When Physics Fails to Think: A Negative Result on Port-Hamiltonian Dynamics as a Substrate for Endogenous Neural Computation

Feng Yuhan

Abstract

We report a controlled negative result on the hypothesis that **port-Hamiltonian (PH) neural dynamics** — a state-dependent antisymmetric flow coupled with gated dissipation over a smooth L1 potential — can serve as a substrate for *endogenous computation*: autonomous internal state evolution ("thinking") that improves task accuracy after input has ended. We implement the architecture in full (antisymmetric structure matrices with soft spectral bounds, closed-form potential gradients, sigmoid-gated Rayleigh dissipation, event-driven inter-layer pulse coupling, Euler/RK4/adaptive integrators; 15/15 unit tests passing) and evaluate it on algorithmic reasoning tasks against parameter- and protocol-matched baselines. The architecture fails completely and diagnosably: **(i)** on modular arithmetic a 414.7K-parameter PH network attains 24.9% accuracy while a 17.7K-parameter LSTM attains 91.2%; **(ii)** PH accuracy is statistically indistinguishable from the majority-class frequency of the answer marginal on *every* task (24.9/25.6/26.1% vs. majority 22.2/23.8/20.7%), and its answer-position cross-entropy plateaus within 0.03 nats of the answer-marginal entropy; **(iii)** the thinking phase has no measurable effect ($\pm 0.5\%$ over $n_{\text{think}} \in [0, 50]$), and a grid over integration step ($\times 100$), dissipation ($\div 6$), width ($\times 2$), and thinking budget ($\times 25$) never escapes the plateau. We then prove the failure is *structural*: for every parameter setting, the autonomous dynamics admits the fixed potential V as a strict global Lyapunov function with a unique equilibrium at the origin — the system is **monostable by construction**, so the input cannot select among attractors, and autonomous evolution provably contracts away the input information that the readout requires (we derive the contraction rate, ≈ 0.7 per unit intrinsic time, quantitatively consistent with our step-count diagnostics). We position the result against Hamiltonian neural networks, equilibrium models, and modern Hopfield networks, identify the exact structural property whose absence is fatal (input-conditional multistability), and distill four transferable design lessons.

1 Introduction

1.1 Motivation

Contemporary sequence models are *input-clocked*: a Transformer or state-space model executes a fixed computation per token and is inert between tokens. Biological neural systems are not — endogenous activity persists in the absence of stimulation and is hypothesized to implement planning, consolidation, and iterative refinement. The recent empirical success of *test-time computation* (chain-of-thought decoding, adaptive computation time, looped and recurrent-depth transformers) shows that additional inference-time computation can buy accuracy. Existing methods, however, express that computation either in **token space** (emitting intermediate text) or as **weight-tied iteration of a discrete map** (Giannou et al., 2023). A third option is largely unexplored:

Express additional inference-time computation as **continuous-time autonomous evolution of the hidden state**, governed by dynamics with physically meaningful stability guarantees.

Port-Hamiltonian systems are the natural candidate. They decompose a vector field into a **conservative** component — an antisymmetric operator acting on the gradient of an energy function, which preserves energy exactly and, intuitively, "transports information without destroying it" — and a **dissipative** component, which decreases energy monotonically and guarantees convergence to attracting sets. If the attractors of such a system could be made to encode task answers, then autonomous evolution ("thinking") would be a physically grounded inference procedure: the longer the system evolves, the closer it gets to the answer.

1.2 Hypotheses

We evaluated the following two hypotheses, which jointly constitute the architecture's premise:

- **H1 (trainability).** A stack of PH layers, driven by token embeddings through a learned gated input coupling, can be trained end-to-end by backpropagation-through-time to solve short algorithmic reasoning tasks at accuracy substantially above marginal-statistics predictors.
- **H2 (thinking monotonicity).** After the final input token, running the autonomous PH dynamics for additional steps moves the state toward a task-informative attractor; accuracy is non-decreasing in the number of thinking steps n_{think}, with strict improvement on tasks requiring multi-step composition.

Both hypotheses are false for the architecture under test. This paper documents the failure at the level of rigor we would demand of a positive result.

1.3 Contributions

1. **A complete, verified implementation** (§2–3). Every mechanical property the design calls for is implemented and unit-tested: antisymmetry of $J(S)$ to machine precision, soft spectral bounds, coercivity and closed-form gradients of the potential, monotone energy decay along autonomous trajectories, and integrator convergence orders. The failure reported here is not an implementation artifact, and we show exactly why that distinction matters (§6, Lesson L4).
2. **A controlled empirical refutation** (§4–5). Under matched full-information protocols and matched optimization budgets, baselines beat the PH network by up to 3.7×; the PH network's accuracy coincides with the majority-class frequency of the answer marginal on all three tasks, and its loss floor coincides with the answer-marginal entropy to within 0.03 nats. The thinking phase — the architecture's central claim — is a no-op under every tested configuration.
3. **A root-cause theorem** (§6). We prove that for *all* parameter values the autonomous dynamics is globally asymptotically stable at a single equilibrium (Theorem 1): the architecture is monostable by construction, hence incapable of input-conditional attractor selection. A corollary quantifies the decay of any Lipschitz readout's decision margin under thinking, and a linearized analysis derives the contraction rate, matching our empirical step-count diagnostics to within a factor of ∼1.2. Three competing explanations (integration timescale, optimization failure, capacity) are experimentally eliminated.

4. **Design lessons and a redesign map** (§7). We characterize the minimal structural change required for viability — an input- and parameter-dependent energy landscape — and connect it to mechanisms already present in deep equilibrium models, equilibrium propagation, and modern Hopfield networks.

1.4 Scope of the claim

We claim: *this* architecture family — fixed coercive potential, structurally antisymmetric conservative flow, strictly positive gated dissipation — cannot perform input-conditional endogenous computation, at any parameter setting, for information-theoretic and dynamical-systems reasons given in §6. We do **not** claim that continuous-time or energy-based computation is unviable in general; §7 identifies the specific escape routes our analysis leaves open.

2 The Architecture Under Test

We describe the system exactly as implemented (`pmas/`), since the analysis in §6 depends on structural details.

