GREENCERT: SIGNED-RESPONSE VERIFICATION OF NEURAL TRAINING TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce GreenCert, a matrix-free verifier for certifying neural-network training trajectories over thousands of optimizer updates from a fixed checkpoint. GreenCert propagates the signed response to a proposed path's one-step defects, preserving cancellation before bounding the remaining nonlinear error. We derive a self-consistency condition guaranteeing an enclosure of the exact-real optimizer trajectory. Output-margin bounds then certify persistent-event times, later reversals, or the absence of a target within the verified window on a fixed evaluation set. When the required conditions cannot be established, GreenCert abstains. The verifier uses explicit derivative and arithmetic-error bounds, together with an ideal-Gaussian assumption for randomized operator estimates. On classification and modular-arithmetic tasks, the signed construction closes in cases where its matched unsigned counterpart fails. In the tested small-network long-window studies, GreenCert completes full-trajectory guarantees within budgets exhausted by the sequential comparators. The cost advantage depends on the problem and horizon: direct continuation remains cheaper on the larger digits task through 1,024 updates. In a separate short-window experiment, GreenCert certifies a persistent-event onset to a single optimizer update in a million-parameter LayerNorm Transformer, matching the onset subsequently observed in a separately executed float64 continuation.

1 INTRODUCTION

A training forecast can predict an accuracy gain and still miss a later reversal. To establish either claim, we need to bound the trajectory between the checkpoint and the proposed event. A numerical reference suggests where training will go; a certificate establishes where the specified optimizer trajectory can go, under explicit mathematical and numerical assumptions.

GREENCERT measures each reference step's mismatch with the optimizer and tracks its linear effect on later states. These effects remain signed vectors, so they can cancel before their size is bounded. The remaining task is to control the nonlinear error. A self-consistency test establishes that the exact-real trajectory from the stored checkpoint stays within a computed neighborhood of the corrected reference. Bounds on prediction margins then establish accuracy-threshold events, reversals, or target absence on a fixed evaluation set.

The computational contribution is a whole-window enclosure obtained through verified vector products. Applying the finite-window response operator and its adjoint requires Hessian–vector products, rather than a dense Hessian at every update. The implementation bounds the errors in those products and stops gain estimation when the enclosure test passes. Its randomized gain estimate assumes ideal Gaussian probes; the enclosure theorem itself is deterministic given valid bounds. We distinguish that assumption from outward arithmetic and from the validity of the neural derivative bounds.

The primary comparison asks every method to enclose the same reference through 8,192 updates. GREENCERT completes four of five selected fresh checkpoints in 13.91–18.69 CPU minutes, including required replay. Neither direct interval continuation nor residual-verified QR completes the same task within four CPU hours per case. A separate optional-policy cohort completes three of five cases and retains two construction errors. These results concern a 528-parameter LayerNorm Transformer and a continuation rate fixed after development. The advantage is not universal: direct continuation is cheaper on the larger digits task through 1,024 updates and in the early-event comparisons.

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.

A separate experiment tests parameter scale rather than a long window. For a 1,008,288-parameter, four-block LayerNorm Transformer, two executions of one construction return the bracket $[43, 43]$ before a separate future observation is opened. The observed onset is $+43$. This result uses outward local arithmetic conditional on supplied neural bounds and the Gaussian projection event; the subsequent observation is float64, with independent output replay. We report the scale case separately from the fresh cohorts.

Section 2 defines the trajectory problem; Section 3 gives the method, theorem, and output guarantees. The experiments compare completed guarantees and retain abstentions, construction errors, resource stops, and nontriggering seeds.

2 VERIFICATION PROBLEM

Let $x_{t+1} = G(x_t)$ be a differentiable deterministic optimizer map. The state includes parameters and any auxiliary optimizer variables. Fix a stored checkpoint x_a, a window of H updates, and a proposed path $c_0, \dots, c_H$, called the *reference*. The verification problem is to bound $x_{a+j} - c_j$ throughout the window, not merely at its endpoint. A certificate is a checked bound under stated assumptions; an abstention means those conditions were not established.

Stored binary64 checkpoint entries, data, and optimizer coefficients are interpreted as exact dyadic numbers. The target is the iteration of G in real arithmetic from those inputs. A floating-point continuation instead iterates $\widehat{G}$; agreement with it is empirical and does not transfer the theorem to that program. Such a transfer would require verified bounds on $\widehat{G} - G$. Lemma 1 handles the represented initial offset introduced by rounded momentum coordinates.

Output claims follow from the trajectory enclosure. On a fixed set of N evaluation examples, let n_j be the number classified correctly at x_{a+j}. For a required count r and persistence $K \geq 1$, define

$$T = \inf\{j \geq 0 : n_{j+\ell} \geq r \text{ for all } 0 \leq \ell < K\}. \tag{1}$$

Here $\inf \varnothing = +\infty$; no monotonicity is assumed. A bracket $[L, U]$ bounds the first persistent onset, $L \leq T \leq U$. Within a verified window, a complete block must begin no later than $H - K + 1$. These are fixed-evaluation-set events under unchanged optimizer dynamics.

3 GREENCERT

A one-dimensional example. Two nonzero defects can leave zero error at a later state. Consider the scalar update $G(x) = x/2$ and an anchored reference with two one-step defects, $s_0 = 0.02$ and $s_1 = -0.01$. Its true error starts at $h_0 = 0$ and follows $h_{j+1} = h_j/2 + s_j$: the first error is 0.02, but the second is zero. The second defect exactly cancels the propagated first one. Adding magnitudes instead gives $0.02/2 + 0.01 = 0.02$ at the second state. This cancellation does not remove the first state's error; it shows why the timing and signs matter. For a nonlinear update, a Taylor remainder enters each step as well. The verifier propagates the known linear response, then checks whether the nonlinear remainder fits inside a surrounding tube. We develop these two steps in turn.

3.1 REFERENCE CONSTRUCTION AND THE KNOWN RESPONSE

Call the proposed path the *reference*, $c_0, \dots, c_H$, anchored at the stored state $c_0 = x_a$. If this were a genuine training trajectory, applying one optimizer update to c_j would land exactly at c_{j+1}. The difference is its reference-path defect. The Jacobian describes how a small displacement changes that update:

$$s_j = G(c_j) - c_{j+1}, \qquad J_j = DG(c_j). \tag{2}$$

Propagate each known defect through this linearized dynamics, in order:

$$z_0 = 0, \qquad z_{j+1} = J_j z_j + s_j, \qquad c'_j = c_j + z_j. \tag{3}$$

Here z_j is the signed linear response and c'_j is the corrected reference. Its one-step defect is $G(c_j + z_j) - G(c_j) - J_j z_j$, a Taylor remainder. This identity also permits proposal-level sweeps when constructing a reference. Those unverified sweeps are distinct from the verified response computed after the final reference has been fixed (Appendix B.1).

The operator K_H is shorthand for this propagation: for any sequence of injected errors u, it returns the sequence $(z_1, \ldots, z_H)$ solving $z_0 = 0$, $z_{j+1} = J_j z_j + u_j$. It is called a finite-window Green operator; no stored matrix is needed to apply it. The true reference error is $h_j = x_{a+j} - c_j$. What the linearization misses is

$$N_j(v) = G(c_j + v) - G(c_j) - J_j v. \tag{4}$$

The total error therefore has two parts. The first is predictable: the reference's defects and their linearized directions are already known. The second is the unknown nonlinear remainder. Exactly,

$$\underbrace{h}_{\text{total error}} = \underbrace{K_H s}_{\text{known directional response}} + \underbrace{K_H N(h)}_{\text{unknown nonlinear remainder}}. \tag{5}$$

GREENCERT computes the first part and bounds the second. By contrast, the unsigned ablation replaces the known response norm by $\kappa \|s\|_2$, where $\kappa \geq \|K_H\|_2$ allows any direction of injected error. Norms below are Euclidean over all coordinates and times; $N(h)$ stacks the transition remainders for $j = 0, \ldots, H-1$ with $h_0 = 0$.

3.2 FINITE-WINDOW CLOSURE

We seek a tube wide enough to contain every error that the tube itself permits. Let $\widetilde{z}$ be the computed signed response and let E bound the error around $c + \widetilde{z}$. The test adds three contributions: error in computing the response, nonlinear error at the corrected reference, and additional nonlinear error allowed within distance E. Their sum must not exceed E. The entire neighborhood must also remain inside the radius ρ where the derivative bounds hold.

In the theorem, M bounds Jacobian drift, α bounds response-computation error, and β bounds the propagated nonlinear error at the corrected reference. All bound constants are nonnegative.

Theorem 1 (Response-centered trajectory enclosure)**.** *For $1 \leq j < H$ and $\|v\|_2 \leq \rho$, suppose $\|DG(c_j + v) - J_j\|_2 \leq M \|v\|_2$. Let $\kappa \geq \|K_H\|_2$ and let the anchor-fixed response $\widetilde{z}_0 = 0$ satisfy $\|K_H s - \widetilde{z}\|_2 \leq \alpha$. Put $p = \max_j \|\widetilde{z}_j\|_2$, $q_j = N_j(\widetilde{z}_j)$, and assume $\beta \geq \|K_H q\|_2$. If $E \geq 0$ satisfies*

$$p + E \leq \rho, \qquad \alpha + \beta + \kappa M p E + \tfrac{1}{2}\kappa M E^2 \leq E, \tag{6}$$

then the realized trajectory obeys $\|h - \widetilde{z}\|_2 \leq E$, and hence $\|h_j\|_2 \leq p + E$ for every $1 \leq j \leq H$.

The anchor is exact: $h_0 = \widetilde{z}_0 = 0$, so no drift bound at transition zero is needed. The final state is not a transition input. Lemma 1 gives the represented-initial-offset extension used by the scaled-momentum implementation. Appendix A gives the proof and arithmetic interface.

The implementation searches for a nonnegative radius satisfying these inequalities. Appendix B.4 gives the explicit smaller-root formula and its feasibility conditions. Unavailable bounds or a failed test cause abstention.

3.3 PRIMARY IMPLEMENTATION

The original long-window policy is the primary implementation evaluated here. Its checkpoint, dynamics, horizon, domain and probability budget are fixed before verification. Its implementation is:

1. Form a binary64 optimizer rollout and apply one proposal-level variational sweep. Freeze the resulting dyadic reference; it is not yet a validated trajectory.
2. Compute its signed response with outward gradients and HVPs, retaining recurrence residuals and the physical anchor offset. Obtain parameter-ball derivative envelopes at radius 10^{-8}.
3. Acquire a gain bound from one Gaussian probe, testing closure after each complete forward or adjoint product, up to eight products. Allow derivative bounds to vary by update and use parameter displacements in the nonlinear bound (Appendix A.2), including all arithmetic residuals.
4. Stop acquisition at the first valid tube; issue only after mandatory arithmetic replay. Exhausted bounds or resources give no certificate; construction errors are recorded separately from mathematical abstentions.

The theorem is deterministic given valid bounds. Randomized gain acquisition uses one fixed-operator projection event; stopping at a complete product retains its failure budget (Proposition 2). The measured-defect fallback is an optional extension, not part of this primary policy. Historical event studies retain their separately frozen variants.

3.4 Output guarantees

To connect model-state uncertainty to accuracy, consider a *margin*: the true class's output score minus a competitor's. Positive margins against every competitor imply a correct prediction. For margin m_{iq}, let $m^{-}_{iq,j} \le m_{iq}(x_{a+j}) \le m^{+}_{iq,j}$. Strict positive lower margins establish correctness; one strict negative upper margin establishes incorrectness. These give counts $n^{-}_{j} \le n_j \le n^{+}_{j}$. Apply Eq. (1) separately to n^{+} and n^{-} to obtain L and U. If both endpoints exist within the complete-block horizon, $L \le T \le U$. A valid tube does not require an event bracket. If n^{+} has no complete K-state target block, then $T > H - K + 1$, not $T = \infty$; missing output bounds leave the event unresolved. The bracket can also support update-budget decisions; the retrospective policy analysis is in Appendix D.9.

3.5 Computational structure

Applying K_H is a forward Jacobian recurrence; applying its adjoint is a backward recurrence. For momentum, both use objective Hessian–vector products rather than a full Hessian. Verified recurrence residuals bound errors in these actions, and gain acquisition ends once its bound suffices for nonlinear closure. With q_{G} complete products over H updates, gain acquisition uses $q_{\mathrm{G}}H$ objective-HVP calls. The tested QR comparator instead assembles each p-parameter Hessian with p coordinate HVPs per update, before verified frame transport: pH such calls for a complete window, with $p = 528$ in our comparison.

Direct interval continuation also avoids dense Hessians. Against that baseline, Green exchanges sequential interval propagation and precision retries for signed-response construction, gain products and replay; a small product count alone does not establish lower cost. The relevant comparison charges reference construction, neural bounds, output checks and replay to the same completed guarantee. HVP cost depends on model structure, training-set size and precision. The implementation stores an $O(Hp)$ reference but no dense Hessian or whole-window residual inverse. The measured operating range therefore depends on both derivative cost and enclosure quality, not parameter count alone (Sections 5.1 and E.8).

4 Experimental setup

The primary task is full-window verification: each method must establish total parameter distance at most 10^{-8} from the same reference at every state, starting from the same stored physical checkpoint. The small model has 528 parameters, two Transformer blocks, and LayerNorm, on addition modulo 3. Its checkpoint-generation rate is 0.003; the continuation rate 0.00015 was chosen during development and fixed before the fresh studies. The larger digit-parity task uses a 10,052-parameter tanh MLP, 1,077 training examples, and 361 evaluation examples at the unchanged task rate 0.05.

Direct continuation propagates outward state intervals. Residual-verified QR retains an affine frame and verifies its transport with a whole-domain interval Hessian and inverse-residual bound. The historical reference-centered Taylor comparator is specified separately in Appendix C. All current comparisons charge reference construction, precision retries, required outputs, and mandatory replay. A useful prefix must come from one anchored attempt; prefixes are never joined across retries. A resource stop is distinct from a failed enclosure inequality. Caps and unlaunched arms are reported as such, not as completion times.

The sequential comparators use deterministic outward arithmetic. Green's primary construction additionally assumes ideal Gaussian probes, with failure allowance $1/30{,}000{,}000$ per fixed window. Theorem 1 requires valid neural and arithmetic bounds in either case; authenticating stored bounds does not independently recompute their derivation. Historical float64 constructions and later independent event checks are distinguished in Appendix D.5. The scale result states its own numerical contract next to its outcome.

Fresh-study settings and seed lists were fixed locally before initialization, without an external timestamp. An outcome seal withholds the separately executed future observation, not evaluations of the optimizer at reference states. Development studies, repeated executions, and multiple thresholds from one seed are not counted as independent replications. Complete protocols and per-case outcomes are in Appendix D.

5 RESULTS

5.1 WHOLE-WINDOW VERIFICATION AND OUTPUT CONSEQUENCES

GREENCERT verifies all 8,192 updates on four of five selected fresh checkpoints. Successful runs take 13.91–18.69 CPU minutes, including mandatory replay. Neither direct interval continuation nor residual-verified QR completes the same task within its four-hour CPU allowance. Every method starts from the same checkpoint and must establish parameter distance at most 10^{-8} from the same reference. Three additional registered seeds do not meet the checkpoint-selection rule. Each method receives 14,400 process CPU seconds and 12 GiB RSS, including retries; sequential retries restart at the original anchor.

Table 1: Full-window completion and cost: 528 parameters, $H = 8192$, total parameter distance $\leq 10^{-8}$ from the same reference. Each cohort registers eight seeds, selects five and retains three nontriggers. Green times cover successful constructions including mandatory replay; 240 is the per-job CPU cap, not a completion time. Direct/QR are deterministic outward methods; Green additionally assumes ideal Gaussian probes with failure budget $1/30{,}000{,}000$ per window. The optional fallback has no fresh rescue.

Frozen cohort	Method	Complete	CPU minutes	Other outcomes
Primary	Green	4/5	13.91–18.69	1 abstention
	Direct	0/5	240 (cap)	5 CPU limits
	Residual-QR	0/5	240 (cap)	5 CPU limits
Fallback-enabled	Green	3/5	13.91–21.35	2 construction errors
	Direct	0/5	240 (cap)	5 CPU limits
	Residual-QR	0/5	240 (cap)	5 CPU limits

Both studies froze policies and seeds before initialization and compare the same reference and endpoint (Table 1). All three successful optional-policy runs use ordinary closure; its two construction errors occur before the fallback can be attempted. The strongest fresh direct prefix is 6,366. Figure 1 and Appendix D.1 retain every seed, stopping reason and resource record. Frozen local records were not publicly timestamped.

The development checkpoints A–C identify a concrete cost difference. Residual-QR on A retains a maximum distance bound of 7.41×10^{-15} through 539 updates before its four-hour CPU cap. Completed-step Hessian construction consumes 4,541.48 CPU seconds; coordinate transport and inverse-residual checks consume 4,966.70 and 2,355.85. Its enclosure is still informative, rather than lost to wrapping. Green completes the same 8,192-update target in 972.55 CPU seconds, including replay. This evidence concerns the cost of obtaining the enclosure with these implementations, not an isolated causal effect of signed cancellation. Appendix C.2 retains the other development checkpoints, direct/centered comparisons and component-accounting scope.

