Combining Theory and Benchmarks for Length Generalisation: Formal Certificates Meet Large-Scale Evaluation

Zacharie Bugaud

Abstract

We present a case study in tight theory–benchmark coupling for length generalisation in RNNs. Starting from switched dynamical systems analysis, we derive a gap suppression mechanism and a margin condition that together certify generalisation from weights alone—zero false positives at $H \geq 4k$ (100/100; preconditions verified empirically on all two-phase models); certified models achieve 100% accuracy at L=1,000,000 (66,667× training length, 10/10 seeds). Inverting the certificate into a training recipe yields ≥99/100 across k=3–10, 12, including a blind prediction at k=12 (20/20 vs 0/20 baseline), with extension to 42 absorbing-state regex patterns (349/350 certified, 0 FP). Equally informative are the failures: non-absorbing DFAs defeat the recipe (0% vs 80–100% baseline), gated architectures resist certification (1/20 GRU, 0/20 LSTM) even when the recipe partially works, and the certificate produces false positives when capacity is insufficient. These breakdowns delineate exactly where the theory–benchmark loop needs new theoretical input.

1. Introduction: The Theory–Benchmark Loop

A productive relationship between theory and benchmarks requires each to constrain the other: theory should predict benchmark outcomes, and benchmark failures should expose theoretical gaps. We demonstrate a complete iteration of this loop—theory → certificate → recipe → benchmark → refined theory—for length generalisation in RNNs, connecting to grokking (Power et al., 2022; Nanda et al., 2023), formal language learning (Delétang et al., 2023), and switched dynamical systems (Liberzon, 2003).

Astera Institute, Berkeley, CA, USA. Correspondence to: Zacharie Bugaud <zacharie@astera.org>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea.

Setting. Elman RNNs on binary substring detection (k consecutive 1s), L_{train}=15, eval up to L=50,000. Trained with Adam (lr $= 10^{-3}$) for 10,000 epochs on 256 freshly sampled sequences per epoch. The RNN defines a switched dynamical system: $f_0(h) = \tanh(W_{hh}h + u_0)$, $f_1(h) = \tanh(W_{hh}h + u_1)$.

The puzzle that starts the loop. Baseline training produces bimodal outcomes (∼80% of seeds generalise at k=3, the rest fail entirely). We measure **persistence**: the probability that a random post-detection trajectory stays correctly classified. Every standard diagnostic fails to predict it: spectral radius ($r = -0.22$), eigenvalue danger score ($r = -0.17$), accept-state saturation ($r = -0.06$), and DFA extraction accuracy (7 DFA-perfect models still fail). Standard IFS contraction conditions (Barnsley, 1988) are violated ($\sigma_{\max}(W_{hh}) \in [2, 4.2]$), yet many models generalise. The loop begins: what formal property actually determines generalisation?

2. The Formal Certificate

Theorem 1 (Confinement ⇒ Generalisation). Let $C_k \subset \mathbb{R}^H$ satisfy $f_0(C_k) \cup f_1(C_k) \subseteq C_k$ and $w^\top h + b_{\text{fc}} > 0$ for all $h \in C_k$. If the hidden state enters C_k upon pattern detection, the network correctly classifies at all subsequent lengths.

The key insight is *how* confinement arises despite violated contraction conditions: through tanh saturation geometry rather than spectral properties.

2.1. Gap Suppression (Proof Sketch)

For pre-activation $z_x(h)[d] = (Wh + u_x)[d]$, if the pre-activations share sign for both inputs at all reachable states, the per-dimension output gap between inputs x=0 and x=1 satisfies:

$$|\tanh(z_0[d]) - \tanh(z_1[d])| \leq \tanh'(z_{\min}[d]) \cdot |z_0[d] - z_1[d]| \tag{1}$$

by the mean value theorem, where $z_{\min}[d] = \min_{h \in C_k, x} |z_x(h)[d]|$ (the same-sign condition ensures $|\xi| \geq z_{\min}$ for the intermediate value ξ). Since $\tanh'(z) = 1 - \tanh^2(z)$, at $|z_{\min}| \geq 3$ the derivative is ≤ 0.0099,