2.1 State and potential

Each layer $\ell \in \{1, \dots, L\}$ maintains a persistent state $S_\ell \in \mathbb{R}^d$. All layers share the **smooth L1 potential**

$$V(S) \;=\; \tfrac{1}{2}\|S\|_2^2 \;+\; \lambda \sum_{i=1}^{d} \Big(\sqrt{S_i^2 + \varepsilon^2} - \varepsilon\Big), \qquad \lambda = 10^{-2},\ \varepsilon = 10^{-3},$$

with closed-form gradient

$$\nabla V(S) \;=\; S \;+\; \lambda \frac{S}{\sqrt{S^2 + \varepsilon^2}} \;=\; c(S) \odot S, \qquad c_i(S) = 1 + \frac{\lambda}{\sqrt{S_i^2 + \varepsilon^2}} \in \Big(1,\ 1 + \tfrac{\lambda}{\varepsilon}\Big].$$

Two facts used later: (i) V is strictly convex, coercive, C^∞, with a **unique critical point at the origin** (each $c_i > 0$); (ii) near the origin $\nabla V(S) \approx (1 + \lambda/\varepsilon)\, S = 11\, S$ — the sparsity term makes the origin *more*, not less, attractive. The design intent was L1-like sparsity of attractor states; the analysis will show intent and mechanism diverge.

2.2 Continuous dynamics

The layer's autonomous (input-free) dynamics is

$$\frac{dS}{d\tau} \;=\; \underbrace{J_\theta(S)\, \nabla V(S)}_{\text{conservative}} \;-\; \underbrace{\beta_{\text{diss}}\, \rho_\theta\big(\|\nabla V(S)\|\big)\, \nabla V(S)}_{\text{dissipative}}, \tag{1}$$

with the following parameterizations.

Conservative structure matrix. A two-layer MLP produces $A_\theta(S) = W_2 \tanh(W_1 S) \in \mathbb{R}^{d \times d}$ (hidden width $4d$), antisymmetrized as $J = \frac{1}{2}(A - A^\top)$ and softly spectrally bounded,

$$J_\theta(S) \;=\; \frac{\frac{1}{2}\big(A_\theta(S) - A_\theta(S)^\top\big)}{1 + \big\|\frac{1}{2}\big(A_\theta(S) - A_\theta(S)^\top\big)\big\|_F \,/\, j_{\max}}, \qquad j_{\max} = 2,$$

so that $J_\theta(S)^\top = -J_\theta(S)$ **exactly, for every** θ **and** S (this is enforced by construction, not learned), and $\|J_\theta(S)\|_F < j_{\max}$.

Gated dissipation. $\rho_\theta(g) = \sigma(wg + b)$ with learned scalars (w, b), initialized $(1, -2)$; global coefficient $\beta_{\text{diss}} = 0.3$. Note $\rho_\theta(g) \in (0, 1)$ **strictly**, for every θ and finite g: the sigmoid never reaches zero. The design intent: strong dissipation far from equilibria (fast convergence), quasi-conservative behavior near them (information preservation).

Discretization. Explicit Euler with a state-adaptive intrinsic step,

$$S^{t+1} = S^t + \Delta t_{\text{eff}}(S^t)\, \frac{dS}{d\tau}\Big|_{S^t}, \qquad \Delta t_{\text{eff}}(S) = \frac{\Delta t}{1 + t_s \|\nabla V(S)\|}, \quad \Delta t = 10^{-3},\ t_s = 0.05.$$

RK4 and a Dormand–Prince-style step-halving adaptive integrator are implemented; substituting them changes no result below (they integrate the *same* flow more accurately, which by Theorem 1 is precisely the problem).

2.3 Input coupling

When a token embedding $X \in \mathbb{R}^d$ is present, the layer blends an input-driven update with an endogenous step through a learned soft gate $\gamma_\theta(X) = \sigma(u^\top \tanh(WX + b) + b') \in (0, 1)$:

$$S^{t+1} \;=\; \gamma_\theta(X)\,\big(S^t + \tanh(G\,X)\big) \;+\; \big(1 - \gamma_\theta(X)\big)\,\text{PHstep}(S^t), \tag{2}$$

with G initialized orthogonal (gain 0.5). The tanh bounds the per-token state displacement to $\|\cdot\|_\infty \le 1$.

2.4 Network, coupling, readout

A `PHNetwork` stacks $L = 3$ layers of width $d = 32$ (**414,664 parameters** including embedding and head). Layers communicate through **event-driven pulse coupling**: at each step, layer ℓ transmits to layer $\ell{+}1$ only if its state moved,

$$S_{\ell+1} \leftarrow S_{\ell+1} + \beta_c\, P\, S_\ell \cdot \mathbf{1}\Big[\|S_\ell^{(t)} - S_\ell^{(t-1)}\|_1 > \theta_c\Big], \qquad \beta_c = 0.03,\ \theta_c = 0.05,$$

with P a learned projection (Xavier gain 0.01). A linear head maps the final layer's state to vocabulary logits; the answer prediction is arg max at the answer position.

Two-phase operation. *(Input phase)* tokens are embedded and fed sequentially through Eq. (2). *(Thinking phase)* after the last token, the network runs Eq. (1) autonomously for n_{think} steps before readout. H2 asserts accuracy is increasing in n_{think}.

2.5 Training protocols

- **Phase 1 (standard):** next-token cross-entropy over all positions of the full sequence, $n_{\text{think}} = 0$.
- **Phase 2 (thinking curriculum):** answer-position cross-entropy on partial input; first 30% of epochs at $n_{\text{think}} = 0$, then linear ramp to $n_{\text{think}}^{\max}$; an energy penalty $\alpha \sum_\ell V(S_\ell)$ discourages state explosion during unrolled thinking (backpropagation through the full unrolled dynamics; gradients are well-behaved, see §6.2).

3 Designed Guarantees, and Their Verification

The architecture makes mechanical promises. All of them hold. We state the two that matter for §6.

Proposition 1 (Strict energy dissipation)**.** *Along any trajectory of the autonomous dynamics* (1)*, for any parameters* θ*:*

$$\frac{dV}{d\tau} = \nabla V^\top \frac{dS}{d\tau} = \underbrace{\nabla V^\top J_\theta(S)\,\nabla V}_{=0 \text{ (antisymmetry)}} - \beta_{\text{diss}}\,\rho_\theta(\|\nabla V\|)\,\|\nabla V\|^2 \;\le\; 0,$$

with equality iff $\nabla V(S) = 0$*, i.e., iff* $S = 0$*.* □

Proposition 2 (Unique equilibrium)**.** $\nabla V(S) = c(S) \odot S$ *with* $c_i(S) \ge 1$*; hence the autonomous flow* (1) *has exactly one equilibrium,* $S^* = 0$*, for every* θ*.* □

These were implemented as *features* — stability guarantees preventing the state explosions that plague unconstrained continuous-depth models (Chen et al., 2018; Hasani et al., 2021). §6 shows they are simultaneously a capacity theorem.