The optional fallback and its post-outcome rescue of the known primary abstention are defined in Appendix E.3. The original abstention remains in Table 1.

The development tubes also certify output consequences. On two fixed evaluation examples, A has one correct prediction at $+614$: $+\ 5179$ and zero at $+5180$: $+\ 8192$; B retains one from $+643$ and C has zero throughout. These post-hoc consequences include a reversal, persistent gain and target absence. Endpoint displacements are approximately 2.065, 2.335 and 1.831; Appendix D.3 adds sampled checks on all nine task inputs. Separate event-only baseline runs certify $[614, 614]$ and $[643, 643]$ in 3.4–13.1 CPU minutes. The completion advantage above concerns the registered full-window task, not these early events or a completed-runtime ratio against budget-censored jobs.

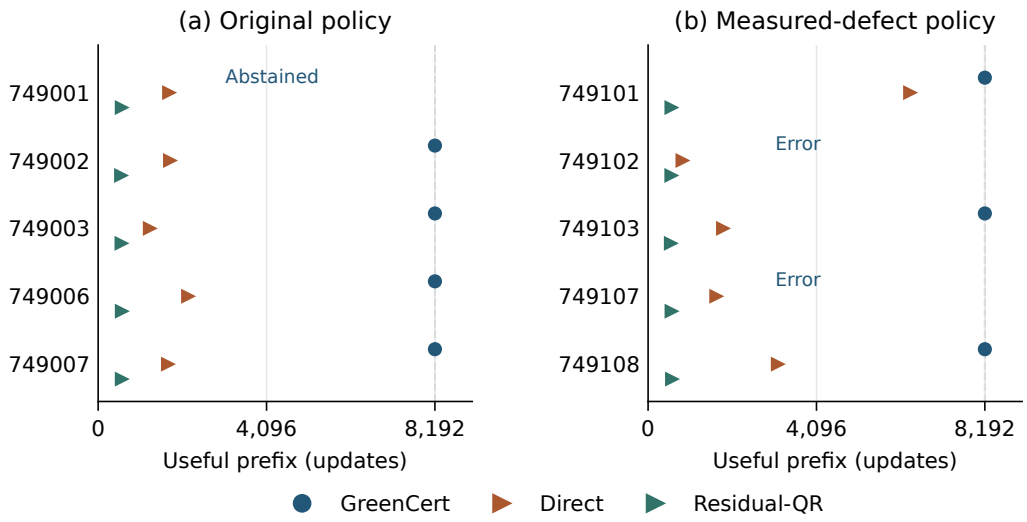

Figure 1: Completed fresh studies at $H = 8192$: original primary policy and optional measured-defect extension. Each registers eight seeds and selects five; nontriggers are retained in Table 1. Circles mark full Green enclosures; triangles mark sequential useful prefixes at the four-hour CPU limit. Green abstentions and execution failures have no certified prefix. Green additionally assumes ideal Gaussian probes; the sequential contracts are deterministic.

5.2 Larger digits comparison

A 10,052-parameter digits study retains task rate 0.05 and 361 evaluation examples. Both completed fresh pairs agree at all 1,025 states, but direct is cheaper: Green/direct CPU ratios are 1.45 and 1.50. Table 2 retains all four registered cases. Fixed direct–Green–QR ordering and changing host availability preclude a four-case algorithmic reliability ranking; resource stops are not closure failures (Appendices D.4–D.4.1).

Table 2: Fresh digits comparison, $H = 1024$. Plain values are CPU minutes for completed windows; parentheses show expenditure before a resource stop, not a completion time. "Not run" denotes the host-reserve launch guard. Costs include retries and mandatory replay. Direct/QR are deterministic outward; Green additionally assumes ideal Gaussian probes.

Case	Direct	GreenCert	Residual-QR
1	161.38	234.26	RSS cap (0.67)
2	161.22	242.53	RSS cap (0.64)
3	135.06	Host reserve (31.65)	Not run
4	Not run	Not run	Not run

5.3 Signed-response ablation and early-event checks

The historical ablation replaces the signed response norm by its operator-gain upper bound. Both versions issue 56 WDBC brackets; signed and unsigned versions issue seven and six on digits. The difference is the digits $[147, 147]$ event: the signed enclosure statistic is 0.2369, versus 13.018 unsigned. The conventional sequential Taylor verifier also retains this event and all 63 issued WDBC/digits brackets. That issuance-selected panel uses outward derivatives rather than the original float64 premises; it tests event retention, not matched cost or superiority over other validated solvers (Appendix C.3).

In the post-seal fixed-radius Transformer audit, unsigned propagation retains one of nine signed brackets. On 18 evaluable candidates, the replacement inflates the response bound by median 319.7

(range 83.0–411.5). A separate cohort's response-centered rule issues 11/19 brackets versus 10/19 for its matched fixed-radius rule. These historical Transformer calculations use float64 without complete arithmetic-error enclosures. All 11 response-centered remainder diagnostics exceed their nominal radii by at most 3.07×10^{-14}; the unchanged brackets and padding audit do not turn them into outward proofs. Appendix D.5 retains the full inventory and numerical contracts.

Direct continuation is also cheaper when only an early event is needed. The four-fresh-seed study selects two seeds and tests four windows at $H = 32, 64$; both methods certify the sole positive-future event. Median Green/direct costs, including replay, are 125.38/15.72 seconds at $H = 32$ and 241.22/26.72 at $H = 64$. A reduced-probe follow-up issues two of six brackets, both matched by direct. These complete short-window comparisons are reported in Appendices D.7 and E.5.

5.4 Million-parameter verification and cost

Table 3: Separate scale demonstration. Two executions share one checkpoint and probe set. Costs include interruption and recovery; they are not isolated speed measurements. Full coefficients are in Table 15.

Model/window	1,008,288 parameters; four LayerNorm blocks; $H = 64$
Event	$35/58$ correct for $K = 5$ states
Bracket / observation	$[43, 43]$ / $+43$ in separate float64 continuation
Numerical contract	Outward local arithmetic; supplied neural bounds; ideal-Gaussian projection event, $\delta = 1/30{,}000{,}000$
Construction/replay	113.42 h parallel; 127.07 h serial, with interruptions

GreenCert closes an event bracket for a 1,008,288-parameter Transformer with four blocks and learned LayerNorm, using full-batch momentum on addition modulo 17. At saved update 10,275, the event requires at least 35 of 58 evaluation examples to be correct for five consecutive states.

Two audited constructions returned $[43, 43]$ before the future was revealed. The subsequent one-shot 64-update continuation placed the first persistent onset at update 10,318, exactly $+43$. All 65 observed counts lie within the issued bounds; independent output replay reproduces 64,090 logits without disagreement or argmax ties. The five-state block ends at $+47$ (Appendix Figure 2).

The serial and parallel executions test reproducibility of one case, using the same checkpoint and probes. The certificate is conditional on supplied neural bounds whose stored identity is checked, outward local arithmetic, and the stated Gaussian projection event. The subsequent optimizer trajectory is CPU float64 with independent point-output replay. Appendix D.8 records architecture dimensions, archive checks, and execution amendments. Appendix Table 15 gives every closure coefficient and slack. A compact rational-arithmetic checker reproduces closure and margin transport from authenticated supplied summaries (Appendix D.8.1); Appendix E.6 retains the same-case structural ablations.

The 1.008M amended budget clocks record 113.42 hours for parallel construction/replay and 127.07 hours for serial, including interruption and recovery. Concurrent execution prevents interpreting their ratio as an isolated speedup. Both complete closure and output-margin checks; Appendix D.8 separates these costs from later closeout.

6 Related work and scope

Shadowing validates approximate trajectories, neural verification bounds outputs under perturbations, and grokking research studies learning transitions. GreenCert connects these objects through a persistent-event claim on the optimizer trajectory starting at a stored checkpoint.

The closest comparisons are fixed-initial-value validation and structure-preserving shadowing, rather than a norm-only error bound. Radii-polynomial methods validate initial-value solutions near a numerical reference (Lessard & Reinhardt, 2014); componentwise shadowing and containment retain directional structure (Van Vleck, 2000; Hayes & Jackson, 2003). Classical shadowing also admits nearby-orbit formulations (Chow et al., 1989; Chow & Palmer, 1992), including in optimization (Orvieto & Lucchi, 2019). Fixing the anchor selects the unique causal inverse of the finite-window

residual map; freeing it introduces a homogeneous response that can change the bound. Appendix A.5 makes this distinction explicit and explains what remains shared with fixed-anchor solvers.

GREENCERT exposes signed optimizer-defect propagation, a matrix-free gain with an arithmetic-error contract, and transport to persistent neural margin events. Adjoint methods already estimate first-threshold timing error (Chaudhry et al., 2021); here the output is a two-sided discrete persistent-event bracket without monotonicity. The implementation avoids a dense residual inverse. The unsigned ablation isolates discarded defect direction; the sequential studies evaluate alternative enclosures of the same trajectory or event. The fresh full-window comparison is limited to the implemented methods, registered references and resource allowance.

Checkpoint verification addresses a related but distinct question. Naughton (2026) certifies full-domain correctness and circuit agreement for a fixed trained grokking checkpoint; GREENCERT instead encloses a future optimizer trajectory to bracket a persistent event.

Neural robustness and abstract gradient training certify different perturbation sets or outputs (Weng et al., 2020; Sosnin et al., 2026). Taylor-model and derivative-preserving neural verification provide related enclosure tools (Berz & Hoffstätter, 1998; Schilling et al., 2022; Sharifi & Fazlyab, 2024). Empirical clocks and mechanistic grokking studies explain or predict learning transitions (Power et al., 2022; Nanda et al., 2023; Notsawo et al., 2023); the proposed verifier complements rather than replaces those explanations.

The guarantees concern fixed evaluation sets and deterministic exact-real continuation. Full minibatch, dropout and AdamW closure remain open. The 8,192-update studies use 528-parameter models and a continuation rate chosen during development; the million-parameter construction covers 64 updates. Combining these horizons and scales remains untested. The two 8,192-update studies retain all 16 initialized seeds, including six nontriggers, one Green abstention and two Green execution failures. Later development repairs produce abstentions, not new certificates. Green's ideal-Gaussian condition differs from deterministic interval continuation; Appendix D.5 separates the historical numerical contracts.

7 CONCLUSION

GREENCERT propagates signed defects to enclose a fixed-anchor trajectory and its output consequences. The primary policy verifies 8,192 updates on four of five selected fresh seeds; both sequential comparators reach their CPU limits. Direct is cheaper on both completed fresh digits pairs through 1,024 updates and in the early-event comparisons. Separately, the million-parameter construction certifies the sealed $+43$ event over 64 updates. The advantage is problem- and horizon-dependent.

AI USE STATEMENT

Generative AI tools (Codex and Claude) were used as auxiliary tools for literature discovery, limited brainstorming and implementation support, and occasional drafting or language polishing. Their outputs were treated as unverified suggestions and were checked, revised, or independently reproduced as appropriate before inclusion. The authors determined the research questions, methodology, mathematical claims, experimental design and interpretation, and final presentation, and take responsibility for the complete contents of the paper.

REPRODUCIBILITY STATEMENT

Appendix A gives the complete enclosure and output proofs; Appendix B specifies verified computation; Appendix C defines the baselines. Protocols and complete outcomes are in Appendix D, and additional variants are in Appendix E. Appendix F maps claims to artifact components and distinguishes record checks, arithmetic replay, and neural reconstruction. A complete 8,192-update Green case has been reconstructed in a clean dependency environment. The million-model reference has also been regenerated exactly, but a new full neural construction has not been performed. The source includes vector figures and their Matplotlib code; experimental inputs and evidence are in the separate anonymous supplement. Identifying links are omitted for double-blind review.

REFERENCES

Martin Berz and Georg Hoffstätter. Computation and application of taylor polynomials with interval remainder bounds. *Reliable Computing*, 4(1):83–97, 1998. doi: 10.1023/A:1009958918582.

Daniel Carando and Jorge Tomás Rodríguez. Symmetric multilinear forms on hilbert spaces: where do they attain their norm? *arXiv preprint arXiv:1810.09373*, 2018. URL https://arxiv.org/abs/1810.09373.

Jehanzeb H. Chaudhry, Donald Estep, Zachary Stevens, and Simon J. Tavener. Error estimation and uncertainty quantification for first time to a threshold value. *BIT Numerical Mathematics*, 61: 275–307, 2021. doi: 10.1007/s10543-020-00825-0.

Shui-Nee Chow and Kenneth J. Palmer. On the numerical computation of orbits of dynamical systems: The higher dimensional case. *Journal of Complexity*, 8(4):398–423, 1992. doi: 10.1016/0885-064X(92)90004-U.

Shui-Nee Chow, Xiao-Biao Lin, and Kenneth J. Palmer. A shadowing lemma with applications to semilinear parabolic equations. *SIAM Journal on Mathematical Analysis*, 20(3):547–557, 1989. doi: 10.1137/0520038.

Wayne B. Hayes and Kenneth R. Jackson. Rigorous shadowing of numerical solutions of ordinary differential equations by containment. *SIAM Journal on Numerical Analysis*, 41(5):1948–1973, 2003. doi: 10.1137/S0036142901399100.

Jacek Kuczyński and Henryk Woźniakowski. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM Journal on Matrix Analysis and Applications*, 13(4): 1094–1122, 1992. doi: 10.1137/0613066.

Jean-Philippe Lessard and Christian Reinhardt. Rigorous numerics for nonlinear differential equations using Chebyshev series. *SIAM Journal on Numerical Analysis*, 52(1):1–22, 2014. doi: 10.1137/13090883X.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Brian Naughton. Certified grokking: A machine-checked certificate for the nanda et al. modular-addition transformer, July 2026. URL https://doi.org/10.5281/zenodo.21893060. Exact-real full-domain checkpoint and circuit certificate.

Pascal Junior Tikeng Notsawo, Hattie Zhou, Mohammad Pezeshki, Irina Rish, and Guillaume Dumas. Predicting grokking long before it happens: A look into the loss landscape of models which grok. *arXiv preprint arXiv:2306.13253*, 2023. URL https://arxiv.org/abs/2306.13253.

Antonio Orvieto and Aurelien Lucchi. Shadowing properties of optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Computation*, 6(1):147–160, 1994. doi: 10.1162/neco.1994.6.1.147.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022. URL https://arxiv.org/abs/2201.02177.

Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010. URL https://www.tuhh.de/ti3/rump/intlab/ActaNumerica2010.pdf.

Christian Schilling, Marcelo Forets, and Sebastián Guadalupe. Verification of neural-network control systems by integrating taylor models and zonotopes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8169–8177, 2022. doi: 10.1609/aaai.v36i7.20790. URL https://ojs.aaai.org/index.php/AAAI/article/view/20790.

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.

Philip Sosnin, Matthew Wicker, Josh Collyer, and Calvin Tsay. Abstract gradient training: A unified certification framework for data poisoning, unlearning, and differential privacy. *Journal of Machine Learning Research*, 27(128):1–58, 2026. URL https://www.jmlr.org/papers/v27/25-2206.html.

Erik S. Van Vleck. Numerical shadowing using componentwise bounds and a sharper fixed point result. *SIAM Journal on Scientific Computing*, 22(3):787–801, 2000. doi: 10.1137/S1064827599353452.

Tsui-Wei Weng, Pu Zhao, Sijia Liu, Pin-Yu Chen, Xue Lin, and Luca Daniel. Towards certificated model robustness against weight perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6356–6363, 2020. doi: 10.1609/aaai.v34i04.6105.

A ENCLOSURE THEOREM AND OUTPUT GUARANTEES

This appendix gives the enclosure proof, operator-error interface, and arithmetic assumptions used by the main algorithm. The research archive contains specialized implementation refinements and their experiment records, separate from the manuscript source package.

A.1 RESPONSE-CENTERED ENCLOSURE

Let $X = (\mathbb{R}^d)^H$ and $U = (\mathbb{R}^d)^H$, both equipped with their Euclidean sequence norms. Causality gives Eq. (5) exactly. For $e = h - \widetilde{z}$, define the map

$$F(e) = (K_H s - \widetilde{z}) + K_H N(\widetilde{z} + e). \tag{7}$$

For $\|e\|_2 \le E$, every transition input lies within the prescribed domain because $\|\widetilde{z}_j + e_j\|_2 \le p + E \le \rho$. The integral derivative bound gives

$$\|N_j(\widetilde{z}_j + e_j) - N_j(\widetilde{z}_j)\|_2 \le M \|\widetilde{z}_j\|_2 \|e_j\|_2 + \tfrac{1}{2} M \|e_j\|_2^2, \tag{8}$$

$$\|N(\widetilde{z} + e) - N(\widetilde{z})\|_2 \le Mp \|e\|_2 + \tfrac{1}{2} M \|e\|_2^2. \tag{9}$$

The second inequality uses $(\sum_j \|e_j\|_2^4)^{1/2} \le \sum_j \|e_j\|_2^2$. Equation (6) makes the closed radius-E ball invariant under F. Set $e^{(0)} = 0$ and $e^{(m+1)} = F(e^{(m)})$. Since F_j depends only on $e_1, \dots, e_{j-1}$, induction fixes coordinate j after j substitutions. Thus $e^{(H)}$ lies in the ball and is the unique fixed point, namely the realized trajectory error. This proves Theorem 1 without a contraction assumption.