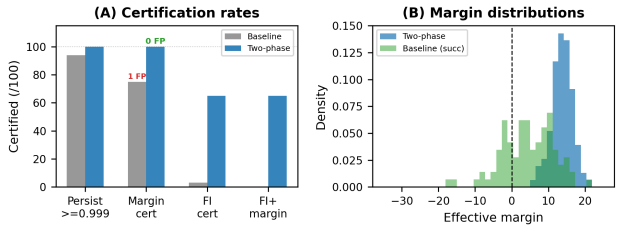

Figure 1. Illustrative; values match reported statistics. **(A)** Certification rates: 100/100 two-phase, 0 FP; 75/100 baseline, 1 FP. **(B)** Margin distributions: two-phase concentrated at $+13.5$, baseline spread at $+4.7$.

Table 1. Certification rates on 200 held-out models.

Condition	Baseline (100)	Two-phase (100)
Persist ≥ 0.99	94/100	**100/100**
Margin cert	75/100 (1 FP)	**100/100** (0 FP)
FI box cert†	3/100 (2 FP†)	65/100 (0 FP)
FI + margin	0/100 (0 FP)	65/100 (0 FP)

†FI checked by random sampling; 2 FPs likely from missed violations.

yielding $>99\%$ gap suppression. We call a dimension **tight** when its gap is below 0.2. The approximation $\text{gap}_d \approx \tanh'(\bar{z}_d) \cdot |\Delta z_d|$ predicts tight dimensions from weights alone ($r = +0.743$, F1 $= 0.872$).

2.2. Margin Condition (Proof Sketch)

Proposition 2. Partition hidden dimensions into tight (gap < 0.2) and loose. Decompose the classifier output:

$$w^\top h + b_{\text{fc}} = \underbrace{\textstyle\sum_{d\in\text{tight}} w_d h_d + b_{\text{fc}}}_{\text{nearly input-invariant}} + \underbrace{\textstyle\sum_{d\in\text{loose}} w_d h_d}_{\text{variable}} \quad (2)$$

For tight dimensions, h_d varies by at most $\text{gap}_d/2$ from the midpoint regardless of input, so the tight-dimension contribution spans an interval of width $\leq \sum_{d\in\text{tight}} |w_d| \cdot \text{gap}_d$. The loose-dimension swing is bounded by $\sum_{d\in\text{loose}} |w_d|$ (since $|h_d| \leq 1$). If the minimum tight-dimension contribution exceeds the maximum loose-dimension swing, the classifier output is positive for all states within the bounding box defined by the gap ranges, certifying confinement when tight dimensions remain within these bounds under switching (verified empirically and rigorously by the FI box check below). This is a weight-only certificate requiring no test data (though identifying which dimensions are tight requires running the dynamics on a reference input).

Causal validation. Extending the fixed-point analysis of Sussillo and Barak (2013) to switching dynamics, we use surgical clamping to provide causal evidence: of 10 failure models, clamping one rogue dimension raises persistence $0.0 \rightarrow 1.0$ in 6; of 10 success models, clamping one confining dimension breaks 3 ($1.0 \rightarrow <0.5$). Clamping $h_d = c$ also drives other dimensions via $W_{hh}[:, d] \cdot c$; the bidirectional results mitigate but do not eliminate this confound. Neural network verification methods (Singh et al., 2019; Katz et al., 2019) inspire our certificate (Figure 1, Table 1), though our condition is specialised to recurrent dynamics.

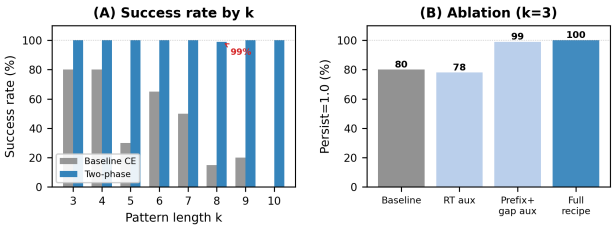

Figure 2. Illustrative. **(A)** Benchmark across k=3–10: $\geq$99/100 with two-phase. **(B)** Ablation confirms each theory-derived component is necessary.