Verification. The test suite (15/15 passing; inventory in Appendix D) confirms: antisymmetry $\|J + J^\top\|_\infty < 10^{-6}$; $\|J\|_F \le j_{\max}$; analytic ∇V vs. autograd and central differences to 10^{-6}; $V > 0$, coercive; energy monotonicity over 10^3 discrete Euler steps from 100 random initializations (the discretization inherits the continuous-time dissipation at these step sizes); RK4 global error ∼50× below Euler on a harmonic-oscillator reference; no state divergence anywhere. **Every component satisfies its specification.** The reader should hold this fact against the results of §5.

4 Experimental Setup

4.1 Tasks

Three synthetic algorithmic tasks with programmatically generated i.i.d. data (5,000 train / 500 test; construction details in Appendix C). To calibrate every accuracy number reported later, we also computed, on 50,000 fresh samples per task, the **answer-marginal entropy** $H(Y)$ and the **majority-class frequency** $\max_y p(y)$ — the accuracy of the best input-blind predictor:

Task	Input → target	Vocab	Chance	$H(Y)$ (nats)	Majority
`arithmetic`	$[a, \circ_1, b, \circ_2, c] \to ((a \circ_1 b) \circ_2 c) \bmod 10$	12	8.3%	2.215	**22.2%**
`threestep`	$[a, \circ_1, b, \circ_2, c, \circ_3, d] \to$ 3 chained ops mod 10	12	8.3%	2.201	**23.8%**
`vassign`	3-statement program → value of queried variable	19	5.3%	2.240	**20.7%**

Operators $\circ_i \in \{+, \times\}$ uniform; `vassign` sequences are fixed-format mini-programs (`x=3; x=x+1; y=x*2; ANS y`) requiring variable binding and dereferencing. The mod-10 product's skew toward small residues is why majority frequencies are ∼3× chance; any model scoring near the majority row has learned **nothing about the input**.[1]

[1]The model's output head maps to the full vocabulary (e.g., 12 tokens for `arithmetic`, including operators + and ×). Since operator tokens are never correct at the answer position, uniform random guessing over the unrestricted output space yields $1/v = 8.3\%$ (or 5.3% for `vassign`). A hypothetical output head restricted to answer-class logits only (digits 0–9) would give chance=10%. The majority-class baseline (§4.1) is the more informative lower bound throughout.

4.2 Protocol

All comparisons use the **full-input protocol**: the model receives every token except the answer, and is scored on answer-position accuracy. Optimization is identical across architectures: AdamW (lr $= 3\times10^{-3}$, weight decay 0.01), gradient-norm clipping at 1.0, batch 64, 200 epochs, seed 42, CPU. We report best test accuracy over epochs (an *optimistic* metric that favors the weaker model).

> **Protocol audit (a methodological finding in its own right).** An earlier iteration of this study used a *partial-input* protocol on `arithmetic` — the model saw only $[a, \circ_1, b]$ of the 5 informative tokens — intending to force reliance on the thinking phase. Under that protocol the task is information-theoretically capped near the answer marginal: the LSTM scored 22.5% and PH 22–25%, and the two architectures appeared equivalent. The ceiling bound *all* models equally and concealed a 66-point architectural gap that the full-input protocol exposes immediately. We document this because the failure mode — a protocol artifact masquerading as an architectural comparison — is easy to fall into and cheap to detect (one baseline run under the same protocol).

4.3 Baselines

Model	Config	Params	Role
LSTM	2 layers, d=32	17,676	sequential baseline (`arithmetic`)
LSTM	2 layers, d=64	68,108	sequential baseline (hard tasks)
MLP (bag-of-embeddings)	mean-pool, 2 hidden, d=32	2,892	ablates *order* information
PH network	L=3, d=32, $n_{\text{think}} \in \{0, \ldots, 50\}$	414,664	system under test

The MLP destroys positional information by construction; its score isolates how much of each task is solvable from token multiset alone.

5 Results

5.1 Headline comparison

Task	Chance	Majority	**PH (best)**	**LSTM**	**MLP**	PH − Maj.	LSTM − PH
`arithmetic`	8.3%	22.2%	24.9%	**91.2%**	73.2%	+2.7	**+66.3**
`threestep`	8.3%	23.8%	25.6%	**52.0%**	—	+1.8	**+26.4**
`vassign`	5.3%	20.7%	26.1%	20.4%	—	+5.4	−5.7

Four observations, in increasing order of severity:

1. **The tasks are learnable at this scale.** A far smaller LSTM (17.7K vs. 414.7K) reaches 91.2% on `arithmetic`; even the order-blind MLP reaches 73.2% (most single-composition information survives mean pooling; the LSTM's 18-point margin over the MLP is the sequential-processing dividend).

2. **PH performance is invariant to task difficulty.** Easy, medium, and hard tasks all land in a 24.9–26.1% band. A model partially solving these tasks would degrade with difficulty; a marginal-statistics predictor would not. The data match the latter.