Taylor's bound also gives

$$\|q\|_2 \le \frac{M}{2} \left(\sum_j \|\widetilde{z}_j\|_2^4 \right)^{1/2} \le \frac{M}{2} pZ. \tag{10}$$

Substitution yields the quadratic whose rationalized smaller root appears in Eq. (29). A nonnegative radicand alone is insufficient: the linear coefficient and domain conditions must also hold. When $Y = 0$ and $bp < 1$, the displayed expression gives $E = 0$.

A.2 THE CLOSURE USED BY THE SCALE IMPLEMENTATION

The scale calculation refines Theorem 1 in two ways: it retains input-specific derivative bounds, and only parameter displacements enter the nonlinear remainder. In scaled momentum coordinates $x = (\theta, w)$, $w = \eta v$, the optimizer is

$$G(\theta, w) = (\theta - \mu w - \eta \nabla \ell(\theta), \mu w + \eta \nabla \ell(\theta)). \tag{11}$$

Here ℓ includes the quadratic weight penalty. It follows that N_j depends only on the parameter displacement; its nonlinear injection has the form $(-\eta r_j, \eta r_j)$. If $L_{3,j}$ bounds the third derivative of the training loss throughout the parameter ball, $M_j = \sqrt{2}\eta L_{3,j}$ is valid. The producer averages bounds over all 173 training examples; the quadratic penalty contributes no third derivative. Momentum enters the Green gain rather than this nonlinear injection constant. All state and sequence norms remain Euclidean in (θ, w), with parameter projection of norm one.

Let $p_j \ge \left\|\widetilde{z}_j^{\theta}\right\|_2$, $p_0 = 0$, $P = \max_{0 \le j \le H} p_j$, and let τ_{joint} and γ bound the complete response recurrence residual and the unrepresented physical-anchor injection error, respectively. For the $H - 1$ transition inputs $j = 1, \dots, H - 1$, define

$$\alpha = \kappa(\tau_{\text{joint}} + \gamma), \qquad \beta_* = \frac{\kappa}{2} \left(\sum_{j=1}^{H-1} M_j^2 p_j^4 \right)^{1/2}, \tag{12}$$

$$b_* = \kappa \max_{1 \le j < H} (M_j p_j), \qquad c_* = \frac{\kappa}{2} \max_{1 \le j < H} M_j. \tag{13}$$

Empty sums and maxima are zero. Thus, for $H = 1$, only the known first-step forcing remains; there is no post-anchor nonlinear input. Specifically, for the stored rounded response,

$$\tau_{\text{joint}} \geq \left(\sum_{j=0}^{H-1} \|\widetilde{z}_{j+1} - J_j \widetilde{z}_j - s_j\|_2^2 \right)^{1/2}. \tag{14}$$

Each step encloses s_j using an interval gradient and $J_j \widetilde{z}_j$ using an interval Hessian-vector product, then bounds the distance of their enclosing sum from the rounded response. Thus τ_{joint} includes defect-evaluation uncertainty, response-action uncertainty, and local arithmetic rounding. It replaces the separate $\sigma + \tau$ bound in Appendix A.3; those uncertainties are not added a second time. The known velocity offset is encoded in $\widetilde{z}_0$, whose parameter component is zero. Only the unrepresented offset tail contributes to γ through its first-step injection. These statements require valid gradient/HVP enclosures; the scalar audit does not validate the derivative callbacks.

Lemma 1 (Represented initial offset)**.** *Let $\Delta_0 = x_a - c_0$, $a_0 = \Delta_0 - \widetilde{z}_0$, and $d_j = \widetilde{z}_{j+1} - J_j \widetilde{z}_j - s_j$. Suppose $\|d\|_2 \leq \tau_{\text{joint}}$ and $\gamma \geq \|J_0 a_0 + N_0(\Delta_0)\|_2$. Then the known forcing in the post-anchor error $e = (h_j - \widetilde{z}_j)_{j=1}^H$ has norm at most $\kappa(\tau_{\text{joint}} + \gamma)$ after Green propagation. For a pure-momentum offset, $\Delta_0^\theta = \widetilde{z}_0^\theta = 0$, the scale map has $N_0(\Delta_0) = 0$; one may use $\gamma \geq \|J_0 a_0\|_2$ and Eq.* (13)*. Its closure then bounds this e.*

Proof. Subtracting the response recurrence from the actual error recurrence gives

$$e_1 = J_0 a_0 - d_0 + N_0(\Delta_0), \tag{15}$$

$$e_{j+1} = J_j e_j - d_j + N_j(\widetilde{z}_j + e_j), \quad 1 \leq j < H. \tag{16}$$

Write $b_0 = J_0 a_0 + N_0(\Delta_0)$ and $b_j = 0$ for $j > 0$. The represented initial offset has canceled. With N_+ denoting the remainder sequence whose first entry is zero,

$$e = K_H(b - d) + K_H N_+(\widetilde{z} + e), \qquad \|K_H(b - d)\|_2 \leq \kappa(\gamma + \tau_{\text{joint}}). \tag{17}$$

Apply the closed-ball and causal-uniqueness argument of Theorem 1; e includes states $1{:}H$, not the fixed time-zero mismatch a_0. For $a_0 = (0, a_0^w)$, $J_0 a_0 = (-\mu a_0^w, \mu a_0^w)$, so $\gamma = \sqrt{2}|\mu| \, \|a_0^w\|_2$ suffices. □

The physical parameter checkpoint is used for the time-zero event test. The scale response represents the velocity offset exactly, making the remaining tail and γ zero; its post-anchor residual is still enclosed by Eq. (14). This is the nonzero-start extension of the main theorem, not its literal zero-start case.

The implemented test is

$$\alpha + \beta_* + b_* E + c_* E^2 \leq E, \qquad P + E \leq \rho. \tag{18}$$

To see the connection to Theorem 1, bound each center remainder by $M_j p_j^2/2$, retain the Euclidean sum of these bounds, and bound its variation by $M_j p_j \left\|e_j^\theta\right\|_2 + M_j \left\|e_j^\theta\right\|_2^2/2$. Since $\left\|e^\theta\right\|_2 \leq \|e\|_2$, the proof in Appendix A gives precisely Eq. (18). The same causal induction identifies the physical trajectory. Replacing all M_j by their maximum and p_j by full-state response bounds gives valid coarsenings. The additional inequality $(\sum_j p_j^4)^{1/2} \leq (\max_j p_j)(\sum_j p_j^2)^{1/2}$ recovers the simpler main-text forcing bound. The scale result uses Eq. (18), not an unreported substitution into the uniform-M formula.

Inclusive-anchor indexing in the long-window implementation. The long-window code retains the physical states $0{:}H$ in its error sequence. Its Green action is $w_0 = u_0$, $w_{j+1} = J_j w_j + u_{j+1}$ for $0 \leq j < H$: the first identity block propagates the initial-state uncertainty. The gain bound is computed for this $(H + 1)$-block operator, not borrowed from the H-block operator in the main theorem. The response residual stacks the unrepresented anchor error with all H local recurrence residuals; their Euclidean norm supplies the forcing bound. Input-specific nonlinear bounds cover the physical transition inputs $0{:}H - 1$. Prepending a fixed zero coordinate allows the existing scalar closure routine to handle this sequence. The same causal proof applies: its first remainder coordinate is the fixed initial mismatch, and each later coordinate depends only on earlier ones. For an interval initial state, the argument holds for each enclosed realization. The padded coordinate is algebraic only and never enters output or event assembly. Thus a record with 8,193 Green blocks certifies 8,192 optimizer updates and the checkpoint state, not 8,193 updates.

A.3 INEXACT RESPONSES

If the computed defect $\widetilde{s}$ has $\|s - \widetilde{s}\|_2 \le \sigma$, and

$$d_j = \widetilde{z}_{j+1} - J_j \widetilde{z}_j - \widetilde{s}_j, \qquad \|d\|_2 \le \tau, \tag{19}$$

then $\widetilde{z} = K_H(\widetilde{s} + d)$ and $\alpha = \kappa(\sigma + \tau)$ is valid. Similarly, if a computed $\widetilde{q}$ and response $\widetilde{y}$ satisfy $\|q - \widetilde{q}\|_2 \le \sigma_q$ and recurrence residual τ_q, then $\beta = \|\widetilde{y}\|_2 + \kappa(\sigma_q + \tau_q)$ is valid. The signs in the recurrence matter; the enclosing norm is applied to the remaining error, not to an arbitrarily replaced correction.

An equivalent convention sets $\widetilde{z}_0 = 0$ and replaces the first forcing by $s_0 + G(x_a) - G(c_0)$. The response residual must be computed against that modified forcing. This replaces the nonzero-start convention in Lemma 1; adding the full injection again when it is already represented in $\widetilde{z}_0$ would double count it. A general offset includes $N_0(\Delta_0)$; for the scale map's pure-velocity offset it is zero. Both conventions test the physical parameter checkpoint at time zero.

A.4 MARGIN AND PERSISTENCE TRANSPORT

For evaluation example i, let $f_i : \mathbb{R}^{d_\theta} \to \mathbb{R}^C$ be the full logit vector and let $\bar{\theta}_j$ be the parameter component of the reference. Define the two derivative norms by

$$B_{1,ij} \ge \sup_{\|v\|_2 = 1} \left\| Df_i(\bar{\theta}_j)v \right\|_2, \tag{20}$$

$$B_{2,ij} \ge \sup_{\left\|\theta - \bar{\theta}_j\right\|_2 \le \varrho} \sup_{\|u\|_2 = \|v\|_2 = 1} \left\| D^2 f_i(\theta)[u, v] \right\|_2. \tag{21}$$

Thus $B_{1,ij}$ is an induced Euclidean operator bound and $B_{2,ij}$ is a vector-valued bilinear operator bound, uniform on the whole ball. For $m_{iq} = (e_{y_i} - e_q)^\top f_i$ and $\left\|\theta_{a+j} - \bar{\theta}_j\right\|_2 \le \varrho$, Taylor's theorem gives

$$|m_{iq}(\theta_{a+j}) - m_{iq}(\bar{\theta}_j)| \le \sqrt{2}\big(B_{1,ij}\varrho + \tfrac{1}{2}B_{2,ij}\varrho^2\big). \tag{22}$$

The factor is $\|e_{y_i} - e_q\|_2 = \sqrt{2}$. A bound on one scalar class Hessian cannot be substituted for $B_{2,ij}$: per-class bounds L_q instead give a vector bound $(\sum_q L_q^2)^{1/2}$, or one may bound the margin Hessian directly by $L_{y_i} + L_q$ with no extra $\sqrt{2}$. The minimum competitor margin obeys the same uniform perturbation bound. A response-centered output evaluation may use $\bar{\theta}_j + \widetilde{z}_j^\theta$ and radius E, with derivative inflation from the original center and the same domain checks.

Every realized K-persistent event satisfies the possible-count condition, so the first possible block begins no later than T. A guaranteed block forces a realized block, so T occurs no later than the first guaranteed block. This proves $L \le T \le U$ without assuming monotonicity. Endpoint searches stop at $H - K + 1$; an incomplete block at the window end cannot certify persistence. Therefore $U + K - 1 \le B$ implies completion by budget B on the same validity event. Reusing a bracket for multiple budgets introduces no new randomized query.

A.5 FIXED ANCHORS AND STRUCTURE-PRESERVING VALIDATION

For a fixed $x_0 = x_a$, define the finite-window residual $\mathcal{R}(x_1, \dots, x_H)_j = x_{j+1} - G(x_j)$ for $0 \le j < H$. Its derivative at the reference is the square block lower-bidiagonal operator

$$(Lv)_j = v_{j+1} - J_j v_j, \qquad v_0 = 0. \tag{23}$$

The diagonal blocks are identities, so L is invertible for every finite window, even when an individual J_j is singular or unstable. Forward substitution gives $L^{-1} = K_H$. Since $\mathcal{R}(c) = -s$, $K_H s$ is the anchor-fixed Newton correction. Invertibility does not imply a small gain or successful nonlinear closure.

If the anchor is instead allowed to move by a, the same linear recurrence has solutions $v_j = \Phi(j, 0)a + (K_H s)_j$, where $\Phi(j, 0) = J_{j-1} \cdots J_0$. Thus a small nearby-orbit shadow need not enclose the specified anchor's orbit. For $J_j = \lambda > 1$, $s_0 = \varepsilon$, and later defects zero, $a = 0$ gives $v_j = \varepsilon\lambda^{j-1}$, whereas $a = -\varepsilon/\lambda$ gives $v_j = 0$ for all $j \ge 1$. This elementary distinction applies to

Table 4: Closest verification objects. The comparison concerns the stated outputs, not an impossibility of adapting a general solver to neural events.

Approach	Shared or differing structure	Relation to this work
Fixed-initial-value radii-polynomial validation (Lessard & Reinhardt, 2014)	Fixed initial data; a posteriori validation near a reference	Shares the fixed-anchor validation objective; GREENCERT implements a discrete optimizer-to-event interface
Componentwise shadowing and containment (Van Vleck, 2000; Hayes & Jackson, 2003)	Retains geometry beyond a single scalar error magnitude	Directional information is an antecedent; the unsigned ablation is not this method class
First-threshold error analysis (Chaudhry et al., 2021)	Targets an event time rather than only trajectory error	GREENCERT transports strict neural margins to a first complete persistent block, permitting nonmonotone counts
Sequential Taylor comparator in this paper	Fixed anchor; retained linear structure; outward remainder	Retains all 63 selected non-modular events; unmatched candidates and costs prevent a verifier ranking

free-anchor formulations; fixed-initial-value solvers already impose the same anchor constraint as GREENCERT.

The contribution supported here is the implemented combination of causal signed-response validation, neural derivative transport, an inexact matrix-free gain interface, and outcome-sealed persistent-event bracketing. Fixed anchors, Newton correction, and directional numerical enclosures individually have classical precedents. The scale result demonstrates that this combination can close a neural event at a million parameters under its stated contract; it does not establish an intrinsic efficiency or generality advantage over all validated solvers.

B VERIFIED COMPUTATION

This appendix collects the reference procedure, gain estimates, stopping proof, and neural derivative bounds used by the enclosure theorem. Experimental comparisons between alternative gain providers are in Appendix E.

B.1 IMPLEMENTATION BEHIND THE INTERFACE

The reference uses the optimizer map at proposed states without reading the subsequently observed trajectory. The original protocols fix three variational sweeps for digits and four for the other headline cohorts; the scale construction is specified below. One sweep propagates the known defect and leaves a Taylor remainder, as in Eq. (3). Relinearization, structured momentum forcing, prefix scheduling, derivative reuse, and higher-order neural jets are implementation options documented in the archival theorem package. They do not redefine the target event, and their later evaluations are not pooled with frozen cohorts as if one final implementation generated every record.

Jacobian-vector products propagate a perturbation forward; vector-Jacobian products apply the transposed map in reverse. For gradient descent these reduce to Hessian-vector products (Pearlmutter, 1994). This avoids storing full propagators, but evaluating valid derivative bounds and output margins can still dominate cost. Gaussian projection bounds use these products to enclose the operator gain. Their probability event and their arithmetic errors are separate requirements, formalized next.

Exact scale reference procedure. Let F denote momentum in unscaled coordinates (θ, v), and let a be the stored checkpoint. All proposal calculations below use float64. The initial affine path has constant linearization $A = DF(a)$ and initial slope $b = F(a) - a$: $u_0 = 0$, $u_{j+1} = Au_j + b$, $c_j^{(0)} = a + u_j$. Four prescribed sweeps follow:

$$z_0^{(k)} = 0, \qquad s_j^{(k)} = F(c_j^{(k)}) - c_{j+1}^{(k)}, \tag{24}$$

$$z_{j+1}^{(k)} = DF(c_j^{(k)}) z_j^{(k)} + s_j^{(k)}, \qquad c_j^{(k+1)} = c_j^{(k)} + z_j^{(k)}, \quad k = 0, 1, 2, 3. \tag{25}$$

Each sweep relinearizes along its own input path. Streaming the loops in time and then sweep order preserves these dependencies without storing all five paths. The four-sweep path $c^{(4)}$ supplies the scale selector's clock. After conversion $w = \eta v$, one further float64 sweep using G supplies the reference passed to the outward producers. The verified signed response is then computed about that final reference. The selected window has $H = 64$; four proposal sweeps and one scaled sweep are fixed, not stopped when an event or small defect is found. The proposal checks finiteness and a 10^6 correction-norm cap. Failure does not authorize another sweep budget.

The historical decision analysis instead uses each cohort's recorded clock: three sweeps for digits and four for the other historical cohorts. Recorded sweep diagnostics retain maximum defect and correction norms, HVP counts, and reached horizons; intermediate event predictions and separate per-sweep timings were not retained. We do not reconstruct those missing measurements from final-reference bounds. The information barrier permits G and DG at proposed future states; it withholds the separately executed physical continuation and its outcome join, not all computation of future dynamics.

