. Classical recurrent learning and dynamics also establish why preserved state matters (Elman, 1990; Hochreiter & Schmidhuber, 1997; Pascanu et al., 2013; Sussillo & Barak, 2013).

Verification and model reduction. Abstract interpretation and counterexample-guided refinement are established principles (Cousot & Cousot, 1977; Clarke et al., 2000). Neural verification and differential verification supply related correlation-preserving bounds (Katz et al., 2017; Dvijotham et al., 2018; Paulsen et al., 2020; Mohammadinejad et al., 2020). The shared-input affine recurrence is not claimed as a new algebraic identity. Exact and approximate reduction have long been studied in control and weighted automata (Moore, 1981; Antoulas, 2005; Kiefer, 2020; Balle et al., 2015; 2019); approximate simulation and bisimulation supply further semantic comparisons (Girard & Pappas, 2011; Zamani et al., 2015). We preserve a prescribed mask and its shared inputs, rather than optimizing an unrestricted lower-dimensional realization.

Control-theoretic assumptions. Athanasopoulos et al. (2017) study automaton-constrained additive disturbances, with assumptions on compact convex disturbance sets containing the origin in their interiors, stability, and strongly connected graphs stated in Section 2.1. Their Theorem 1 approximates the minimal invariant multiset by finite reachability plus a geometric remainder. That is the closest predecessor to our fading-history argument. Our scalar proof accepts reachable graphs with transients, deterministic offsets and rank-deficient inputs, and constructs seeds from the specified initial set. These conveniences do not imply a fundamentally new convergence principle. The main change is to put persistent retained memory into the finite graph before requiring contraction. Once compiled, established invariant-multiset algorithms are also applicable to the resulting error system. Finite controller automata have also been used to construct implicit invariant sets (Liu et al., 2021). We do not claim that augmenting a dynamics graph with finite memory is itself new; the application here preserves a prescribed circuit, its legal shared-input histories, and the distinction between bounded-memory upper bounds and feasible two-sided certificates.

Peak-to-peak and switched infinity gains are the corresponding input-output perspective (Dahleh & Pearson, 1987; Naghnaeian & Voulgaris, 2016; Naghnaeian et al., 2018). Extremal storage functions and minimal graph-dependent realizations (Philippe et al., 2016b) are especially relevant to reduction before bounding. Invariant polytopes, multinorms and path-complete Lyapunov functions can improve our simple sufficient norm tests (Guglielmi & Protasov, 2013; Philippe et al., 2016a; Ahmadi et al., 2014). We compare to a block-template invariant construction but do not claim an exhaustive benchmark against these more general methods.

Structural support. For nonnegative homogeneous recurrent matrices with zero ablation, exact fidelity through H requires retaining every coordinate on a legal source-to-readout path of length at most H. Path expansion proves sufficiency; positivity and a source basis vector prove necessity. This is a horizon- and protocol-restricted reachable/observable support characterization, closely related to classical structured and positive systems (Lin, 1974; Reinschke, 1988; Dion et al., 2003; Farina & Rinaldi, 2000). It is useful preprocessing, not the principal novelty claim. A delay line shows why

initial distinguishability alone does not account for the relay coordinates needed by a persistent mask, but that elementary construction is not a substitute for the coupled-memory or learned-certificate evidence.

Fixed-mask temporal hardness. The following background result describes worst-case verification cost. Here NP denotes nondeterministic polynomial time.

Proposition 7. *For rational nonnegative switched systems, deciding whether a prescribed mask exceeds a rational tolerance within a unary-encoded horizon is NP-complete, even with two modes and a mask deleting only one coordinate.*

Proof. Membership follows by guessing the finite mode word and evaluating its exact rational box-optimal discrepancy in polynomial time. To reduce Boolean satisfiability (SAT) with v variables and m clauses, create, for every clause and each time $t < v$, two coordinates recording whether the clause has already been satisfied. Initially the unsatisfied coordinate of each clause equals one. Mode zero or one at time t routes that mass to the next layer's appropriate satisfaction flag using the assignment of variable $t + 1$. At the final transition, each satisfied clause routes its mass to one shared output coordinate. All matrices are nonnegative; earlier readouts vanish. The readout at time v is the number of satisfied clauses. Masking only the final output coordinate makes the circuit's readout zero, since that coordinate is clamped after every transition. Set $\varepsilon = m - 1/2$ and $H = v$. A violation exists exactly when the formula is satisfiable. The construction uses $2mv + 1$ coordinates and two modes. Scaling both mode matrices by a sufficiently small dyadic number makes their norms less than one while scaling the tolerance by its vth power, which has polynomial bit length. Thus contraction does not eliminate all small-margin hardness. □

This direct reduction complements, rather than replaces, the broader complexity theory of matrix products and switching systems (Tsitsiklis & Blondel, 1997; Karp, 1972). It does not establish that each observed unresolved case is intrinsically hard; the cancellation quotient experiment demonstrates the opposite.