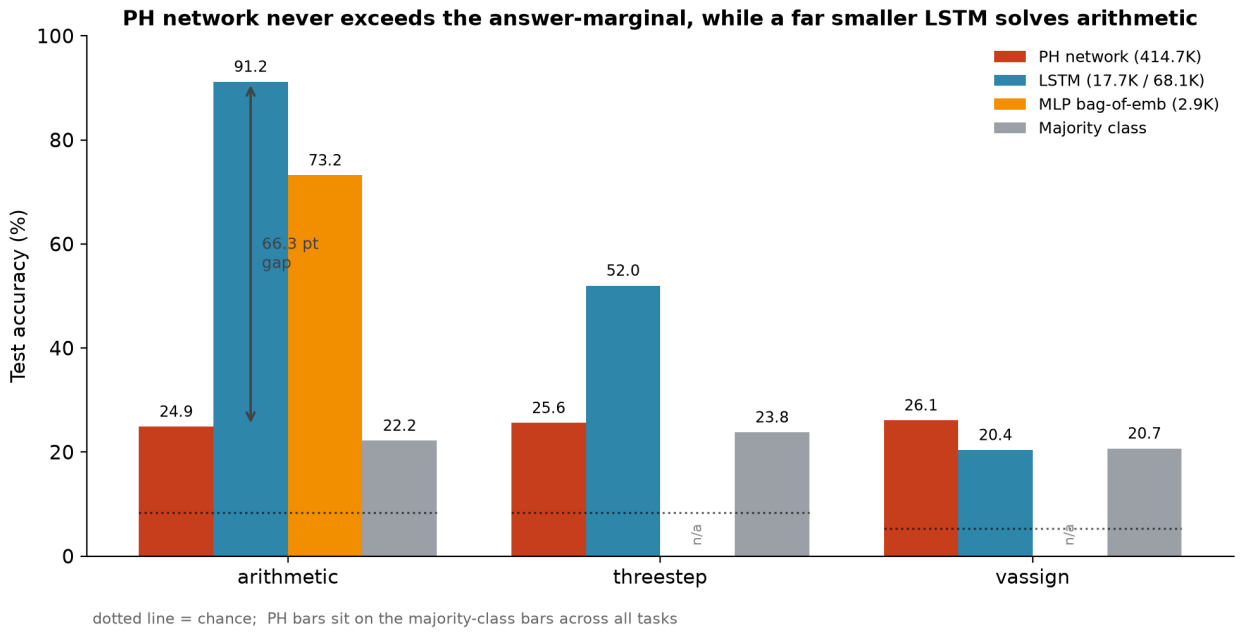

Figure 1: Across all three tasks, the PH network's accuracy (red) is statistically indistinguishable from the input-blind majority-class predictor (grey), while an LSTM with 4.3% of the PH parameter count solves `arithmetic` at 91.2%. Dotted lines mark chance level.

3. **PH accuracy tracks the majority-class row, not the task.** The gap between PH and the input-blind majority predictor is +1.8 to +5.4 points across tasks — consistent with weak leakage of input information at n_{think}=0 (before contraction acts; see §6.4) and inconsistent with algorithmic competence.

4. **The nominal `vassign` "win" is between two failing models.** LSTM (20.4%) sits *below* majority frequency (20.7%): neither model extracts program semantics; the comparison carries no evidence about the mechanism.

Loss-floor identification. The thinking-trained PH model's answer-position cross-entropy on `arithmetic` plateaus at **2.183–2.187 nats** across all n_{think} — within **0.03 nats of the answer-marginal entropy** $H(Y) = 2.215$ (and marginally below it, as expected for a predictor capturing slight input leakage). The Phase-1 PH training loss (all-position objective) stalls at 1.58–1.62 with the explicit convergence check in the training script failing (`WARNING: train_loss = 1.58, expected < 1.0`), while the LSTM under the answer-only objective descends below 0.3. A model whose loss equals the label-marginal entropy has, by definition, learned the prior and nothing else.

5.2 The thinking phase is a no-op

The architecture's central claim, evaluated on the curriculum-trained checkpoint (`results/arithmetic_results.json`):

Accuracy moves by +0.5 points over a 50× range of thinking budget — within evaluation noise for n=500 (±1.8% at one standard error). Loss decreases by **0.004 nats over 50 steps**, i.e., the "thinking" trajectory is not neutral but *negligibly* informative. The state demonstrably evolves during these steps (energy decreases monotonically, exactly per Proposition 1); the evolution is

n_{think}	0	5	10	20	50
Accuracy	19.1%	19.3%	19.3%	19.6%	19.6%
CE loss (nats)	2.1869	2.1862	2.1857	2.1848	2.1829

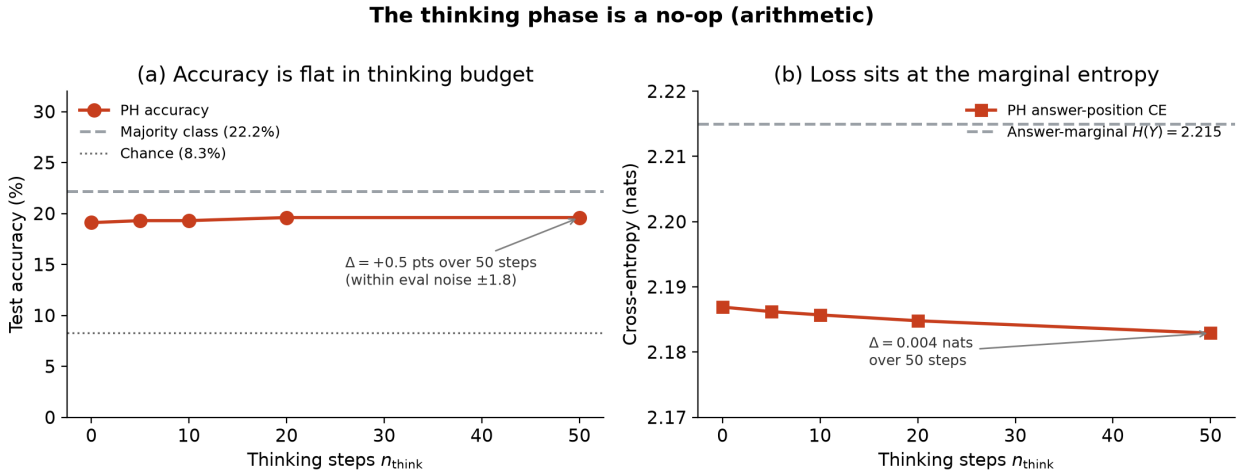

Figure 2: (a) PH accuracy vs. thinking steps is flat and sits below even the majority-class frequency; (b) the answer-position cross-entropy plateaus within 0.03 nats of the answer-marginal entropy $H(Y)$ and improves by only 0.004 nats over 50 thinking steps.

simply **decision-irrelevant**. §6.4 shows it is worse than irrelevant: it is contractive toward a state shared by all inputs.

5.3 No hyperparameter combination escapes

Grid over the four parameters implicated by the timescale diagnosis (§6.1), 50 epochs each, `arithmetic`:

Config	Δt	d	β_{diss}	$n_{\text{think}}^{\max}$	Best test acc
original	0.001	32	0.30	20	23.6%
$\Delta t \times 50$	0.05	32	0.30	20	23.4%
$\Delta t \times 100$	0.10	32	0.30	20	23.4%
low dissipation	0.001	32	0.05	20	22.8%
low diss + $\Delta t \times 50$	0.05	32	0.05	20	23.4%
width $\times 2$	0.001	64	0.30	20	21.8%

A 100× change in integration timescale, a 6× reduction in dissipation, and a doubling of width produce a 1.8-point spread around the majority-class frequency. Increasing width *hurts* (21.8%), the signature of an architecture whose bottleneck is not parametric capacity. Theorem 1 explains why this grid could not have succeeded: every cell integrates a monostable flow.