B.2 Probability and arithmetic

Proposition 1 (Shared-block Gaussian norm bound)**.** *For a fixed operator T, an integer $m \geq 1$, and $0 < \delta < 1$, let $A = T^\top T$ and draw independent standard Gaussian vectors $g_1, \ldots, g_m$. With $c_\delta = \Phi^{-1}((1 + \delta^{1/m})/2)$, the event*

$$\|T\|_2 \leq \left(\frac{\max_i \|A^q g_i\|_2}{c_\delta} \right)^{1/(2q)} \qquad \text{for every integer } q \geq 1 \tag{26}$$

has probability at least $1 - \delta$ in exact arithmetic.

Proof. For a unit top eigenvector v of A, $\|A^q g_i\|_2 \geq \|A\|_2^q |v^\top g_i|$. The probability that all m absolute Gaussian projections are smaller than c_δ is $(2\Phi(c_\delta) - 1)^m = \delta$. The same projection event works for every power. Random-start spectral methods are established tools (Kuczyński & Woźniakowski, 1992); this one-sided form supplies the required verification interface. □

If an operator is selected adaptively from previous results, its new probe block must be independent of that operator conditional on the past. Predictable per-query budgets δ_n with $\sum_n \delta_n \leq \delta$ give a family guarantee by conditional union bound. Reusing probes to choose a new operator without this independence is not justified by the fixed-operator proposition.

Numerical error is a separate requirement. If computed powers have residual $\|\widetilde{v}_{i,\ell+1} - A\widetilde{v}_{i,\ell}\|_2 \leq \xi_\ell$, with $\widetilde{v}_{i,0} = g_i$, any $R > 0$ satisfying

$$c_\delta R^q \geq \max_i \|\widetilde{v}_{i,q}\|_2 + \sum_{\ell=0}^{q-1} \xi_\ell R^{q-1-\ell} \tag{27}$$

certifies $\|T\|_2 \leq \sqrt{R}$ on the projection event. To see this, write $\lambda = \|A\|_2$. The residual expansion bounds $c_\delta \lambda^q$ by the same nonnegative polynomial of degree at most $q - 1$. Dividing by λ^q shows that $\lambda > R$ is impossible; if every coefficient vanishes, the projection event forces $\lambda = 0$. Residual bounds must themselves be verified. Float64 evaluation with ξ_ℓ simply omitted does not meet this interface; historical level-C records are identified accordingly.

What the probability covers. Theorem 1 is deterministic once its bounds are valid; Gaussian sampling is one method for obtaining the gain bound. Fix the checkpoint, reference, operator and all nonrandom validity assumptions. Let $\mathcal{E}$ denote the projection event above. Every issued claim is valid on $\mathcal{E}$, including when issuance depends on the observed products. Consequently, under the ideal-Gaussian sampling model,

$$\Pr(\text{an invalid certificate is issued}) \leq \Pr(\mathcal{E}^c) \leq \delta.$$

This is a joint false-issuance bound, not a claim that $\Pr(\text{invalid} \mid \text{issued}) \leq \delta$. If issuance has positive probability, the latter is bounded only by $\min\{1, \delta/\Pr(\text{issued})\}$ from this argument. Observed coverage conditional on issuance is a separate empirical quantity.

A saved pseudorandom seed makes an execution reproducible; it does not establish that a finite-bit sampler implements a continuous Gaussian law. The ideal-Gaussian premise is therefore retained explicitly, separate from the outward arithmetic checks. Substituting a deterministically verified upper bound on the same operator norm would remove the projection premise if closure still passed. The reported large-model constructions do not establish such a deterministic replacement at comparable cost.

B.3 Alternating products with arithmetic residuals

The long-window implementation can stop after an odd forward/adjoint product. The following result states the arithmetic condition it uses; it does not require an additional independent probe at each stopping test.

Proposition 2 (Complete-product gain bound)**.** *Fix a finite-dimensional operator K, draw $g \sim \mathcal{N}(0, I)$, and choose $c > 0$ with $\Pr(|\langle v, g\rangle| \geq c) \geq 1 - \delta$ for a fixed top right singular vector v of K. Set $A_i = K$ for odd i and $A_i = K^*$ for even i. Suppose $w_0 = g$ and the computed complete products obey $w_i = A_i(w_{i-1} + r_i)$ with $\|r_i\| \leq d_i$, $d_i \geq 0$. Let $y_m \geq \|w_m\|$. If $R > 0$ satisfies $d_1 < c$ and*

$$c - d_1 - \sum_{i=2}^{m} \frac{d_i}{R^{i-1}} - \frac{y_m}{R^m} \geq 0, \tag{28}$$

then $\|K\| \leq R$ on the projection event. This implication holds simultaneously for every complete prefix m, so a stopping time that chooses among these verified bounds retains the same failure budget.

Proof. Write $k = \|K\|$. The claim is immediate for $k = 0$. Otherwise the exact alternating product $x_m = A_m \cdots A_1 g$ has top singular component $k^m \langle v, g\rangle$, hence $\|x_m\| \geq ck^m$ on the event. Expanding the computed recurrence gives

$$\|w_m - x_m\| \leq \sum_{i=1}^{m} k^{m-i+1} d_i, \qquad c \leq d_1 + \sum_{i=2}^{m} d_i k^{-(i-1)} + y_m k^{-m}.$$

The right side is strictly decreasing in $k > 0$ if any nonconstant coefficient is positive, so Eq. (28) implies $k \leq R$. If all those coefficients vanish, $d_1 < c$ contradicts $k > 0$ and forces $K = 0$. The event is the same for every prefix. Residuals may depend on the realized computation; their verified pointwise inequalities, not independence from the probe, are required. □

For the causal inverse, the forward residual equation $Lw = b + r$ gives $w = K(b + r)$; $L^*w = b + r$ gives the adjoint form. The d_i above are *input recurrence residuals*, not product-output errors. If instead $\|w_i - A_i w_{i-1}\| \leq \xi_i$ is available, the sufficient test is $cR^m \geq y_m + \sum_{i=1}^{m} \xi_i R^{m-i}$. These tests must not exchange d_i and ξ_i: the input-residual error acquires another factor k. The implementation evaluates Eq. (28) outward at 256 bits and tests closure only after a complete product. An unchanged operator and probe, authenticated local residuals, and full replay remain necessary. Partially acquired products and probe-adaptive changes to K are not covered.

B.4 Radius selection

A simple executable choice avoids a second operator response. For $Z = \|\widetilde{z}\|_2$, Taylor's theorem gives $\beta = \kappa M p Z/2$. Writing $b = \kappa M$ and $Y = \alpha + bpZ/2$, a sufficient radius is

$$E = \frac{2Y}{1 - bp + \sqrt{(1 - bp)^2 - 2bY}}, \tag{29}$$

when $bp < 1$, the radicand is nonnegative, and $p + E \leq \rho$. Unavailable bounds or failed conditions cause abstention. The fixed-radius historical variant instead tests $Z + \kappa M R^2/2 \leq R$ at $R = 2Z$. These are distinct closure policies, not interchangeable experimental labels.

B.5 Neural-bound production from network inputs

Network and objective. The scale model takes tokens $(a, b, 17)$, with learned token and position embeddings. Each of four pre-LayerNorm blocks applies causal multihead attention and a feed-forward residual, with a separate learned LayerNorm before each. Width is 144, with four heads

and feed-forward width 576. Attention uses $QK^\top/\sqrt{36}$, row softmax over the allowed prefix, and learned affine projections. The activation is exact-erf GELU, $\phi(t) = t(1 + \mathrm{erf}(t/\sqrt{2}))/2$, not its tanh approximation. A bias-free readout of the last token gives 17 logits. LayerNorm uses biased variance $d^{-1}\sum_k (v_k - \bar{v})^2$, $\epsilon = 10^{-5}$, and learned scale and shift. There is no dropout. The loss is mean cross entropy over all 173 training pairs plus $0.01\,\|\theta\|_2^2/2$, including embeddings, biases, and normalization parameters in the penalty. The update uses $\eta = 0.003$ and $\mu = 0.9$.

Whole-ball derivatives. For each reference parameter and example, the producer propagates a point logit enclosure, an activation-displacement bound, and uniform derivative norm bounds A_1, A_2, A_3 on $\left\|\theta - \bar{\theta}_j\right\|_2 \le \rho = 10^{-8}$. These are ball bounds, not derivatives sampled at the center. A trainable layer $g(x, w)$ retains mixed bounds $B_{ab} \ge \left\|D_x^a D_w^b g\right\|_2$, $a + b \le 3$. For incoming activation bounds a_1, a_2, a_3, sufficient outgoing bounds are

$$A_1 = B_{10}a_1 + B_{01}, \tag{30}$$

$$A_2 = B_{10}a_2 + B_{20}a_1^2 + 2B_{11}a_1 + B_{02}, \tag{31}$$

$$A_3 = B_{10}a_3 + 3B_{20}a_1a_2 + 3B_{11}a_2 + B_{30}a_1^3 + 3B_{21}a_1^2 + 3B_{12}a_1 + B_{03}. \tag{32}$$

All right-hand quantities are outward upper bounds on the joint domain. Nonnegative Taylor-coefficient convolution handles products, attention, and trainable normalization; shared parameter paths remain in the point Gram calculation. Verified matrix norms include decomposition reconstruction residuals.

GELU uses derivative bounds $(1.13, 0.8, 2)$ through order three; softmax uses $(1/2, \sqrt{6}/9, 1/2)$ in Euclidean norms. For unlearned normalization, let $r_{\min} > 0$ bound the minimum standard deviation including ϵ throughout the activation ball. Its derivative bounds are $r_{\min}^{-1}$, $2/(\sqrt{3d}\, r_{\min}^2)$, and $3/(d\, r_{\min}^3)$; learned scale and shift enter the mixed terms. The lower bound is checked before reciprocation; invalid domains fail.

Let c_1, c_2, c_3 bound cross-entropy derivatives with respect to logits throughout their displacement enclosure. A training-loss third-derivative bound and the optimizer drift bound are

$$L_{3,j} = \frac{1}{173}\sum_i (c_{3,i}A_{1,i}^3 + 3c_{2,i}A_{1,i}A_{2,i} + c_{1,i}A_{3,i}), \qquad M_j = \sqrt{2}\eta L_{3,j}. \tag{33}$$

For reproducibility, set $A = \log\sum_{q\neq y}\exp(f_q - f_y) + \sqrt{2}D$ for logit displacement bound D, and let q enclose $\mathrm{sigmoid}(A)$ from above. The implemented constants are $c_1 = \sqrt{2}q$, $c_2 = \min(1/2, 2q)$, and $c_3 = \min(\sqrt{6}/9, 2\sqrt{2}\{q + (1-q)q^3\})$. The quadratic penalty has zero third derivative. Evaluation uses the point Jacobian Gram bound for B_1 and the whole-ball second derivative for B_2 in Eq. (21).

Point derivatives and arithmetic settings. Objective gradients and HVPs use outward reverse-over-forward differentiation of the same network, averaging every training example in chunks of eight. Penalty HVPs add $0.01v$ once after the data mean. Point logits use the same erf, softmax, and normalization definitions. Stored binary64 inputs are exact dyadics. Point geometry tries 64, 96, then 192 bits; neural transport and signed-response recurrence use 192 bits, Green HVPs 128 bits, and scalar assembly 256 bits. The gain uses four probes, powers one and two, and failure budget $1/30{,}000{,}000$. Caches are tied to the exact parameter, example, architecture, and precision; a failed fixed precision ladder causes abstention. Validity requires correct primitive envelopes and implementations, as distinct from authentication of their saved results.

B.6 Derivation of the primitive envelopes

These bounds justify the constants in Appendix B.5 for the actual functions used by the implementation. Inputs and outputs carry Euclidean norms, with Frobenius norms for arrays. Throughout, $\|D^k f(x)\| = \sup_{\|v_i\|\le 1}\|D^k f(x)[v_1, \ldots, v_k]\|$. The diagonal-norm identity for symmetric multilinear forms on real Hilbert spaces is stated in Carando & Rodríguez (2018, Eq. (2)). Pairing a vector-valued derivative with a unit output covector extends that identity to the norms used here. Thus diagonal directional derivatives suffice; softmax has the additional symmetry inherited from the next derivative of log-sum-exp.

Bounded cumulants and softmax. A variable supported on an interval of length D satisfies

$$\mathrm{Var}(X) \le D^2/4, \quad |\mathbb{E}(X-\mathbb{E}X)^3| \le D^3/(6\sqrt{3}), \quad |\mathbb{E}(X-\mathbb{E}X)^4 - 3\,\mathrm{Var}(X)^2| \le D^4/8.$$

Here is a proof matching the constants. Rescale to $[0,1]$ and write $m = \mathbb{E}X$, $v = \mathrm{Var}(X)$. Since $X(1-X) \ge 0$, $v \le m(1-m) \le 1/4$. A measure maximizing the third central moment exists by weak compactness. First variation against a point mass shows that its support lies among the global maximizers of $f(x) = (x-m)^3 - 3v(x-m)$ on $[0,1]$. If $v = 0$ the claim is immediate. Otherwise this cubic has at most two global maximizers: its only interior local maximum is $m - \sqrt{v}$; when that point lies in $(0,1)$, strict increase from zero to that point excludes zero as a maximizer. Only that point and the right endpoint remain. Without an interior maximum, only the endpoints remain. The maximizing law therefore has at most two atoms a, b, whose third moment is $p(1-p)(1-2p)(b-a)^3$. Maximization over $p \in [0,1]$ gives $1/(6\sqrt{3})$; reflection gives the negative bound. For the fourth moment m_4, Jensen's inequality and $|X-m| \le 1$ give $v^2 \le m_4 \le v$. Consequently $-1/8 \le -2v^2 \le m_4 - 3v^2 \le v - 3v^2 \le 1/12 < 1/8$.

For $F(x) = \log \sum_i e^{x_i}$, its second through fourth diagonal derivatives are these cumulants of a variable taking values v_i with probabilities $\mathrm{softmax}(x)_i$. A unit vector v has range at most $\sqrt{2}$, by $|v_i - v_j| \le \|e_i - e_j\|$. Since softmax is ∇F, diagonal reduction yields the global vector-valued bounds

$$\|D\,\mathrm{softmax}\| \le \tfrac{1}{2}, \qquad \|D^2\,\mathrm{softmax}\| \le \tfrac{\sqrt{6}}{9}, \qquad \|D^3\,\mathrm{softmax}\| \le \tfrac{1}{2}.$$

Causal masking restricts coordinates and inserts zero outputs, operations of norm at most one. Independent rows preserve these constants in Frobenius norm by Hölder's inequality for the products of row norms.

Normalization on the entire activation ball. Let $P = I - \mathbf{1}\mathbf{1}^\top/d$, $r(x) = \sqrt{\|Px\|^2/d + \epsilon}$, and $f(x) = Px/r(x)$, where $\epsilon > 0$ is the stored normalization constant. For $\|x - c\| \le D$, the reverse triangle inequality and $\|P\| \le 1$ imply

$$r(x) \ge r_{\min} := \sqrt{\max(0, \|Pc\| - D)^2/d + \epsilon}.$$

For several tokens, take the minimum centered reference norm over rows; the total Frobenius displacement bounds each row displacement. For an interval-valued reference, the producer centers with outward arithmetic, sums lower bounds on squared absolute centered coordinates, and takes a downward square root. It subtracts an upward displacement, clips at zero, and evaluates the displayed expression downward. This construction, rather than a positivity check alone, establishes the whole-ball premise.

To bound derivatives, lift to $g(y) = y/\|y\|$ with $y = (Px, \sqrt{d\epsilon})$. Then $f = \sqrt{d}\,Qg(y)$, where Q selects the first d coordinates. Put $R = \|y\|$, $n = y/R$, and $a = \langle n, v\rangle$ for $\|v\| = 1$. Differentiation gives

$$R^2 D^2 g[v,v] = -2av + (3a^2 - 1)n, \qquad R^3 D^3 g[v,v,v] = 3(3a^2-1)v + 3a(3 - 5a^2)n.$$

With $t = a^2 \in [0,1]$, the first expression's squared norm is $1 + 2t - 3t^2 \le 4/3$; one ninth of the second's squared norm is $1 - t(5t^2 - 7t + 3) \le 1$. The latter quadratic is positive, having discriminant -11. Also $\|Dg\| \le 1/R$. Diagonal reduction and $R = \sqrt{d}\,r$, followed by the contractions P, Q, prove

$$\|Df\| \le r_{\min}^{-1}, \qquad \|D^2 f\| \le \frac{2}{\sqrt{3d}\,r_{\min}^2}, \qquad \|D^3 f\| \le \frac{3}{d\,r_{\min}^3}.$$

Exact-erf GELU. For $\phi(x) = x\Phi(x)$ and standard Gaussian density φ, $\phi' = \Phi + x\varphi$, $\phi'' = (2 - x^2)\varphi$, and $\phi''' = (x^3 - 4x)\varphi$. The first derivative has extrema at $\pm\sqrt{2}$, giving $\sup|\phi'| = \Phi(\sqrt{2}) + \sqrt{2}\varphi(\sqrt{2}) < 1.13$. Also $\sup|\phi''| = 2/\sqrt{2\pi} < 0.8$. The extrema of $|x|^k e^{-x^2/2}$ and the triangle inequality give

$$|\phi'''(x)| \le \{3\sqrt{3}e^{-3/2} + 4e^{-1/2}\}/\sqrt{2\pi} < 2.$$

The finite strict inequalities are checked with 192-bit Arb. Coordinatewise application preserves these constants under the stated multilinear norms.