Table 2. Master evaluation across settings and architectures.

Setting	Arch	Baseline	Two-phase	Cert
k=3	Elman	80%	**200/200**	100/100
k=5	Elman	30%	**200/200**	20/20
k=8	Elman	15%	99/100	20/20
k=12*	Elman	0%	**20/20**	20/20
k=3	GRU	15%	16/20	1/20
k=3	LSTM	15%	17/20	0/20
Dyck-2	Elman	**100%**	—	—
Regex (42)	Elman	varies	**350/350**	349/350

*Blind: k=12 never used during development.

3. From Certificate to Recipe to Benchmark

The certificate identifies *what* makes models generalise; inverting it yields a recipe that *causes* generalisation. Three auxiliary losses target the certified geometry:

- **Gap loss**: drives input-invariance in tight dimensions ($\mathcal{L}_{\text{gap}} = \lambda \sum_d |w_d| \cdot (h^\infty_{0,d} - h^\infty_{1,d})^2$). The loss targets constant-input fixed points as a proxy; minimising fixed-point divergence encourages but does not formally guarantee orbit confinement under mixed inputs.
- **Projection loss**: ensures the margin is positive
- **Random-trail loss**: enforces confinement under random input continuations

Two-phase training (cross-entropy + auxiliary, then auxiliary-only) prevents CE from interfering with geometric corrections.

Blind prediction: the loop's strongest test. The k=12 experiment was designed *before* any training run: theory predicted the recipe would succeed because the gap suppression mechanism is k-agnostic, and benchmarks confirmed it exactly (20/20 vs 0/20 baseline). This is theory–benchmark coupling at its tightest—a quantitative prediction made in

Table 3. Ablation at k=3: removing any theory-derived component produces the predicted failure.

Recipe	Grokking	Persist = 1.0	Theory-predicted failure
Baseline CE	$\sim$80/100	$\sim$80/100	No confinement signal
+ RT aux	87/100	78/100	Persist comparable to baseline
Prefix + gap aux	100/100	99/100	1 wrong half-space
Full two-phase	**100/100**	**100/100**	—

advance, then verified.

3.1. Ablation: Each Component Targets a Theoretical Requirement

Each ablation produces the failure mode predicted by the theory (Figure 2, Table 3): without gap loss, dimensions remain loose; without projection loss, the margin can be negative; without random-trail loss, confinement is not enforced under arbitrary continuations. Since the losses were designed to target the theory's features, the ablation validates the recipe design more than the mechanism itself; the blind k=12 prediction provides the stronger independent test.

3.2. Cross-Architecture and Regex Transfer

The training recipe partially transfers beyond Elman RNNs (Table 2)—GRU: $3/20 \rightarrow 16/20$ (5.3×); LSTM: $3/20 \rightarrow 17/20$ (5.7×); the GRU/LSTM difference is not significant at N=20; PyTorch replication: 99/100—though the margin certificate does not (1/20 GRU, 0/20 LSTM). Dyck-2 (context-free) generalises at 16× training length (20/20 with standard training, no recipe needed); we confirm this generalisation but do not claim the input-invariance mechanism explains the context-free dynamics. The recipe extends to 42 absorbing-state regex patterns of the form $p_1(.*)p_2$ (DFA 3–11 states): 350/350 Elman models generalise, **349/350 margin-certified** (0 FP). The tight-dimensions diagnostic predicts generalisation across architectures ($r = +0.59$ Elman, $+0.31$ GRU, $+0.23$ LSTM), with the declining correlation suggesting gated architectures achieve confinement through different mechanisms.

3.3. Where the Loop Breaks

The most informative aspect of theory–benchmark coupling is identifying exactly where the theory's scope ends. Four systematic breakdowns each reveal a distinct limitation, pointing the loop toward its next iteration.

Non-absorbing DFAs. The gap suppression mechanism requires a post-detection absorbing state: once the pattern is detected, the system must reach a region of state space that is invariant to further input. For non-absorbing languages (modular counting, last-k, alternating), no such attractor exists. Theory predicts failure, and benchmarks confirm: 0% two-phase success across all non-absorbing tasks. Critically, the *baseline* achieves 80–100% on these same tasks—the theory correctly identified why the tanh-saturation mechanism is wrong here, not that the tasks are intrinsically hard.