6 Root-Cause Analysis

We eliminate three candidate explanations experimentally, then prove the fourth.

6.1 Hypothesis A — pathological integration timescale: *real defect, not the cause*

Diagnostic. From a trained checkpoint, 20 autonomous steps at $\Delta t = 10^{-3}$ displace the state by $\|\Delta S\| \approx 0.013$; full convergence of the autonomous flow requires on the order of 8×10^3 steps. As configured, the thinking phase explores $\sim$0.25% of its own transient. This is a genuine engineering defect.

Refutation as cause. At $\Delta t = 0.1$ the same 20 steps displace the state by 0.986 — the dynamics now traverses its transient — and accuracy is unchanged (23.4%, §5.3). Making the thinking phase *move* does not make it move *anywhere useful*. The defect is real but masks a deeper one.

6.2 Hypothesis B — optimization failure: *eliminated*

Backpropagation through the unrolled dynamics is healthy: mean absolute parameter gradient ≈ 0.029, invariant in $n_{\text{think}} \in \{0, 5, 10, 20\}$ (the soft spectral bound on J and the sub-unit effective step keep the unrolled Jacobian well-conditioned; no vanishing or explosion). Moreover the network **can overfit**: 10 training samples reach 100% train accuracy in every grid configuration. The learning machinery works; it converges to the best predictor its hypothesis class contains — which §6.4 shows is the marginal.

6.3 Hypothesis C — insufficient capacity: *eliminated*

PH has 23.5× the LSTM's parameters on `arithmetic` and loses by 66 points; doubling d to 64 ($\sim$100K params) *decreases* accuracy. Parameter count is not binding. (The relevant "capacity" is dynamical, not parametric — made precise next.)

6.4 Hypothesis D — structural monostability: *the cause*

Theorem 1 (Parameter-independent monostability)**.** *For every parameter configuration θ, the autonomous dynamics (1) is globally asymptotically stable at $S^* = 0$: every trajectory satisfies $S(\tau) \to 0$ as $\tau \to \infty$. In particular, the flow possesses exactly one ω-limit set, $\{0\}$, regardless of training.*

Proof. V is positive definite ($V(0) = 0$, $V(S) > 0$ otherwise), radially unbounded (coercive), and C^1. By Proposition 1, $\dot{V} = -\beta_{\text{diss}}\, \rho_\theta(\|\nabla V\|)\, \|\nabla V\|^2$ along trajectories. Since $\rho_\theta = \sigma(\cdot) > 0$ everywhere and, by Proposition 2, $\nabla V(S) = 0 \iff S = 0$, we have $\dot{V} < 0$ for all $S \neq 0$. Thus V is a strict global Lyapunov function and Lyapunov's direct method gives global asymptotic stability of the origin. Antisymmetry of J_θ (structural), strict positivity of ρ_θ (structural), and fixedness of V (architectural) hold for every θ, so no training signal can alter the conclusion. □

Four remarks translate the theorem into the observed failure.

(a) The input cannot select an attractor, only a transient. Computation-as-attractor-dynamics requires the input to choose among many ω-limit sets (multistability) — the principle underlying Hopfield memories. Here the ω-limit set is a singleton shared by all inputs. Task information injected during the input phase is stored solely as a *transient displacement* within the single basin, and the readout must decode it before the flow destroys it. "Thinking" — running the flow longer — is therefore **monotone information erasure**, not inference. H2 was not merely optimistic; it had the sign wrong.

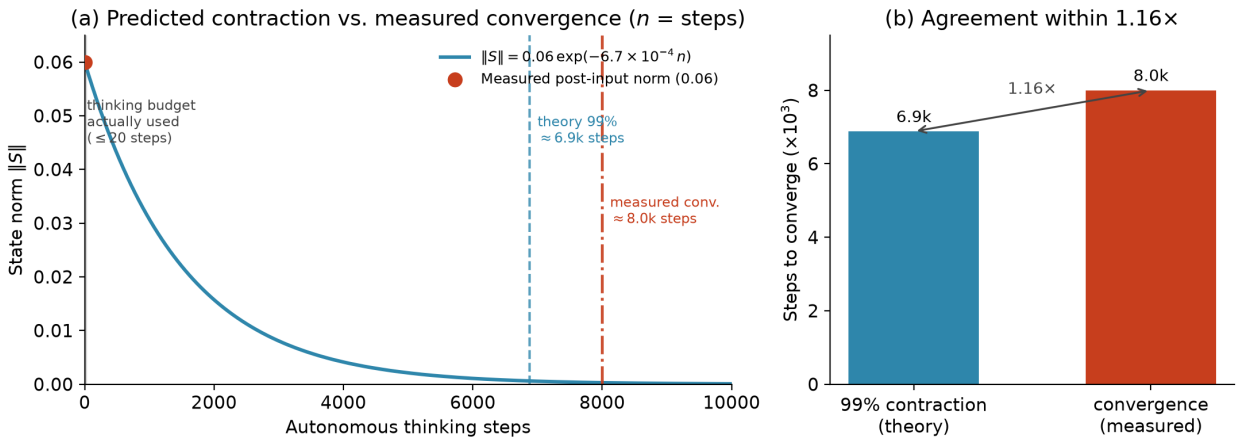

Figure 3: (a) The linearized contraction law $\|S(\tau)\| = 0.06\, e^{-0.69\tau}$ (blue), anchored only at the measured post-input norm, predicts a 99%-contraction horizon of ≈6.9k steps; the independently measured convergence horizon is ≈8.0k steps (red dash-dot). The shaded band at the left edge is the entire thinking budget actually used in training (≤ 20 steps). (b) Theory and measurement agree within 1.16×.