Cross-entropy saturation. Let q bound the total wrong-class probability. For $\ell_y(x) = \log \sum_i e^{x_i} - x_y$, the bounds are

$$\|D\ell_y\| \le \sqrt{2}q, \quad \|D^2\ell_y\| \le \min(1/2, 2q), \quad \|D^3\ell_y\| \le \min\big(\sqrt{6}/9, 2\sqrt{2}\{q + (1-q)q^3\}\big).$$

Indeed, with $p = \mathrm{softmax}(x)$, $I \sim p$, and $W = e_I - e_y$, the gradient is $\mathbb{E}W$, the Hessian is its covariance, and the third derivative is its centered third moment tensor. We have $\|\mathbb{E}W\| \le \sqrt{2}q$ and $\mathbb{E}\|W\|^2 = 2q$. For $I = y$, $\|e_I - p\| \le \sqrt{2}q$; otherwise $\|e_I - p\| \le \sqrt{2}$. Taking the expected cube proves the third bound; its global alternative follows from the cumulant argument above. The expression $q + (1-q)q^3$ is increasing on $[0, 1]$, since its derivative is $(1-q)(4q^2 + q + 1)$. For logit displacement at most D, each wrong-minus-right logit difference increases by at most $\sqrt{2}D$. Thus a whole-ball upper bound is $q = \mathrm{sigmoid}(\log \sum_{i \ne y} e^{c_i - c_y} + \sqrt{2}D)$, evaluated upward.

Learned affine parameters and shared paths. For N tokens the learned normalization is $\gamma \odot f(x) + \beta$, with shared γ, β. Pure activation bounds acquire a factor $\|\gamma_c\|_\infty + \rho$ on the parameter ball. Mixed activation–γ derivatives are derivatives of f; two derivatives in the independent local affine parameters vanish. Bias replication has norm $\sqrt{N}$. At the center, the squared joint (γ, β) gain is the largest eigenvalue among the d blocks

$$\begin{bmatrix} \sum_t f(c_t)_k^2 & \sum_t f(c_t)_k \\ \sum_t f(c_t)_k & N \end{bmatrix}.$$

This accounts for sharing rather than treating tokens as independent parameters. For a composition $F(x(\theta), \theta)$ with mixed bounds B_{ab} and activation bounds A_1, A_2, A_3, the chain rule yields

$$C_1 = B_{10}A_1 + B_{01}, \qquad C_2 = B_{10}A_2 + B_{20}A_1^2 + 2B_{11}A_1 + B_{02},$$
$$C_3 = B_{10}A_3 + 3B_{20}A_1A_2 + 3B_{11}A_2 + B_{30}A_1^3 + 3B_{21}A_1^2 + 3B_{12}A_1 + B_{03}.$$

The point-Gram construction tracks common global parameter paths; whole-ball inflation uses these deterministic mixed-derivative bounds. For mean cross-entropy, average the per-example third-derivative bounds. Quadratic weight decay contributes no third derivative, and the scaled-momentum map contributes the factor $\sqrt{2}\eta$.

Independent implementation checks. An autograd test evaluates 234 directional derivatives at 78 locations: softmax in dimensions 2, 3, and 7; normalization away from its reference center; and composed two-block Transformers, including perturbations at 40% and 90% of a parameter-ball radius. All satisfy their corresponding bounds. These tests use a different differentiation path, not a second outward neural library. They check implementation behavior; the preceding derivations, not sampling, establish whole-ball validity.

C Baseline implementations

All comparisons start from the same stored physical checkpoint and use the same requested event or total-distance target. This appendix separates the historical Taylor baseline from the range-aware direct and affine-frame implementations used in the current trajectory studies. They are implemented comparators, not representatives of every high-order validated solver.

C.1 Direct and reference-centered continuation

Direct continuation carries Arb intervals for parameters and velocity between optimizer updates. It never replaces the state enclosure by its midpoint. Reference-centered continuation instead carries an error enclosure about the common saved reference. Both enclose the same differentiable network functions, including LayerNorm and softmax.

The current range-aware operations use square ranges rather than generic interval multiplication and enforce $\mathrm{Var}(x) + \epsilon \ge \epsilon > 0$ when bounding the reciprocal square root and its derivatives. Softmax uses $p_j = (1 + \sum_{k \ne j} \exp(s_k - s_j))^{-1}$ with $0 \le p_j \le 1$. These are range restrictions on the original functions, not derivatives of clipped surrogates. Earlier implementations lacking these restrictions

remain labeled as historical controls. Their domain failures do not prove that the exact LayerNorm is singular.

Each precision retry restarts from the stored anchor. The long-window policy starts at 384 bits, uses the registered failed-radius heuristic with at least doubling, and caps precision at 262,144 bits. The larger digits studies use the fixed 384/768/1536-bit ladder. A retry follows nonfinite bounds, unresolved required outputs, or failure of the stated distance target; study-specific early-event stopping and no-prefix-growth rules are retained in Appendix D. All attempts count toward the allowance. The reported useful prefix is contiguous within one attempt, never assembled from several retries.

C.2 Residual-verified QR

The affine-frame comparator carries $X = c + Br$ between updates. A whole-domain interval Hessian encloses $DF(X)B$; the parameter-coordinate Hessian is assembled with p coordinate HVPs per update. A floating-point QR factorization proposes a dyadic frame Q. The algorithm carries this frame forward and uses its coordinate hull only for derivative bounds and distance reporting, not to restart the state. QR accuracy and orthogonality are not assumed.

Let M enclose $DF(X)B$ and let d be the mapped-center rounding defect. With $P = Q^\top$, compute outward $b \geq \|I - PQ\|_\infty < 1$ and $u = (PM)r + Pd$. An enclosure of $Q^{-1}(Mr + d)$ is obtained by adding a symmetric coordinate radius $b\|u\|_\infty/(1-b)$ to u. This follows from $Q^{-1} = (I - (I - PQ))^{-1}P$ and the Neumann-series bound (Rump, 2010, Theorem 10.2). The original full-interval-inverse QR implementation is retained only as a development control. An identity-frame control uses the same full Hessian kernel without rotation.

On development case A at $H = 64$, 384 bits and no output checks, the identity and residual-QR controls complete in 869.70 and 1,661.87 CPU seconds, with the same maximum distance bound 3.3221×10^{-15}. The full-inverse variant reaches prefix 49 before its one-hour CPU cap. At $H = 8192$, the identity and residual-QR controls reach prefixes 1,044 and 539 before their four-hour caps, while their distance bounds remain below 1.11×10^{-14} and 7.41×10^{-15}, respectively. Thus the observed limit is derivative/frame cost, not loss of enclosure quality. These development results and component timings are retained in the artifact's original comparator record.

C.3 Historical sequential Taylor verifier

For the binary tanh models, $G(\theta) = \theta - \eta\nabla F(\theta)$. At reference c_j, interval network calculations enclose the mean-loss Hessian H_j, the defect norm d_j, and Jacobian drift M_j throughout the radius-r_j ball. If $\beta_j \geq \|DG(c_j)\|_2$, induction gives

$$r_0 = 0, \qquad r_{j+1} = \beta_j r_j + d_j + \tfrac{1}{2}M_j r_j^2. \tag{34}$$

The eigensolver only proposes a basis Q and diagonal D. Interval operations verify $\epsilon \geq \|Q^\top Q - I\|_2 < 1$ and $R \geq \|H_j Q - QD\|_2$, giving

$$\|I - \eta H_j\|_2 \leq \frac{\sqrt{1+\epsilon}\max_i |1 - \eta D_{ii}| + \eta R}{\sqrt{1-\epsilon}}.$$

The numerator and denominator are enclosed outward. Nonlinear endpoints use 192-bit Arb. Elementwise operations and reductions use explicit outward binary64 expansion; matrix products assume an underflow-aware ordinary-dot model, round-to-nearest arithmetic and gradual underflow. Unmodeled BLAS behavior and flush-to-zero modes are excluded.

This panel was selected by prior Green issuance: 40 WDBC seed-anchor windows carry 56 events, and seven digits windows carry seven events. All 63 are retained. It does not estimate comparator issuance on all 96 original non-modular cases or compare complete construction costs under matched numerical premises. Two retrospective diagnostics preserve the same hashed reference. Seed 509 completes its 156-update window and retains $[147, 147]$; seed 510, whose clock predicts 218 instead of observed 219, reaches update 108 and then exceeds the unchanged 10^4 radius cap with next bound approximately 1.23169×10^{13}. It abstains. Recovery times of 1,697.23 and 411.58 seconds omit unavailable pre-interruption work and are not full-cost measurements. Saved-scalar replay checks the recurrences and event assembly, not the supplied neural premises.

C.4 Accounting and diagnostic controls

Reference work, retries, outputs, evidence production and required replay are charged to the requested guarantee. A resource limit records expenditure, not time to eventual completion. QR's dense setup can reach the worker RSS cap before its first useful prefix; a host-memory reserve stop is a different limit. Neither is relabeled as mathematical abstention. Exact stopped-process CPU may be unavailable, in which case tables use the last OS sample. Partial final steps are not assigned to uninstrumented sub-operations.

On development A, interval-Hessian midpoints agree with an independent autograd path to 1.67×10^{-16}, and an outward Rayleigh witness establishes $\lambda_{\min}(\nabla^2 \mathcal{L}) \le -0.51906$. For momentum, the characteristic factor is $x^2 - (1 + \mu - \eta\lambda)x + \mu$; its larger root exceeds one for $\eta > 0$, $0 \le \mu < 1$ and $\lambda < 0$. The recorded Jacobian spectral radius is at least 1.0007732. This excludes one-step contraction in a fixed induced norm, not multistep or time-varying alternatives. The artifact preserves the complete same-domain HVP/Hessian, range-restriction, and hardware-class controls.

D Experimental protocols and complete results

The primary cohorts, historical event studies and development comparisons are reported separately. Counts refer to initialized seeds, selected checkpoints, or issued events as indicated; these denominators are not interchangeable. Execution logs and their identities are indexed in the artifact rather than reproduced as a launch history here.

D.1 Fresh long-window cohorts

Each cohort registers eight seeds before initialization. The primary policy uses 749001–749008; 749004, 749005 and 749008 do not trigger. The optional fallback policy uses 749101–749108; 749104, 749105 and 749106 do not trigger. Each therefore has five selected checkpoints and fifteen method jobs. All thirty jobs have terminal records. The local freezes have no external timestamp. Nontriggers are neither certificates nor numerical abstentions.

The model is the 528-parameter, two-block, width-four LayerNorm Transformer on mod-3 addition, with five training and two evaluation examples. At checkpoints $16, 32, \ldots, 512$, a 64-update proposal with one variational sweep targets one additional correct evaluation example for two states; checkpoints already correct on both examples are skipped. The first predicted onset in $[16, 48]$ triggers selection. The selector reads current and proposed outputs, not the separately observed future.

Checkpoint generation uses rate 0.003. Continuation retains the physical parameters, unscaled velocity, data and objective, but uses the development-chosen rate 0.00015, momentum 0.9, and quadratic penalty $0.01\|\theta\|^2/2$. Each method verifies all $H = 8192$ updates within 10^{-8} total parameter distance of a common binary64-rollout-plus-one-sweep reference. The target is total distance, not Green's recentered remainder.

Each arm receives 14,400 process CPU seconds, 12 GiB sampled RSS, a two-second soft-stop grace and no wall-clock cap. Reference construction, retries, required outputs and mandatory replay are included. Sequential arithmetic starts at 384 bits and follows the fixed retry rule to at most 262,144 bits. Green uses 192-bit response, 128-bit products and 256-bit scalar arithmetic, the fixed neural-envelope ladder, one probe and at most eight complete products. Jobs rotate over three assigned cores on the shared DGX host. These are resource comparisons, not isolated hardware benchmarks.

The per-window ideal-Gaussian allowance is $1/30{,}000{,}000$. At most eight queries per registered study give a union allowance $8/30{,}000{,}000$; independence between issued claims is not required for this union bound. The direct and QR contracts are deterministic outward. Reusing a tube for several output consequences adds no randomized query.

The primary policy completes four windows and abstains on 749001. The optional policy completes three; 749102 and 749107 fail in reference construction before the fallback is eligible. All three optional successes use ordinary closure. All twenty sequential jobs exhaust their CPU budgets; the largest direct and QR useful prefixes are 6,366 and 575. The two Green policies are not pooled into one success rate.

Table 5: Original policy: all selected seed–method outcomes. CPU and wall time are seconds; RSS is peak sampled GiB. A dagger denotes sampled CPU for a terminated worker, not a completed-verification time.

Seed	Method	Outcome	Prefix	CPU	Wall	RSS
749001	Green	Abstained	–	1,685.40	2,318.35	1.43
749001	Direct	CPU cap	1,718	14,402.08†	14,416.73	1.26
749001	QR	CPU cap	566	14,402.05†	14,416.67	1.90
749002	Green	Enclosed	8,192	1,121.38	1,526.23	1.29
749002	Direct	CPU cap	1,740	14,402.01†	14,415.33	1.23
749002	QR	CPU cap	550	14,400.06	14,417.85	2.00
749003	Green	Enclosed	8,192	976.39	1,286.36	1.19
749003	Direct	CPU cap	1,249	14,402.05†	14,417.62	1.20
749003	QR	CPU cap	560	14,400.93	14,417.67	1.91
749006	Green	Enclosed	8,192	834.36	1,059.97	1.21
749006	Direct	CPU cap	2,180	14,400.24	14,415.47	1.17
749006	QR	CPU cap	568	14,402.05†	14,416.59	1.80
749007	Green	Enclosed	8,192	971.74	1,243.51	1.22
749007	Direct	CPU cap	1,690	14,400.11	14,413.44	1.18
749007	QR	CPU cap	567	14,402.03†	14,420.16	1.87

Table 6: Measured-defect policy: all selected seed–method outcomes. CPU and wall time are seconds; RSS is peak sampled GiB. A dagger denotes sampled CPU for a terminated worker, not a completed-verification time.

Seed	Method	Outcome	Prefix	CPU	Wall	RSS
749101	Green	Enclosed	8,192	834.84	1,054.35	1.17
749101	Direct	CPU cap	6,366	14,400.88	14,415.36	1.16
749101	QR	CPU cap	560	14,402.10†	14,416.39	1.89
749102	Green	Error	–	60.51†	60.73	1.07
749102	Direct	CPU cap	829	14,402.06†	14,416.24	1.38
749102	QR	CPU cap	558	14,400.05	14,416.83	1.87
749103	Green	Enclosed	8,192	1,280.99	1,775.46	1.30
749103	Direct	CPU cap	1,815	14,402.09†	14,416.29	1.19
749103	QR	CPU cap	546	14,402.07†	14,419.06	1.92
749107	Green	Error	–	87.67†	88.13	1.09
749107	Direct	CPU cap	1,651	14,401.24	14,414.65	1.21
749107	QR	CPU cap	563	14,402.01†	14,416.42	1.86
749108	Green	Enclosed	8,192	1,256.03	1,731.07	1.28
749108	Direct	CPU cap	3,148	14,400.39	14,415.24	1.16
749108	QR	CPU cap	575	14,400.05	14,418.34	1.76

D.2 Development long-window comparison

A, B and C reuse seeds 74306, 74307 and 74308 at updates 160, 208 and 64; 74305 is a non-trigger. They use the continuation problem above, selected after earlier rates 0.003 and 0.0003 were examined. Update count alone is therefore not a rate-invariant measure of difficulty. The total-distance bounds are 3.544×10^{-14}, 1.624×10^{-13} and 4.889×10^{-14}, also below 10^{-10} and 10^{-12} without rerunning the comparators at those targets. Outward endpoint displacements lie in $[2.0645346371023576, 2.0645346371024345]$, $[2.335299399506999, 2.33529939950759]$ and $[1.831091146116293, 1.8310911461164299]$. Reference path lengths are approximately 2.481, 2.701 and 2.141; these and reference-loss changes are binary64 diagnostics, not outward loss claims.

Matched direct A/B reach their useful prefixes at 51,968/63,872 bits and then begin higher-precision retries before interruption. A fixed-start 51,968-bit A control also stops at useful prefix 3,422, so the inherited low starting precision does not explain that quality boundary. For residual-QR on A, completed-step Hessians, inverse-residual checks and coordinate transport cost 4,541.48, 2,355.85 and 4,966.70 CPU seconds. Those figures exclude partial final steps and reference construction. Its

Table 7: Range-aware development controls at $H = 8192$. Entries are largest useful prefixes; all runs exhaust four CPU hours. Slower-core direct runs are retained separately from matched-core repetitions.

Case	Direct, slower core	Direct, matched core	Centered, matched core
A	1779	3422	1182
B	1082	3561	1082
C	1047	1047	1047

Table 8: Load-bearing primary closure quantities. Products are complete alternating applications, not Gram powers.