Gated architectures. The margin certificate does not transfer: only 1/20 GRU and 0/20 LSTM models are certified, even though the recipe partially works (16/20 and 17/20). The gap widens on regex patterns: GRU achieves only 8/60 (baseline 7/60) and LSTM 0/60. The diagnostic correlation decay ($+0.59 \rightarrow +0.31 \rightarrow +0.23$) quantifies the divergence: gated architectures may achieve confinement through gate-mediated information suppression rather than tanh saturation, requiring a different certificate.

Basin-of-attraction effects at k=5. Theory predicts that models with many tight dimensions and positive margin should generalise. Yet 5 models at k=5 satisfy both conditions but fail within 27–82 steps: the basin of attraction around the certified region is too narrow for the actual post-detection state distribution. The theory's static margin analysis misses this dynamical subtlety.

Minimal-capacity false positives. At k=6, H=11, the margin condition certifies 21/30 models, yet 0/30 generalise. The certificate assumes sufficient capacity ($H \geq 4k$) to realise the required geometry; when this assumption is violated, the margin is formally positive but the confinement region is too small. At $H \geq 4k$, the certificate maintains 0 FP.

Each breakdown tells the theory exactly what to address next: absorbing-state dependence, gate-specific certificates, dynamical basin analysis, and capacity-aware bounds.

4. Discussion: Lessons for Theory–Benchmark Coupling

Capacity as a theory–benchmark lesson. Theory predicted that hidden-state capacity determines success; benchmarks confirmed that the binding constraint is H, not training time. Extending phase 2 from 10,000 to 20,000 epochs reproduces identical failures; raising H from $\sim 4k$ to $\sim 5k$ resolves k=9 and k=10. At k=3, H=32 (overcapacity), baseline achieves only 22/30 while the recipe achieves 50/50, confirming that mechanism-targeting losses constrain geometry regardless of excess capacity.

Dead ends as contributions. The loop identifies dead ends early, which is itself a contribution. We tested and rejected spectral radius, eigenvalue danger, DFA extraction accuracy, and accept-state saturation—none predicts generalisation. These failures demonstrate that confinement is a multi-step property of switched dynamics, not a single-

step or linearised property. Without the formal framework, each plausible diagnostic could have consumed months of empirical exploration.

Scope of the theory. The theory is descriptive, not predictive of which weights emerge during training: per-dimension init $\rightarrow$ tight correlations are < 0.07; per-model $|r| < 0.20$. The recipe's hyperparameters are tuned for $[1]^k$; other absorbing DFAs with multiple accept states or sparse acceptance may require re-tuning. The FI box certificate is conservative (65/100 two-phase vs 100/100 for margin alone).

The methodology generalises. The loop operates at multiple levels of validation: the margin certificate (100/100, 0 FP), the recipe ($\geq$99/100 across k=3–10, 12, 200/200 at k=3–5), regex generality (349/350 certified, 0 FP across 42 patterns, DFA 3–11), and the blind k=12 prediction. The breakdowns (§3.3) are equally informative, delineating exactly where new theory is needed. The core pattern—derive a mechanism, certify it from weights, invert the certificate into a recipe, validate at scale, and learn from the failures—transfers beyond this specific setting.

References

A. Power et al. Grokking: Generalisation beyond overfitting on small algorithmic datasets. In *ICLR 2022 MATH-AI Workshop*, 2022.

G. Delétang et al. Neural networks and the Chomsky hierarchy. In *ICLR*, 2023.

N. Nanda et al. Progress measures for grokking via mechanistic interpretability. In *ICLR*, 2023.

D. Liberzon. *Switching in Systems and Control*. Birkhäuser, 2003.

D. Sussillo and O. Barak. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Computation*, 25(3):626–649, 2013.

G. Singh et al. An abstract domain for certifying neural networks. In *POPL*, 2019.

G. Katz et al. The Marabou framework for verification and analysis of deep neural networks. In *CAV*, 2019.

M. Barnsley. *Fractals Everywhere*. Academic Press, 1988.