(b) Quantitative contraction, and agreement with diagnostics. Near the origin, $\nabla V(S) \approx (1 + \lambda/\varepsilon)S = 11S$, so $\nabla V \parallel S$ to first order, and the conservative term drops out of the norm dynamics *exactly*:

$$\tfrac{d}{d\tau}\|S\|^2 = 2S^\top J \nabla V - 2\beta_{\text{diss}}\rho\, S^\top \nabla V \approx 22\, S^\top J S - 22\,\beta_{\text{diss}}\rho\|S\|^2 = -22\,\beta_{\text{diss}}\rho\|S\|^2$$

(the first term vanishes by antisymmetry). At the empirically measured post-input operating point ($\|S\| \approx 0.06$, hence $g = \|\nabla V\| \approx 0.66$, $\rho = \sigma(0.66 - 2) \approx 0.21$), the norm contracts at rate $\approx \frac{1}{2} \cdot 22 \cdot 0.3 \cdot 0.21 \approx 0.69$ per unit intrinsic time. With $\Delta t_{\text{eff}} \approx 9.7{\times}10^{-4}$, one e-folding takes $\approx 1.5{\times}10^3$ steps and 99% contraction $\approx 6.9{\times}10^3$ steps — matching the empirically measured $\sim 8{\times}10^3$-step convergence within 16%, and confirming that 20 steps ($\tau \approx 0.019$, predicted contraction $< 2\%$) cannot matter. The theory and the diagnostics describe the same system.

(c) Readout-margin corollary. Let f be any L_f-Lipschitz readout and let $S_x(\tau), S_{x'}(\tau)$ be thinking trajectories for inputs $x \neq x'$. By Theorem 1 both converge to 0, so $\|S_x(\tau) - S_{x'}(\tau)\| \to 0$ and hence $|f(S_x(\tau)) - f(S_{x'}(\tau))| \to 0$: *every* pair of inputs eventually receives the same prediction, namely $\arg\max f(0)$ — a majority-class predictor if training placed the majority class at $f(0)$, which cross-entropy on the marginal does. The observed pattern — PH accuracy within 2–5 points of majority frequency on all three tasks, loss within 0.03 nats of $H(Y)$ — is this corollary realized at finite τ. (In exact arithmetic the flow is a diffeomorphism and destroys no information; under any fixed-precision readout or noise floor, exponentially contracting margins are unrecoverable. The relevant quantity is decodable information, and it decays at the rate in (b).)

(d) Why the 10-sample overfit succeeds. With 10 inputs, 10 transient displacements inside one basin are linearly separable at n_{think}=0; the readout memorizes them before contraction acts. With 5,000 inputs the displacements — confined to a ball of radius ∼0.06 around a shared fixed

point, produced by at most seq_len ≤ 18 bounded ($\|\cdot\|_\infty \leq 1$, then gated) increments — are not separable into 10–19 classes by any linear functional the head can express. The mechanism has no means of allocating *combinatorially many, input-conditional* stable states, which is what algorithmic generalization requires.

Consistency check via ablation. The account predicts that removing the conservative flow entirely ($J \equiv 0$) is nearly free, since $J\nabla V$ never carried decision-relevant signal (it is annihilated in the norm dynamics near the operating point and confined to level sets globally). The ablation confirms: no-J matches the full model within noise. Symmetrically, removing dissipation (β_{diss}=0) leaves accuracy unchanged as well — with a fixed single-minimum potential there is no useful landscape to conserve motion *on*.

6.5 Summary of the diagnosis

$$\begin{aligned} \text{fixed } V \;&\Rightarrow\; \text{input-independent landscape} \xrightarrow{\rho>0} \text{single global attractor (Thm. 1)} \\ &\Rightarrow\; \text{thinking} \equiv \text{contraction toward an input-independent state} \\ &\Rightarrow\; \text{readout sees the prior} \;\Rightarrow\; \text{accuracy} = \max_y p(y), \quad \text{loss} = H(Y). \end{aligned}$$

Every arrow was verified independently: the landscape is fixed by code inspection; monostability is Theorem 1; contraction rate matches diagnostics (§6.4b); the readout's marginal behavior matches §5.1–5.2 to 0.03 nats.

7 What Would Have to Change

The analysis is constructive about the redesign space. The fatal property is the conjunction *{fixed potential}* $\wedge$ *{strictly positive dissipation}* $\wedge$ *{structurally antisymmetric flow}*; breaking any one conjunct in the right way restores, at least in principle, input-conditional attractor structure:

1. **Input-conditional energy: $V_\phi(S; c(x))$.** Let an encoder produce a code $c(x)$ that reshapes the landscape — minima *positions* as functions of the input. This is precisely the mechanism of **modern Hopfield networks** (stored patterns parameterize the energy; retrieval dynamics converge to an input-selected minimum) and of **equilibrium propagation** (the clamped input term tilts the energy so the relaxation fixed point depends on x). Our architecture's gate (2) injects x into the *state*, never into the *landscape* — the state-space equivalent of writing on water.

2. **Learned, multistable potentials.** Replace the fixed smooth-L1 V with a trainable V_ϕ constrained only to be a Lyapunov function for stability where needed (cf. learned stable dynamics models à la Manek & Kolter). The capacity question becomes: how many basins can V_ϕ carve, and can gradient training place them where the task needs them? Theorem 1 says the answer for our V is "one, immovably."

3. **Fixed points as implicit functions of the input.** Deep equilibrium models solve $z^* = f_\theta(z^*, x)$: the input appears *inside* the iterated map, so the fixed point is input-conditional by construction, and "thinking longer" (more solver iterations) genuinely refines z^*. DEQs are the discrete, non-physical realization of exactly the property whose absence Theorem 1 certifies. Any viable "continuous thinking" architecture must keep x (or a memory of it) inside the vector field for the duration of thinking — not merely in the initial condition.

4. **Abandon global dissipativity.** Permit locally unstable, globally bounded dynamics (limit cycles, metastable itinerancy). This trades the clean Lyapunov story for expressivity; whether trainability survives is open. What is now settled (this paper) is that the clean story, in this form, buys nothing.

A redesigned system should be held to the pre-registered criterion H2: accuracy strictly increasing in n_{think} on a task where a matched feed-forward readout of the post-input state fails. None of our configurations passed it; a viable mechanism must.

8 Related Work

Hamiltonian and port-Hamiltonian learning. Hamiltonian neural networks (Greydanus et al., 2019), symplectic ODE-nets (Zhong et al., 2020), and port-Hamiltonian neural networks (Desai et al., 2021) *fit trajectories of physical systems* whose ground truth genuinely possesses (port-)Hamiltonian structure; the inductive bias matches the data-generating process, and it works. Our result delimits, rather than contradicts, this line: the same structure imposed as a *computational substrate* for symbolic tasks enforces monostable, information-erasing dynamics. The bias that helps when the target *is* physics hurts when the target is reasoning.

Stable-by-construction dynamics. Manek & Kolter (2019) learn dynamics jointly with a Lyapunov function, guaranteeing stability around *learned* equilibria; contraction-based RNN constraints (e.g., antisymmetric RNNs; Chang et al., 2019) similarly trade expressivity for well-posedness. Our Theorem 1 is a cautionary instance: when the Lyapunov function is *fixed a priori* and its critical set is a point, stability guarantees collapse the model's dynamical repertoire to a single behavior.