Case	κ	α	E	D	Slack/ρ	Products
A	4.159×10^{7}	4.290×10^{-21}	3.333×10^{-15}	3.544×10^{-14}	0.999996	3
B	4.581×10^{6}	4.166×10^{-22}	1.338×10^{-13}	1.624×10^{-13}	0.999984	5
C	7.846×10^{6}	7.146×10^{-22}	2.066×10^{-14}	4.889×10^{-14}	0.999995	4

enclosure remains informative through prefix 539, whereas Green completes all 8,192 updates in 972.55 CPU seconds, including replay.

Primary closure values. Table 8 exposes the original 8,192-update certificates. α bounds propagated response-computation error; E bounds the remaining nonlinear error around the corrected reference. D bounds *total* parameter distance to the common reference; normalized domain slack is $(\rho - D)/\rho$ for $\rho = 10^{-8}$. The scalar closures were recomputed from the authenticated record at their original 256-bit precision. Every case uses one probe, failure allocation $1/(3 \times 10^7)$, and the conditional outward contract. Displayed values are rounded; execution uses stored outward endpoints.

The maximum signed parameter responses are 3.2104×10^{-14}, 2.8565×10^{-14} and 2.8224×10^{-14}. Input-specific nonlinear forcing bounds are 3.3109×10^{-15}, 8.0504×10^{-14} and 1.9202×10^{-14}. These explain the tested closure without attributing the baseline separation to a single isolated mechanism. The box diagnostic measures range and output-check effects; it does not isolate all correlation or precision-policy effects.

On the original two-example evaluation set, A has one correct prediction at $+614$: $+\,5179$ and zero thereafter; B retains one from $+643$ and C has zero throughout. Four of six separate event-only direct/centered jobs certify $[614, 614]$ and $[643, 643]$ in 3.4–13.1 CPU minutes; both C jobs exhaust four CPU hours. These event-only tasks do not support a Green speedup claim. The full-task sampled counts below are additional consequences of the same tubes.

D.3 Output consequences on the complete finite task

The existing simultaneous trajectory enclosures also support queries on all nine mod-3 inputs. At five recorded updates, a 384-bit outward neural evaluation bounds the true-class margins over a coordinate box of radius 2^{-26} around the original reference. This box contains the certified 10^{-8} Euclidean parameter ball. Strictly positive margins prove a correct prediction; one negative true-class margin proves an incorrect prediction; remaining cases, including ties, are unresolved.

At the final update, complete-task correctness is $5/9$, $6/9$ and $6/9$. On A, non-training correctness falls from $1/4$ at $+4096$ to $0/4$ at $+8192$, while training correctness is $5/5$ at both checkpoints. This distinguishes retained training fit from performance on the complete non-training complement. The zero count on C's original two-example evaluation set does not extend to all non-training inputs: one of the other two is correct at every sampled state.

Exact dyadic margin-sign checks resolve 134 of 135 input–checkpoint pairs; the remaining interval on B is retained without retuning the box. All 15 original-evaluation count checks agree with the full-window output record. These are post-hoc finite-domain consequences of the same tubes, not new trajectory or randomized-gain queries. They do not assert behavior between sampled updates or

Table 9: Certified correct counts on the complete finite task, reusing the three development trajectories. Non-training inputs are the four-input complement of the training set, including the selector-used evaluation set.

Case	Subset	$+0$	$+256$	$+1024$	$+4096$	$+8192$
A	Train / 5	3	3	4	5	5
A	Non-train / 4	0	0	1	1	0
B	Train / 5	3	$[3, 4]$	4	5	5
B	Non-train / 4	0	0	1	1	1
C	Train / 5	4	5	5	5	5
C	Non-train / 4	1	1	1	1	1

provide an independent statistical test set. The additional output evaluation uses 1.42 CPU seconds, excluding the original trajectory construction; no output-only cost comparison is made.

D.4 Larger digits cohorts

The development model is a 64–150–2 tanh parity classifier with 10,052 parameters, 1,077 training and 361 evaluation examples, checkpointed after 100 full-batch gradient-descent updates. Continuation retains rate 0.05 and L2 coefficient 0.001. Every method regenerates the identical one-correction reference and targets 10^{-8} total parameter distance. Six CPU hours, 12 GiB worker RSS and 16 GiB available host memory are required. Sequential precision is 384/768/1536 bits, with all retries charged and no retry after a failed prefix ceases to grow. Green uses 192/128/256-bit response/product/scalar arithmetic, one probe, at most eight complete products, and mandatory full neural and arithmetic replay.

Table 10: Development digits checkpoint. Completed CPU includes replay; stopped QR expenditure is not a completion time.

H	Method	Useful updates	CPU min	Peak GiB	Outcome
256	Direct	256	23.99	0.845	Enclosed
256	Green	256	69.03	0.930	Enclosed
256	QR	0	0.54	12.048	RSS cap
1024	Direct	1024	205.81	1.318	Enclosed
1024	Green	1024	286.72	1.708	Enclosed
1024	QR	0	0.62	12.088	RSS cap

At $H = 256$, direct succeeds at 384 bits. At $H = 1024$, it retains prefix 540 at 384 bits and completes at 768 bits. Green needs two complete products at either horizon, with total parameter bounds below 9.46889×10^{-15} and 1.863311×10^{-14}. Both methods resolve identical counts at every state: 330–335 on the shorter window and 330–344 on the longer. These are descriptive finite-set consequences, not prespecified event gates. The two horizons are one development checkpoint, not independent cases. QR reaches its memory limit before the first prefix-quality check; the exact setup allocation was not logged.

D.4.1 Four registered fresh checkpoints

Four new seeds were fixed locally and all inputs materialized before verification, with no accuracy-based selector. The model, update-100 rule, rate and $H = 1024$ endpoint are unchanged. Methods run direct–Green–QR in that fixed order, with six CPU hours and 12 GiB RSS per arm. The host guard requires 18 GiB available before launch and 16 GiB during execution. No seed is replaced or arm retried after a resource stop. Green's ideal-Gaussian allowance is $1/30{,}000{,}000$ per case, or $4/30{,}000{,}000$ for the cohort.

The two completed pairs favor direct: Green/direct CPU ratios are 1.45165 and 1.50438. Green's mandatory replay consumes 48.85% and 52.18% of its cost. Both Green cases use two products and finish below 1.88×10^{-14} total parameter distance. Direct's initial prefixes are 564, 562 and 524; all

Table 11: All registered fresh digits arms. A dash means not launched. Expenditure before a stop is not a completion estimate.

Case	Method	Updates	CPU min	Peak GiB	Outcome
1	Direct	1024	161.38	1.321	Enclosed
	Green	1024	234.26	1.683	Enclosed
	QR	0	0.67	12.051	RSS cap
2	Direct	1024	161.22	1.354	Enclosed
	Green	1024	242.53	1.745	Enclosed
	QR	0	0.64	12.068	RSS cap
3	Direct	1024	135.06	1.333	Enclosed
	Green	0	31.65	1.460	Host reserve
	QR	–	–	–	Not launched
4	Direct	–	–	–	Not launched
	Green	–	–	–	Not launched
	QR	–	–	–	Not launched

three 768-bit retries complete. Paired methods agree at all 1,025 states, with count ranges 324–350 and 320–343; the third direct path starts at 330 and ends at 347.

The third Green arm stops during response construction, last recorded step 747, below its worker RSS cap. This is not a verified prefix. The cause of transient host pressure is unidentified, and later memory recovery does not override the no-retry rule. Fixed method order and changing availability preclude a four-case reliability ranking. Stored dyadic inputs are authoritative: optional Windows checkpoint regeneration differs by at most 1.12×10^{-16}. Full logs and component costs remain in the artifact.

D.4.2 Longer-window development attempt

The exposed development checkpoint was also tested at $H = 2048$, with three registered arms: inherited direct precision policy, Green, and direct starting at 768 bits with a 1536-bit retry. Each receives 12 CPU hours and 12 GiB RSS under the same 18/16-GiB host guard. The input modules, reference rule and 10^{-8} target remain fixed; stopped arms cannot restart. A declared advantage requires Green at most 0.8 times the cheaper completed direct policy, or Green completion while both direct policies reach CPU/quality limits. Host-resource stops do not meet that criterion.

Table 12: All $H = 2048$ development arms. Green has no issued certificate because required replay is incomplete.

Method	Useful updates	CPU h	Peak GiB	Outcome
Direct, 384-bit start	2048	9.525	2.076	Enclosed
Green	0	6.438	2.898	Host reserve
Direct, 768-bit start	–	–	–	Not launched

Direct takes 34,289.58 CPU seconds: prefixes 540 and 1,333 at 384/768 bits precede full completion at 1536 bits, with distance below 2.51×10^{-14} and singleton counts at all 2,049 states. Green stops after 23,177.90 CPU seconds during mandatory product replay; the third arm never launches. The constructed references agree exactly. The comparison is inconclusive, not a Green speedup or a demonstrated closure failure. The execution archive retains every terminal disposition.

D.5 Historical event studies

WDBC uses 24 fresh seeds, a 266-parameter tanh MLP, disjoint train/trigger/certification splits, three gates, $K = 10$, and $H = 300$. Digits uses 12 fresh seeds, a 538-parameter tanh parity classifier on handwritten digits, disjoint stratified splits, two gates, $K = 10$, and $H = 400$. The selectors do not read future certification trajectories. Two separately sealed 24-seed cohorts use the same 13,792-parameter, one-block causal Transformer on addition modulo 17, with three gates, $K = 25$, and $H = 300$. The first uses the fixed-radius closure; the second uses the response-centered rule

Table 13: Original construction and independent event checks are separate. Cand./cases retains registered denominators; lead is median issued onset. All brackets agree with observed onsets, but rows do not represent a common numerical guarantee or pooled independent trials. FR/RC: fixed-radius/response-centered; B and C are defined below. The small-model outward checks validate events, not every original Green intermediate.

Study	Parameters	Cand./cases	Issued	Lead	Original construction	Independent event check
WDBC	266	71/72	56	79	Float64 (C)	192-bit continuation
Digits	538	24/24	7	70	Float64 (C)	192-bit continuation
Transformer FR	13,792	23/72	9	192	Float64 (C)	Float64 observation
Transformer RC	13,792	19/72	11	118	Float64 (C)	Float64 observation
451k case	451,008	1/1	1	44	Conditional outward (B)	Float64/output replay
1.008M case	1,008,288	1/1	1	43	Conditional outward (B)	Float64/output replay

and an anytime power ladder. Selection in the fixed-radius study can inspect the current certification count, so "prospective" refers to future-outcome sealing, not an untouched certification set. Complete configurations remain in the retained protocol records (Appendix F).

A: Exact-real continuation under an arithmetic contract. All 63 issued WDBC/digits events survive independent sequential outward continuation from stored dyadic anchors. These later audits establish the reported level A; the original issuance arithmetic remains as frozen. Arb evaluates nonlinear endpoints at 192 bits; explicit binary64 enclosures assume ordinary round-to-nearest dot products and gradual underflow. Corrected arithmetic and independent output checks are part of the retained repair record. This establishes the exact-real continuation under those assumptions, not the entire preceding PyTorch/BLAS training execution. It does not validate every randomized Green intermediate merely by validating the same event independently.

B: Conditional outward producer arithmetic. The completed 451k and 1.008M constructions use outward producer bounds conditional on stored inputs and an ideal-Gaussian projection event, with failure budgets 10^{-7} and $1/30{,}000{,}000$, respectively. Their saved-evidence replays recompute local arithmetic and event assembly and authenticate supplied neural bounds; the separate closeout does not independently rerun every neural derivative kernel. Subsequent observations are CPU float64 with point-output audits, not independent exact-real optimizer-state continuations. Neither case is added to the earlier cohorts.

C: High-confidence numerical evidence. The 20 historical Transformer numerical brackets use float64 neural/operator calculations and an ideal-PRNG model, without verified nonzero error enclosures for every operation. In the response-centered cohort, the observed remainder diagnostic exceeds its nominal radius in all 11 cases, by at most 3.07×10^{-14}. Total-state containment and brackets remain unchanged, including a 10^{-8} padding sensitivity audit; that audit is not an outward numerical proof. We do not assign these cohorts level A or B.

D.6 Predeclared short-window grid

Four previously unexamined seeds (74101–74104), checkpoints $\{0, 32, 128\}$ and horizons $\{4, 8\}$ were fixed before initialization. All 24 windows reach state closure and fully resolve their count paths: 19 contain no event, four have an anchor event, and one contains a positive-future event. Seed 74104 at update 128 with $H = 8$ gives $[7, 7]$ under both Green and direct 384-bit continuation. Its remainder radius is 3.68363×10^{-21} and total parameter radius is below 1.14988×10^{-12}. There is one eligible future-event opportunity, not 24 independent event trials. The ideal-Gaussian union bound over these 24 windows is 8×10^{-7}. The recorded median producer-stage times are 7.812 and 16.897 seconds at $H = 4, 8$, versus 0.278 and 0.537 seconds for direct continuation. The Green numbers exclude both proposal and required semantic replay; they are component execution records, not matched pipeline timings.

D.7 FORECAST-SELECTED COMPARISON WITH DIRECT INTERVAL CONTINUATION

The registered seeds are 74301–74304. Checkpoints are every 16 updates from 16 through 512. At each checkpoint, if the current evaluation count is below two, a one-sweep local variational forecast proposes a 64-update count path. The first predicted $K = 2$ event for current count plus one with onset in $[16, 48]$ triggers the case. This rule reads current and proposed logits, not a true future rollout. Every seed, including any without a trigger, is retained. Both $H = 32$ and $H = 64$ are attempted for every triggered seed; windows from the same seed are dependent.

The Green reference is a float64 rollout with one scaled-coordinate sweep. The original neural producers, outer radius 10^{-8}, precision settings, four probes and powers one and two are unchanged. Each call has failure allocation $1/(3 \times 10^7)$ under the ideal-Gaussian model. The direct method carries Arb parameter and velocity boxes and restarts from the exact dyadic anchor at precisions $384, 768, 1536, 3072, 6144$ bits. A restart follows unresolved output bounds, a nonfinite enclosure or a LayerNorm variance that the box cannot prove positive. A fully resolved window with no eligible event needs no higher precision. A positive first-passage bracket stops the method immediately, without validating an irrelevant suffix. Hence this comparison does not attribute failure at one chosen precision to direct continuation in general.

Common training and checkpoint selection precede both verifier processes; selection time is logged separately rather than charged to one method. Both complete child processes have a 600-second wall cap and 2-GiB resident-memory cap. Cold wall time runs from process launch to its disposition and exit. Green time includes input checks, reference, producers, graph authentication and required first semantic replay; there is no optional-repeat charge. Reported medians use these per-run totals, not sums of component medians. The operating system's peak working set is sampled every 50 ms. Method order alternates by seed. Workers use one computational thread, but other laptop work remains active: repeated isolated timings are still needed for a controlled speed benchmark.

Seed 74301 triggers at update 32 with forecast onset 33. Seed 74302 does not trigger. Seed 74303 does not trigger. Seed 74304 triggers at update 256 with forecast onset 39.

Table 14: All selected windows. A dash denotes no issued positive-future bracket. Memory is the Green/direct peak working set in MiB. The direct method's interval bracket and later float64 observation are separate.

Seed	H	Green	Direct	Bits	Time G/D (s)	Memory G/D
74301	32	–	–	768	125.6/13.9	319/276
74301	64	$[33, 33]$	$[33, 33]$	768	239.5/14.4	322/276
74304	32	–	–	1536	125.2/17.5	318/281
74304	64	–	–	3072	243.0/39.0	322/288

Only after both method dispositions is a separate float64 continuation observed. It is not used as exact-real ground truth. Direct interval certificates establish the exact-real event intervals independently of the Green closure. Very small radii may underflow in float64 diagnostic fields; the actual continuation retains Arb intervals without midpoint replacement. The comparison concerns this elementary box-continuation implementation, not all structure-preserving validated solvers.

D.8 SEPARATE SCALE EXPERIMENTS AND NUMERICAL CLOSURE

The 451,008-parameter case uses four width-96 LayerNorm blocks, full-batch momentum and mod-17 addition. At update 3925, the five-state 35-of-58 event is bracketed by $[44, 44]$; the 64-update float64 observation gives onset 3969. The 1,008,288-parameter case uses four width-144 blocks, four heads, feed-forward width 576, exact-erf GELU, seed 94101 and checkpoint 10275. The train/trigger/evaluation split is 173/58/58. Appendix B.5 specifies the network, loss and derivative production.

The two million-case executions share the checkpoint and probes, both returning $[43, 43]$ before observation. Each scale observation has 65 states and 64,090 independently recomputed point logits; neither observation is an independent exact-real state continuation. Hardware transfer for 451k, extended compute allowances for 1.008M, and a metadata admission bridge left the scientific

Table 15: Inspectable 1.008M closure. Values are rounded for display; the checks use exact stored endpoints. The input-specific, parameter-projected condition is Eq. (18), derived from Theorem 1 in Appendix A.2. These are conditional arithmetic checks, not new neural-bound validation.

Quantity	Recorded or recomputed value
Horizon H; gain bound κ	64; 4799.61491454
Green probes; Gram powers; failure budget	4; $\{1, 2\}$; $1/30{,}000{,}000$
Joint residual τ_{joint}; anchor error γ	$1.07999593 \times 10^{-30}$; 0
Parameter-response maximum P; domain ρ	$4.01534753 \times 10^{-15}$; 10^{-8}
Error radius E	$1.3928598156237363 \times 10^{-16}$
Forcing α; β_*	$5.18356457 \times 10^{-27}$; $1.37193345 \times 10^{-16}$
Coefficients b_*; c_*	0.0147503454; $1.96488688 \times 10^{12}$
Closure slack $E - \alpha - \beta_* - b_* E - c_* E^2$	$> 1.47529 \times 10^{-32}$
Domain slack $\rho - P - E$	$> 9.99999584 \times 10^{-9}$
Minimum strict margin slack, states 1:64	$> 6.20890 \times 10^{-4}$
Count path; persistent bracket	34 at 0:42, 35 at 43:64; $[43, 43]$

numerical inputs fixed. Their source checks and execution amendments are retained in the artifact. Local hashes establish content identity, not external execution attestation.

D.8.1 What the compact inspection checks

The standard-library rational checker interprets stored endpoints as exact dyadics and uses integer-square-root enclosures on a 400-bit grid. It recomputes closure, domain validity and all 64 post-anchor margin transports from 65 response bounds, 63 drift bounds and 64×58 output rows. The anchor count is supplied; the future observation is never read. The small closure slack reflects a radius rounded just above the smaller root. Display-rounded coefficients cannot replace the exact saved endpoints.

The checker verifies the identity of supplied gain, residual, neural-jet and output summaries, not how every neural bound was produced. Thus its independent scalar checks do not alter the conditional outward contract. Full reconstruction requires the original checkpoint, reference-generation rule, data and neural source. New reference regeneration has matched the original hash, but a new full million-model neural construction has not been performed. Construction/replay budget clocks are 113.42 and 127.07 h, including interruption and recovery; active work versus waiting was not separately measured. The 389.48-second later closeout is not construction.

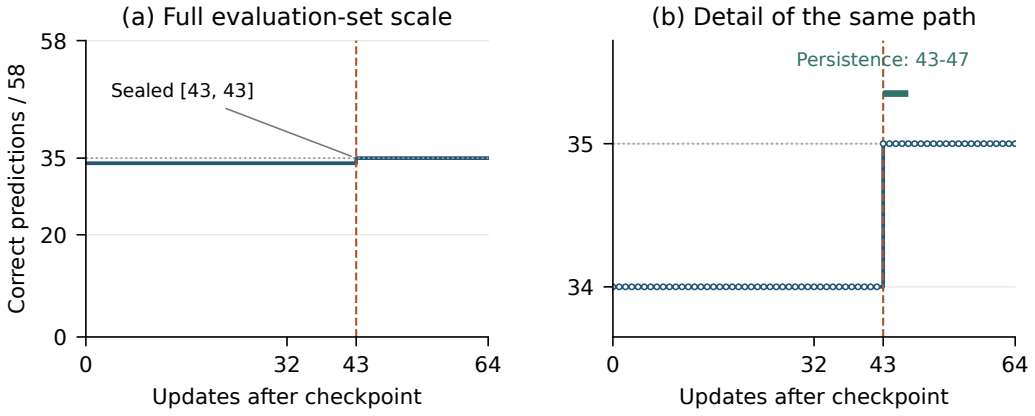

Figure 2: Separate scale observation: full evaluation-count scale and detail of the same path. The sealed $[43, 43]$ bracket matches the float64 onset; the persistence block is $43{:}47$. Independent output replay does not change the supplied-neural-bound or Gaussian premises.

D.9 Retrospective update-budget analysis

Given a clock prediction $\widehat{T}$, a raw positive commitment is $\widehat{T} + K - 1 \leq B$. Given a certified bracket $[L, U]$, the positive commitment is $U + K - 1 \leq B$. Otherwise our evaluated policy abstains; it does not declare that the event cannot occur. For example, $[35, 40]$ with $K = 5$ supports a commitment at budget 44, not at budget 40. The event bracket can be reused for every budget without another trajectory computation or probability charge.

The budget counts optimizer updates, not certificate-construction time.

Across 24 digits candidates from 12 seeds and 401 budgets each, the raw clock makes four false commitments. Both an uncertified screen and the full verifier reject the two inaccurate forecasts responsible. Later stages remove no further false commitments and withhold 1,143 more feasible ones. This retrospective attribution is not a measured decision-cost saving; Appendix D.9 retains the dependent-pair counts and policy curves.

We compare the raw clock, an uncertified pre-screen-only policy, and the complete verifier on all 24 digits candidates and all 401 integer budgets. The added policy applies the original preliminary gates, then uses the raw forecast if they pass. It reads no future outcome; outcomes enter only the separate retrospective scoring step. Table 16 reports all 9,624 dependent candidate–budget pairs, drawn from 12 seeds.

Table 16: Digits policy attribution. Retained and withheld counts concern feasible commitments. Both screened policies reject the same two inaccurate forecast cases; these descriptive counts are not independent trials.

Policy	False commitments	Retained feasible	Withheld feasible
Raw clock	4	3966	66
Pre-screen only (uncertified)	0	3340	692
Complete verifier	0	2197	1835

Eight candidates have no predicted future event, four fail preliminary closure, twelve reach a randomized gain query, and seven issue. Both inaccurate forecasts fail the preliminary signed-response screen $2M \|\widetilde{z}\|_2 > 1$, using $\kappa \geq 1$. The later stages remove no further observed false commitments and withhold 1,143 additional feasible ones; they supply the full verification conditions. This retrospective analysis does not measure incremental stage cost or economic savings. The budget counts training updates, not time spent constructing the certificate. Choosing a resource-allocation policy additionally requires a cost or utility model.

E Additional variants and development findings

These variants do not replace the primary algorithm or overwrite earlier outcomes. Each keeps its original development or fresh-study status. The complete pre-revision records are retained in the artifact relocation index, including unsuccessful tests and execution-only amendments.

E.1 Deterministic gain providers

For scaled momentum, let $L_j \geq \|\nabla^2 \ell(c_j^\theta)\|_2$ and set

$$B_j = \begin{pmatrix} 1 + \eta L_j & |\mu| \\ \eta L_j & |\mu| \end{pmatrix}.$$

Block-norm induction gives $\|K_H\|_2 \leq \|\overline{K}_H\|_2$, where $\overline{K}_H$ is the causal Green matrix of the nonnegative B_j with identity diagonal blocks. Two-component forward/adjoint recurrences give its column and row sums, hence the deterministic bound $\kappa = \sqrt{\|\overline{K}_H\|_1 \|\overline{K}_H\|_\infty}$. The sums use exact fractions and the square root is rounded upward.

The jet-only provider uses $L_j = \lambda + n_{\mathrm{tr}}^{-1} \sum_i (d_{2,ij} a_{ij}^2 + d_{1,ij} b_{ij})$, where a, b bound logit derivatives and d_1, d_2 bound cross-entropy derivatives. The point-Hessian provider takes its minimum with an

outward Frobenius bound from 384-bit coordinate HVPs. For entry bounds $A_{ab} \geq |H_{ab}|$, the latter is $(\sum_{a,b} A_{ab}^2)^{1/2}$, computed with an exact squared sum and an upward square root. The reference, response, ball-valid nonlinear envelopes and closure rule are unchanged. This removes the Gaussian premise only when the resulting deterministic test succeeds; dense Hessian work remains small-model work.

Table 17: All cases in the deterministic gain comparisons. Entries count trajectory closures, not new event certificates. Shared construction costs are not standalone-method timings.

Study	Window/cases	Jet-only	Point-Hessian	Gaussian
Retrospective grid	$H = 4, 12$	11	12	12
	$H = 8, 12$	6	12	12
Fresh rollout	$H = 8, 4$	2	4	4
	$H = 64, 4$	0	2	2

The retrospective dense calculation recomputes 120 Hessians with 63,360 coordinate HVPs in 888.6 CPU seconds. Rational replay checks stored-entry norms, gains and radius tests, not the supplied nonlinear envelopes. The initial fresh modal-reference protocol completed one $H = 8$ case and stopped on a correction-cap failure at the next $H = 64$ proposal, leaving six cases unattempted. The separate rollout protocol then fixed four new seeds at checkpoint 128 and $H = 8, 64$, without a trigger or outcome join. All eight attempts complete; point-Hessian and Gaussian methods retain the same six closures. The full comparison takes 2,341.1 CPU seconds, including shared construction and 280 dense Hessians. Both protocols, including the stopped predecessor, remain in the artifact.

E.2 Time-resolved gain bounds and fresh controls

The preceding comparison separates gain acquisition from the enclosure theorem. We also tested whether retaining the time at which a nonlinear remainder is injected improves closure. These additional studies concern 64-update trajectories of the same 528-parameter model family, not new events or a replacement for the primary long-window policy.

Deterministic temporal bounds. Let $K_H[:, i]$ denote the block that propagates injection i into the post-anchor sequence. The classical backward Gramian recurrence gives

$$V_H = I, \qquad V_j = I + J_j^\top V_{j+1} J_j, \qquad \|K_H[:, i]\|_F^2 = \operatorname{tr} V_{i+1}.$$

Outward bounds $t_i \geq \operatorname{tr} V_{i+1}$ supply $C_i \geq \sqrt{t_i}$ and $\kappa \geq \sqrt{\sum_i t_i}$. For the response and derivative bounds of Appendix A.2, a sufficient test is

$$Y + bE + cE^2 \leq E, \qquad P + E \leq \rho,$$

where

$$Y = \kappa \tau_{\text{joint}} + C_0 \gamma + \tfrac{1}{2} \sum_{j=1}^{H-1} C_j M_j p_j^2,$$

$$b = \left(\sum_{j=1}^{H-1} (C_j M_j p_j)^2 \right)^{1/2}, \qquad c = \tfrac{1}{2} \max_{1 \leq j < H} (C_j M_j).$$

The triangle inequality bounds the center forcing; Cauchy–Schwarz bounds the linear remainder, and $\sum_j \|e_j^\theta\|^2 \leq E^2$ bounds the quadratic remainder. The causal invariant-ball proof then applies. We evaluate the Gramians with 384-bit point-Hessian enclosures. This implementation is dense in parameter dimension, although it does not store the whole Green matrix.

Following a development rescue of an already observed abstention, we froze this policy and initialized four different seeds. Each uses training checkpoint 128, $H = 64$, and the unchanged rollout-plus-one-sweep reference. There is no accuracy trigger or separate outcome observation. Every case receives every gain provider. The local source freeze precedes initialization; it is not an independently timestamped registration.

Fresh case	Jet-only	Point-Hessian	Time-resolved	Gaussian
T1	No	Yes	Yes	Yes
T2	No	Yes	Yes	Yes
T3	No	Yes	Yes	Yes
T4	No	Yes	Yes	Yes
Closures	$0/4$	$4/4$	$4/4$	$4/4$

The deterministic methods retain all four randomized closures. The time-resolved refinement adds no closure over the point-Hessian comparator in this fresh cohort. The full study takes 6,416.25 CPU seconds and 6,649.38 wall seconds, including common construction, replay, 252 point Hessians for the comparator, and 252 recomputed signed Hessians for the temporal method. It is not a standalone-method timing comparison. A separate audit authenticates all 504 stored Hessian arrays and checks the gain conversions and eight accepted deterministic scalar inequalities; it does not reexecute their matrix products or neural derivatives.

Sensitivity-weighted matrix-free control. A separate extension estimates $K_H D_w$ and $K_H D_v$, where the input diagonals have $w_j = M_j p_j$, $v_j = M_j$ for $1 \le j < H$ and zero first entries. If their gains are bounded by k_w, k_v, respectively, then

$$\alpha + \tfrac{1}{2} k_w \|p_{1:H-1}\|_2 + k_w E + \tfrac{1}{2} k_v E^2 \le E, \qquad P + E \le \rho$$

is sufficient. The center remainder factors through D_w with input norm at most $\|p_{1:H-1}\|_2/2$. Split each remainder difference into vectors with bounds $M_j p_j \|e_j^\theta\|$ and $M_j \|e_j^\theta\|^2/2$, using proportional allocation of its norm. Their normalized sequence norms are at most E and $E^2/2$. This gives the displayed inequality without assuming an additional off-center Hessian Lipschitz bound. Upper-rounded diagonals are valid by factorization through the larger weights.

After developing this policy on a known case, we froze a different four-seed cohort at the same checkpoint and horizon. Weighted acquisition uses two four-probe families; its unweighted control uses one eight-probe family. Both use four Gram powers, 2048-bit whole-image products and exactly 4,032 new HVPs per case. The shared reference, response, derivative envelopes and original gain are identical. Policy order alternates.

Fresh case	Original gain	Weighted	Equal-query unweighted
W1	No	No	No
W2	Yes	Yes	Yes
W3	Yes	Yes	Yes
W4	Yes	Yes	Yes
Closures	$3/4$	$3/4$	$3/4$

Thus the extra acquisition yields no additional closure in this cohort. The complete comparison takes 1,563.70 CPU seconds and 1,787.09 wall seconds, with 16,128 new HVPs per compared policy. Each policy's per-case failure allocation, including the shared cap, is 10^{-7}; the union allocation over all compared claims on four cases is $20/30{,}000{,}000$. The ideal-Gaussian and operator-independence premises remain. The exact-rational audit checks 12 calibrations under that law, 48 Gram supersolutions, 512 normalized diagonal entries, and all accepted radius/domain inequalities. It does not independently evaluate the weighted neural operator images.

Both cohorts are retained in the anonymous addendum, including their abstentions and unchanged comparisons. Separate-process common arithmetic replay reproduces the recorded result on all eight cases, using the supplied neural enclosures. The compact checker authenticates selected records and repeats scalar checks; full action arrays, individual neural jets and dense matrices require reconstruction from the provided inputs and frozen sources. These results leave the primary policy and the million-model numerical contract unchanged.

E.3 Measured-defect fallback and reference repair

The optional fallback measures the corrected reference's residual rather than substituting a Taylor upper bound for it. It was developed after observing the primary-policy abstention on seed 749001.

The following derivation states the initial-offset and derivative-shift conditions; the original fresh outcome remains an abstention.

E.3.1 Closure with a measured corrected-path defect

The fallback retains the original reference c_j, Jacobians $J_j = DG(c_j)$ and Green bound. For the stored signed response z_j, let $a_j = c_j + z_j$ denote the exact mathematical sum, evaluated outward rather than rounded back to binary64. Let P_θ project the full scaled state onto its parameters, and take $p_j \geq \|P_\theta z_j\|_2$. On the parameter ball of radius ρ around $P_\theta c_j$, the supplied neural envelope satisfies

$$\|D^2G(x)[u,v]\|_2 \leq M_j \|P_\theta u\|_2 \|P_\theta v\|_2.$$

Velocity is unrestricted because the momentum map is affine in it. Compute outward bounds $D_j \geq \|d_j\|_2$ on the full-state residual $d_j = G(a_j) - a_{j+1}$. For $e_j = x_j - a_j$, exact subtraction gives

$$e_{j+1} = J_j e_j + d_j + R_j(e_j), \qquad \|R_j(e)\|_2 \leq M_j p_j \|P_\theta e\|_2 + \tfrac{1}{2} M_j \|P_\theta e\|_2^2.$$

The linear shift term remains because J_j is evaluated at c_j, not a_j. Removing that term would invalidate reuse of the operator. The residual includes both parameter and scaled-velocity coordinates.

For this full-state sequence bound, include the initial injection explicitly:

$$(\mathcal{L}e)_0 = e_0, \qquad (\mathcal{L}e)_{j+1} = e_{j+1} - J_j e_j, \qquad K = \mathcal{L}^{-1}, \quad \kappa \geq \|K\|_2.$$

This has $H+1$ injection blocks, unlike the post-anchor convention in Lemma 1. Its forward and adjoint actions, for $0 \leq j < H$, are

$$\begin{aligned} (Kr)_0 &= r_0, \quad (Kr)_{j+1} = J_j (Kr)_j + r_{j+1}, \\ (K^* v)_H &= v_H, \quad (K^* v)_j = v_j + J_j^\top (K^* v)_{j+1}. \end{aligned} \tag{35}$$

The stored gain bounds this larger operator, with $H+1$ full-state injection blocks and the actions in Eq. (35). The $H = 8192$ rescue uses 8,193 blocks of dimension 1,056 and the original eight-product gain $\kappa \approx 2.8444680763 \times 10^6$. The artifact records its probe shape, endpoint copies and source bindings.

For comparison, the zero-anchor inverse satisfies $K_H = \Pi K \iota$, where $\iota r = (0, r)$ and Π discards the initial output block. Hence $\|K_H\|_2 \leq \|K\|_2$, not conversely: a transition-only gain cannot in general bound an uncertain initial injection. Such a gain would require a separate bound on the initial-state response, or a zero-initial-error restriction. Here the original query already bounds K; changing the residual forcing leaves its Jacobians, dimensions, Euclidean norm and ideal-Gaussian probability event unchanged.