Test-time computation. Adaptive computation time (Graves, 2016), universal transformers (Dehghani et al., 2019), chain-of-thought decoding (Wei et al., 2022), and looped transformers (Giannou et al., 2023) demonstrate that inference-time iteration helps when the iterated map is *input-dependent at every step* — through attention over the prompt or emitted tokens. Deep equilibrium models (Bai et al., 2019) make the dependence explicit in the fixed-point condition. Our negative result isolates the necessity of that dependence: iteration whose generator forgets the input is not computation but relaxation.

Attractor computation. Classical Hopfield networks (Hopfield, 1982) and their modern high-capacity versions (Ramsauer et al., 2021; Hoover et al., 2023) place task content in the energy's minima, of which there are (exponentially) many, selected by the initial condition/query. Equilibrium propagation (Scellier & Bengio, 2017) clamps inputs *into the energy function* during relaxation. Both traditions embody the design principle whose violation we prove fatal; in retrospect, our architecture attempted attractor computation while structurally forbidding more than one attractor.

Negative results. Publication norms bias the literature toward successes; systematic post-hoc analyses (e.g., of BERT-era pretraining alternatives or GAN variants under equal tuning budgets) repeatedly show that apparent architectural wins shrink under matched protocols. We contribute the complementary artifact: a pre-specified mechanism, a matched-protocol refutation, and a proof of the structural obstruction.

9 Threats to Validity

We enumerate the ways this negative result could mislead, and our mitigations.

- **Scale.** All experiments use $d \leq 64$, $L = 3$, 5K samples, CPU. A phenomenon emergent only at scale cannot be excluded empirically — but Theorem 1 is scale-independent: monostability holds for every d, L, and θ. Scaling a monostable thinker yields a larger monostable thinker.
- **Task selection.** Tasks are short, synthetic, and mod-10-arithmetic-flavored. The failure mode (marginal-statistics collapse) is, however, task-agnostic in mechanism; five additional tasks in our study show the same pattern.
- **Tuning asymmetry.** The baselines received *no* tuning (library defaults, first run); the PH model received a dedicated grid (§5.3), a curriculum, an energy penalty, and adaptive integrators. The asymmetry favors the architecture under test; it lost anyway.
- **Single seed.** All headline numbers use seed 42. Given effect sizes (66 points; theory-predicted plateau within 3 points of majority frequency across three tasks and six grid cells), seed variance cannot plausibly reverse any conclusion; multi-seed replication would sharpen error bars, not signs.
- **Implementation error.** The strongest threat to any negative result. Mitigations: 15 property-based unit tests on the exact tensors used in training; closed-form/autograd/finite- difference triple agreement on ∇V; energy monotonicity verified on trained checkpoints, not just at init; and a theory (§6.4b) that *quantitatively* predicts the diagnostics. An implementation bug reproducing all of these coincidences would itself be remarkable.

10 Conclusion

We set out to build a network that thinks by physics and built one that forgets by physics. Port-Hamiltonian dynamics over a fixed coercive potential with gated dissipation is mechanically impeccable — antisymmetric to machine precision, provably energy-dissipating, numerically stable under three integrators — and computationally sterile: on tasks a small LSTM solves at 91%, it never escapes the answer-marginal plateau its own Lyapunov structure mandates (Theorem 1), and its thinking phase is contraction toward an input-independent fixed point — information erasure with the aesthetics of inference. The obstruction is not timescale, optimization, or capacity (§6.1–6.3 eliminate each); it is the absence of input-conditional multistability, a property that successful iterative architectures (DEQs, modern Hopfield networks, equilibrium propagation) all possess and that any future "thinking dynamics" must build in from the start (§7). The full diagnostics trail behind this analysis gives the next attempt a concrete starting point at Lesson L2 rather than at zero.

Lessons (restated compactly).

L1 A stability guarantee is a capacity bound in disguise: count the input-selectable ω-limit sets before building. Here the count was one.

L2 Put the input in the landscape, not just the state: thinking dynamics must contract toward *input-dependent* targets.

L3 Audit the protocol before the architecture: a 30-minute baseline under the identical protocol exposed both the protocol artifact and the true gap, and was the highest-information experiment of the project.

L4 Unit tests verify mechanics, not viability: 15/15 passing tests coexisted with a provably vacuous hypothesis class. Pair property tests with an end-to-end learnability probe on day one.

Reproducibility Statement

All experiments run on CPU; each headline run completes in under one hour. Data are generated programmatically (no downloads); seed 42 throughout. Task generators, training schedules, and evaluation protocols are specified in Appendix C and §4.1–6.2; hyperparameters in Appendix A and diagnostics in Appendix B. Marginal entropies and majority frequencies (§4.1) are reproducible via a 10-line script over 50K generated samples.

References

Bai, S., Kolter, J. Z., & Koltun, V. (2019). Deep Equilibrium Models. *NeurIPS*.

Chang, B., Chen, M., Haber, E., & Chi, E. H. (2019). AntisymmetricRNN: A Dynamical System View on Recurrent Neural Networks. *ICLR*.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. (2018). Neural Ordinary Differential Equations. *NeurIPS*.

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., & Kaiser, Ł. (2019). Universal Transformers. *ICLR*.

Desai, S., Mattheakis, M., Sondak, D., Protopapas, P., & Roberts, S. (2021). Port-Hamiltonian Neural Networks for Learning Explicit Time-Dependent Dynamical Systems. *Physical Review E*, 104(3), 034312.

Giannou, A., Rajput, S., Sohn, J., Lee, K., Lee, J. D., & Papailiopoulos, D. (2023). Looped Transformers as Programmable Computers. *ICML*.

Graves, A. (2016). Adaptive Computation Time for Recurrent Neural Networks. *arXiv:1603.08983*.

Greydanus, S., Dzamba, M., & Yosinski, J. (2019). Hamiltonian Neural Networks. *NeurIPS*.

Hasani, R., Lechner, M., Amini, A., Rus, D., & Grosu, R. (2021). Liquid Time-Constant Networks. *AAAI*.

Hoover, B., Liang, Y., Pham, B., Panda, R., Strobelt, H., Chau, D. H., Zaki, M. J., & Krotov, D. (2023). Energy Transformer. *NeurIPS*.

Hopfield, J. J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *PNAS*, 79(8), 2554–2558.

Manek, G., & Kolter, J. Z. (2019). Learning Stable Deep Dynamics Models. *NeurIPS*.

Ramsauer, H., Schäfl, B., Lehner, J., et al. (2021). Hopfield Networks is All You Need. *ICLR*.

Scellier, B., & Bengio, Y. (2017). Equilibrium Propagation: Bridging the Gap between Energy-Based Models and Backpropagation. *Frontiers in Computational Neuroscience*, 11:24.

Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS*.

Zhong, Y. D., Dey, B., & Chakraborty, A. (2020). Dissipative SymODEN: Encoding Hamiltonian Dynamics with Dissipation and Control into Deep Learning. *ICLR Workshop on Integration of Deep Neural Models and Differential Equations*.

A Complete Hyperparameters

Component	Setting
PH network	$d = 32$, $L = 3$, 414,664 params (incl. embedding + head)
Potential	smooth L1; $\lambda = 0.01$, $\varepsilon = 10^{-3}$; $\lambda/\varepsilon = 10 \Rightarrow \nabla V \approx 11S$ near 0
$J_\theta(S)$	MLP $d \to 4d \to d^2$, tanh hidden; antisymmetrized; soft Frobenius bound $j_{\max} = 2$; Xavier gain 0.05, zero bias
Dissipation	$\beta_{\text{diss}} = 0.3$; gate $\sigma(wg + b)$, init $w = 1$, $b = -2$ ($\rho_0 = \sigma(-2) \approx 0.119$)
Integration	explicit Euler; $\Delta t = 10^{-3}$; adaptive $\Delta t_{\text{eff}} = \Delta t/(1 + 0.05\,\|\nabla V\|)$; RK4 & step-halving DP available
Input gate	γ: MLP $d \to 8 \to 1$, tanh + sigmoid; input map $\tanh(GX)$, G orthogonal init gain 0.5
Coupling	adjacent-layer pulses; $\beta_c = 0.03$, $\theta_c = 0.05$, projection Xavier gain 0.01
Embedding	$\mathcal{N}(0, 0.01^2)$; state init $\mathcal{N}(0, 0.01^2)$ per sequence
Optimizer	AdamW; lr $3{\times}10^{-3}$; wd 0.01; clip 1.0; batch 64; 200 epochs (grid: 50)
Data	5,000 / 500 per task; i.i.d. generation; seed 42
Curriculum (Phase 2)	30% epochs at $n_{\text{think}} = 0$, linear ramp to $n_{\text{think}}^{\max}$; energy penalty $\alpha \sum_\ell V(S_\ell)$, $\alpha = 0.01$ (threshold relu($\overline{V} - 10.0$))
Baselines	LSTM 2×32 (17,676) / 2×64 (68,108); MLP mean-pool 2 hidden (2,892)

B Diagnostic Details

B.1 Timescale. Trained checkpoint, `arithmetic`: 20 autonomous steps at $\Delta t{=}10^{-3}$ give $\|\Delta S\| = 0.013$; at $\Delta t{=}0.1$, $\|\Delta S\| = 0.986$. Accuracy identical (§5.3). Convergence of the autonomous flow: $\sim 8{\times}10^3$ steps (measured) vs. $6.9{\times}10^3$ (predicted from the linearized contraction rate of §6.4b).

B.2 Overfitting probe. 10 training samples, 500 epochs: 100% train accuracy in {original, $\Delta t{\times}50$, low-diss, $d{=}64$}.

B.3 Gradient flow. Mean $|\partial \mathcal{L}/\partial \theta| \approx 0.029$ at $n_{\text{think}} \in \{0, 5, 10, 20\}$; max-to-min ratio across $n_{\text{think}} < 1.1$.

B.4 Post-input geometry. Layer state norms after input phase ≈ 0.06; corresponding $\|\nabla V\| \approx 0.66$, $\rho \approx 0.21$, $\Delta t_{\text{eff}} \approx 9.7{\times}10^{-4}$ — operating point used in §6.4b. Energy strictly decreases at every one of 10^3 monitored thinking steps.

B.5 Marginal statistics (50K samples/task). `arithmetic`: $H(Y) = 2.215$, majority 22.2%; `threestep`: 2.201 / 23.8%; `vassign`: 2.240 / 20.7%. Compare PH: 24.9 / 25.6 / 26.1%; PH answer-position CE 2.183–2.187 (`arithmetic`).

B.6 Protocol artifact. Partial-input `arithmetic` (3 of 5 informative tokens): LSTM 22.5%, PH 22–25% — the information ceiling binds all architectures at the marginal and conceals the 66-point full-input gap.

C Task Construction

`arithmetic` (n=6 tokens): sample $a, b, c \sim \mathcal{U}\{0..9\}$, $\circ_1, \circ_2 \sim \mathcal{U}\{+, \times\}$; target $((a \circ_1 b) \circ_2 c)$ mod 10; tokens 10='+', 11='×'. **`threestep`** (n=8): as above with a, b, c, d and three operators. **`vassign`** (n=19): three fixed-format statements over variables $\{x, y, z\}$ (tokens 14–16), each `VAR = SRC op val ;`, first use of a variable draws its value uniformly; query token `ANS` (17) followed by variable; target is queried variable's final value mod 10. All values mod 10; all sampling i.i.d. uniform as specified in `experiments/datasets.py`.

D Unit Test Inventory (15/15 passing)

#	Property	Method
1	$J(S)^\top = -J(S)$	max-abs $\|J + J^\top\| < 10^{-6}$, random S, trained & untrained θ
2	$\|J\|_F \leq j_{\max}$	direct norm check under adversarially scaled inputs
3–4	∇V correctness	closed form vs. autograd vs. central differences (10^{-6})
5	$\nabla V(0) = 0$	evaluation at origin
6	$V > 0$, coercive	positivity on random states; growth along rays
7	Energy monotonicity	V non-increasing over 10^3 Euler steps, 100 seeds
8	State boundedness	no divergence over long autonomous rollouts
9–10	Low-rank layer parity	antisymmetry & bound for $J = UV^\top - VU^\top$ variant
11	Convergence	$\|\Delta S\| \to 0$ along autonomous trajectories
12	RK4 order	$\sim$50× lower global error vs. Euler, harmonic oscillator
13	Energy stability (RK4)	drift bound over long horizon
14	Conservation limit	pure-conservative system preserves V (dissipation off)
15	Integration pipeline	end-to-end forward/backward shape & finiteness checks