Let $\gamma \geq \|x_0 - a_0\|_2$; for an interval anchor this bound must cover every admissible state, not just its midpoint. All stacked norms are Euclidean over the $H+1$ full states. Numerical error in measuring $G(a_j) - a_{j+1}$ is included in D_j; the old response-residual forcing is not added again. Define

$$Y = \kappa \sqrt{\gamma^2 + \sum_{j<H} D_j^2}, \quad b = \kappa \max_{j<H} M_j p_j, \quad q = \frac{\kappa}{2} \max_{j<H} M_j. \tag{36}$$

The executable sufficient condition is

$$Y + bE + qE^2 \leq E, \qquad \max_{0 \leq j \leq H} p_j + E \leq \rho. \tag{37}$$

The linear remainder contributes at most $\max_{j<H}(M_j p_j)E$ in sequence norm. For the quadratic remainder,

$$\left(\sum_{j<H} \|P_\theta e_j\|_2^4 \right)^{1/2} \leq \sum_{j<H} \|P_\theta e_j\|_2^2 \leq E^2.$$

Equation (37) makes the continuous finite-dimensional recurrence map send the closed sequence ball into itself. A fixed point exists, and the fixed anchor and causal update uniquely determine it as the exact-real trajectory. Thus $\|e\|_2 \leq E$. The terminal p_H enters the domain check, but no derivative transition starts at H. This is the same self-mapping argument as Theorem 1, with the corrected residual measured directly instead of bounded through its Taylor majorant. The original anchor uncertainty, derivative domain and probability event are retained; the calculation requires no additional Green query.

E.3.2 Development rescue and subsequent confirmation

On 749001 at $H = 8192$, measured forcing falls from $2.44006199 \times 10^{-13}$ to $1.09571392 \times 10^{-20}$, with unchanged $b = 0.4887118995$ and $q = 7.20679106 \times 10^{12}$. The radius is $E = 2.14304664 \times 10^{-20}$ and total parameter distance is at most $3.39669267 \times 10^{-14}$. Outward 192-bit gradients cost 111.60 additional CPU seconds and no new Green query. Original construction plus the successful calculation costs 1,797.00 seconds, or 1,810.16 seconds including a failed wrapper attempt. An independent rational checker verifies the scalar inequality, including a looser $E = 2^{-65}$ witness, but does not recompute the neural bounds. The separate fresh fallback cohort has three ordinary closures and two pre-fallback construction errors, not a new rescue.

The failed proposal corrections on 749102/749107 reach parameter norms about 1.816×10^{20}/1.029×10^{35} while their uncorrected float64 rollouts remain bounded in diagnostics. Two development repairs reject entire corrections exceeding a 2.5×10^{-9} full-state trust budget, retaining the final 10^{-8} enclosure domain. One returns to the rollout; the other constructs a 192-bit point-update reference. Neither replaces validated continuation or proves a clipped correction.

Table 18: All repair attempts abstain before a Green query. Stop is a construction transition, not a certified prefix. Frozen errors are unchanged.

Seed	Proposal	Stop	CPU s
749102	Safeguarded rollout	830	27.47
749107	Safeguarded rollout	2503	63.64
749102	Nonlinear point	1177	93.45
749107	Nonlinear point	2543	123.61

A successful-case regression on 749101 retains its full 8,192-update certificate in 841.01 CPU seconds, with unchanged reference and a different probe seed. This is a known-case regression. Neither failed fresh case is rescued; changed references would require new matched comparator runs.

E.4 Reference refinement and horizon extensions

At development rate 0.0003 and $H = 512$, A gives $[309, 309]$ with direct/Green/predictive-direct CPU costs 306.41/570.11/317.91 s; B gives $[324, 324]$ at 951.66/520.94/522.80 s. C has a Green abstention at 867.38 s and two direct 1,800-s CPU limits. Reducing the rate to 0.00015 moves the events to $+614$/$+643$. These are development choices, not fresh confirmation of ordinary unchanged-rate training.

The refined-reference variant adds one outward-driven correction before recomputing all derivatives and drawing its probe. Each A–C ladder attempts $H = 1024, 4096, 16384, 32768$ and stops at first failure; all twelve closures pass. The $H = 4096$ comparisons reuse probe seeds at changed references and remain matched-probe development diagnostics. The 32,768-update runs use new distinct probe seeds with allowance $1/(3 \times 10^7)$ per window.

Table 19: Development $H = 32768$ comparison, including required replay. This is within Green, not a baseline speed comparison.

Case	Default	CPU min	Refined	CPU min	Products
A	Enclosed	68.31	Enclosed	73.49	2
B	Abstained	118.14	Enclosed	83.63	3
C	Enclosed	88.93	Enclosed	85.03	3

Attempted extension to 65,536 updates. A final development test applies the unchanged refined implementation to the previously observed checkpoint B at $H = 65{,}536$, retaining its optimizer, derivative routines, precisions, and single additional reference sweep. The proposed path remains finite but reaches coordinates of order 10^{15}; its next variational refinement becomes nonfinite at update 53,906. The worker stops during reference construction after 1,518.96 sampled CPU seconds, before any Green query. This is a recorded construction error, not a closure abstention, certified prefix,

or evidence that the actual optimizer diverges. No retry is included. The successful 32,768-update development results remain the longest completed windows. The supplement retains the frozen sources, inputs, failure record and last failed proposal arrays. A separate eight-update reconstruction preserves the smoke test's numerical results; it does not validate the failed longer window.

E.5 Reference controls and reduced-query policies

The development fixture, seed zero at update 156 of the 528-parameter model, has $H = 8$, $K = 2$ and bracket $[4, 4]$. Four preliminary affine-reference corrections plus a scaled sweep give total parameter radius below 9.81841×10^{-16}. Replacing the preliminary sweeps by a float64 rollout, with the same final sweep and all bounds recomputed, retains the bracket; direct 384-bit continuation also retains it. All intermediate proposal clocks predict $+4$. Thus extra proposal sweeps add no issuance advantage on this known fixture. The implementation may abstain when the earliest possible persistent onset is zero, although Eq. (1) allows an anchor event.

Reducing the probe policy from four probes and two Gram powers to one probe and one power recalibrates the same failure allowance $1/(3 \times 10^7)$. At $H = 64$, it cuts construction gain-HVP queries from 1,008 to 126 and retains the development $[33, 33]$ bracket in 76.38 rather than 239.46 cold-process seconds. All neural bounds and mandatory replay remain. The shared-host records are not an isolated speed benchmark.

This policy is fixed before seeds 74305–74308 are initialized. 74305 does not trigger; the other seeds trigger at updates 160, 208 and 64. All six windows complete. At $H = 64$, 74306/74307 give $[34, 34]$/$[36, 36]$ under both methods; their shorter windows end before the events, and neither window of 74308 contains an eligible event. Median Green/direct full costs are 43.44/14.03 s at $H = 32$ and 76.05/17.63 s at $H = 64$. Direct is cheaper on every window. A baseline-first development policy also returns $[33, 33]$ in 16.09 s without a Green query; it does not establish a Green advantage.

On the saved million case, the four-probe first-power gain $75{,}072.8742$ retains $[43, 43]$, with remainder 3.19809×10^{-15}, using 504 rather than 1,008 gain HVPs. The historical run acquired both powers: this is a conditional query reduction, not a measured construction speedup. Neither available one-probe setting passes that case's scalar test.

E.6 Scale sensitivity under fixed supplied bounds

All four combinations of input-specific/maximum drift and parameter/full-state response norms retain the million-case bracket. Their remainder radii are 1.392860, 1.456195, 1.393196 and 1.456543, times 10^{-16}, respectively. Thus these refinements tighten the radius but do not change issuance on this case. The comparison keeps reference, gain, residual, derivative domain and event fixed and rechecks every scalar inequality with rational enclosures.

Discarding the signed response is a different comparison. The uncentered test uses $Y = \kappa(S + A)$ and $Y + \kappa M R^2/2 \le R$, where S bounds the actual reference defects and A the full physical-anchor injection; the response recurrence residual is not substituted for S. At declared prefixes $H = 16, 32, 48, 64$, the unsigned discriminants are $-73.05, -104.67, -128.33, -148.19$, whereas signed closure passes at each prefix. Only 48 and 64 contain the complete $[43, 43]$ event block. The full-window gain bounds each prefix by zero extension/restriction. For $H = 64$, unsigned feasibility requires $\kappa \le 392.9542$, compared with supplied 4799.6149. No operator lower bound excludes a tighter gain. These tests therefore separate signed and unsigned use of the supplied bound, not all possible verified solvers or gain providers. The comparable total sequence error is $\|\widetilde{z}\|_2 + E \le 2.28892 \times 10^{-14}$, not the recentered remainder alone.

E.7 Compute engineering and its limits

The million-case 81.018-GB evidence graph is reduced to 48.710 GB by lossless byte-lane compression. A bounded-memory saved-graph replay reproduces the original result exactly with 3.53 GiB sampled peak RSS; batched interval decoding then reduces replay CPU from 13,225.55 to 8,790.36 seconds, with 3.52 GiB peak RSS. These are stored-evidence sizes, not peak RAM. Replaying supplied neural intervals is not new neural differentiation; the original 113–127-hour construction clocks are unchanged.

On the development digits checkpoint, two timing replicates per policy reduce complete $H = 256$ CPU from mean 2,795.74 to 1,792.48 seconds (35.9%). The candidate uses outward Arb response/product precisions 128/64 bits and training chunks of 128, while neural envelopes and scalar closure remain at 192/256 bits. Both policies need two products and return the same 257 count bounds. The original implementation and outcomes are preserved.

Million-model component tests retain all parameters and training examples. Factoring the quadratic penalty and increasing response chunks from eight to 64 reduces all-64-step candidate-kernel CPU from 5,971.49 to 2,996.10 s. Every candidate derivative interval contains its separately evaluated original-graph 384-bit interval and retains the recorded recurrence bounds. That cross-check costs an additional 11,942.38 CPU seconds. These audits do not repeat Green products or whole-ball envelopes. Component gains cannot be multiplied into an end-to-end speedup.

Frozen-policy comparison on two new checkpoints. The combined compute policy was next fixed before initializing two additional digits-parity seeds. Both use the same 10,052-parameter family, 1,077 training examples, 361 evaluation examples and update-100 checkpoint rule. Each arm must enclose all $H = 512$ updates within 10^{-8} parameter distance of an identical regenerated reference. Per-arm limits are 8,500 CPU seconds and 8 GiB RSS. Green retains mandatory neural and arithmetic replay; direct continuation uses the fixed 384/768/1536-bit retry ladder. Direct's arithmetic contract is deterministic; Green additionally uses the ideal-Gaussian model with allocation $1/(3 \times 10^7)$ per window.

All four arms complete. Table 20 retains their execution order. Both methods return identical finite-set count bounds at every one of the 513 states for each seed. Green/direct CPU ratios are 2.18 and 2.21, with median 2.19: direct continuation is cheaper on both checkpoints. This is a comparison at a fixed shorter horizon, separate from the 528-parameter long-window studies and the same-checkpoint implementation timings above. No event gate was registered for this study.

Table 20: Complete $H = 512$ constructions on two newly initialized digits checkpoints. CPU includes reference construction and mandatory replay. Wall times are measured on a shared host, not isolated timing trials. All four useful prefixes reach 512. RSS is the sampled process maximum.

Seed	Method	CPU (s)	Wall (s)	RSS (GiB)
1	Direct	1678.38	1677.46	1.058
1	Green	3659.19	3679.96	1.231
2	Green	3627.25	3646.35	1.231
2	Direct	1643.63	1642.32	1.059

The protocol was frozen locally, without an external timestamp. A pre-worker supervisor-limit failure and an execution-only amendment are retained in the artifact; the amended execution reuses both stored checkpoints and all numerical sources and takes 10,617.28 CPU seconds.

Budget-limited extension to 1,536 updates. We also applied the same numerical policies to the previously observed 10,052-parameter digits checkpoint at $H = 1536$. This development comparison was fixed after the earlier $H = 256$ and $H = 1024$ results and partial progress of the separate $H = 512$ cohort were known. Both arms regenerate the same reference, use a 10^{-8} Euclidean parameter-distance target, and receive 8,500 process-CPU seconds and an 8 GiB RSS cap. They run concurrently on the shared host. Precision retries, reference construction, and mandatory Green replay are charged to the corresponding arm.

Neither arm completes the full-window guarantee. Direct continuation uses 8,498.91 recorded CPU seconds and retains a useful prefix of 1,333 updates at 768 bits. Its 384-bit and 768-bit attempts first exceed the distance target at updates 541 and 1,334, respectively; the prescribed 1,536-bit retry reaches update 87 before the CPU limit. Green uses 8,499.42 recorded CPU seconds and stops during mandatory Green-product replay, at row 824 of the first replayed product. Consequently it issues neither a full-window certificate nor a certified positive-length prefix. Peak sampled RSS is 1.79 GiB for direct continuation and 2.34 GiB for Green. Elapsed times are 8,501.33 and 8,565.46 seconds, respectively; these are concurrent-run measurements rather than isolated wall-time benchmarks. The aggregate supervisor records 16,998.32 child-CPU seconds.

The direct prefix has exact correct-count bounds at all 1,334 retained states, increasing from 330 to 345 of 361 examples between its endpoints. No event gate was registered for this study. The record therefore adds neither an event certificate nor a completed-method runtime ratio. Both CPU-censored outcomes, all precision attempts, and the mandatory replay stage are retained in the supplement. This comparison does not establish a full-window computational advantage at this checkpoint and budget.

E.8 Cost dependence

With q complete products, H updates and p parameters, gain acquisition requires $O(qH)$ objective HVP calls and stores no dense $p \times p$ Hessian or $(Hp) \times (Hp)$ residual inverse. The current reference still takes $O(Hp)$ storage. Response construction visits every update; neural envelopes visit training and evaluation examples, and margin transport visits $O(Hn_{\mathrm{ev}})$ example–state pairs. Required replay revisits the operations. An HVP's cost depends on depth, width, data and arithmetic precision, so these counts do not establish linear runtime in p.

For completed work, $C_G = C_{\mathrm{reference}} + C_{\mathrm{response}} + C_{\mathrm{envelopes/output}} + C_{\mathrm{products}} + C_{\mathrm{replay}} + C_{\mathrm{other}}$. Direct cost includes every precision attempt. The two completed fresh digits pairs would require total reductions 31.11%/33.53% to tie direct; replay-only tie factors are 2.754/2.798. Even a twofold replay improvement would leave Green 9.71%/11.19% more expensive, with other work fixed. This is a sensitivity calculation, not a measured optimization or a prediction that increasing the horizon will produce a crossover.

F Reproduction guide

The anonymous supplement is distinct from the manuscript source. Its top-level README and manifests identify the original inputs, numerical sources, results and logs for each study. The source package also includes `ARTIFACT_INDEX.md` and a relocation map linking every moved topic to its appendix and original record. Full pre-revision source passages are retained as noncompiled artifact documentation; none is experimental data generated by this edit.

Three different operations. A record check authenticates supplied data and recomputes stated scalar or counting consequences. Arithmetic replay rechecks saved operator actions and local residuals using their supplied neural intervals. Neural reconstruction starts from the original checkpoint and data and recomputes those intervals. Hashes establish input identity, not derivative validity or external execution attestation. The table names fast entry points; each capsule's README gives its separate reconstruction command and dependencies. Run from a fresh extraction and use new output directories.

Table 21: Claim-to-check index. All commands start with `python -B -S`. Capsule names below are anonymous artifact paths.

Claim	Capsule	Fast entry point / scope
Primary cohorts, scale and event inventory	`experimental_materials.zip`	`review.py -output NEW_DIR`; record, scalar and input checks
Causal/offset proof fixtures	`causal_proof_audit.zip`	`verify.py`; exact fixture checks
Primitive envelopes	`primitive_bound_algebra.zip`	`verify.py`; algebra and rational constants
Alternative gains	`gain_comparison_audit.zip`	`verify.py`; calibrations and saved scalar inequalities
Compute changes / fresh $H = 512$	`engineering_audit.zip`, `fresh_compute_audit.zip`	`verify.py`; each retains inputs, policy and all outcomes
Unsuccessful $H = 65536$ / 1536	`horizon_attempt_audit.zip`, `digits_h1536_audit.zip`	`verify.py`; failure and resource accounting
Saved million-model replay	`compact_completion_audit.zip`	`verify_compact_completion.py`; result equality, not new neural construction

Python 3.12 or later suffices for the standard-library checks. Full numerical execution additionally needs the capsule-pinned PyTorch, NumPy and python-flint/Arb environment; plot regeneration uses Matplotlib. The stored dyadic inputs remain authoritative if optional platform-dependent binary64 checkpoint generation differs. The ideal-Gaussian calibration checker verifies selected scalar inequalities, not the distribution of a finite-bit sampler or probe independence.

A clean-install reconstruction of development A at $H = 8192$, including neural production and required replay, matches its checkpoint, reference, closure, counts, three-product stopping point and $[614, 614]$ bracket. It takes 970.08 CPU seconds, 1,206.54 wall seconds and 1.18 GiB peak RSS on the same DGX host, using a new dependency environment. It is one complete case, not an independent backend or external human validation. The original million-model reference has been regenerated exactly in 121.58 seconds, but no new full neural construction is claimed. Its original-input package does not require the large graph to construct new evidence; replay of the original execution does require that evidence.

The artifact retains failed launches, amended execution allowances, interrupted diagnostics, complete terminal tables, file manifests and component logs. Missing future events remain right-censored. Recovered runs, repeated thresholds and later scalar audits do not increase fresh sample sizes. Lightweight extraction and record checks do not silently retrain models or validate omitted neural calculations.