The Fake Mirror Effect: Foreign Feedback Disrupts Self-Correction in Minimal Recurrent Networks

Anonymous authors
Paper under double-blind review

Abstract

We construct a minimal recurrent network (35 neurons) and mechanistically dissect how it learns to correct predictions through output feedback. In a variable-noise task, the network achieves a correction gain of +0.119 (N=20, 95% CI excluding zero) by accumulating evidence across timesteps; ablating the recurrent loop eliminates this gain. Replacing a model's own feedback with another independently trained model's output ("clone feedback") degrades performance below the no-feedback baseline at the tested minimal scale—even when the clone's output has valid statistics. This "fake-mirror" effect is graded: self-feedback > same-model-wrong-trial feedback (positive) > clone and shuffled feedback (negative), with feedback interpolation revealing smooth monotonic degradation. Progressively stronger static aligners (30 to 17,925 parameters) recover most of the clone-feedback loss but plateau at ∼85% recovery, indicating that, within the tested alignment family, open-loop mapping is insufficient to fully restore closed-loop compatibility. Parameter-matched and compute-matched feedforward controls produce zero gain, and the core findings replicate on MNIST (120k parameters), though geometric mismatch harm attenuates at larger scales. These results provide evidence for *feedback-contract specificity*: the recurrent weights are co-adapted to the model's own output geometry, and the fidelity of this alignment—not merely the presence of feedback—determines whether iterative refinement succeeds.

1 Introduction

1.1 The Inseparability Question

Chain-of-thought prompting, self-consistency decoding, and test-time compute scaling all assume that a model benefits from processing its own prior output—but would the output of an equally competent stranger work just as well? If a model's recurrent weights are co-adapted to its own output geometry, substituting another model's output should degrade performance even when that output is statistically valid. If, on the other hand, any well-formed feedback signal suffices, the benefit is generic rather than self-specific. The question has practical relevance: methods relying on external feedback—such as RLHF or distillation—implicitly assume this signal is compatible with the model's internal processing, an assumption our results suggest deserves scrutiny at scale.

An analogous coupling exists in biological systems: damage to prefrontal monitoring circuits impairs not only self-monitoring but also task performance more broadly (Fleming & Dolan, 2012). Whether artificial neural networks exhibit a similar inseparability—where removing the self-referential pathway degrades the system—remains experimentally untested.

[0]During the preparation of this manuscript, large language models were used to provide language refinement and structured feedback on the experimental plan. All scientific decisions, experiments, and claims are the author's responsibility.

1.2 Why 35 Neurons

Our architecture sits at a critical intersection: it is *task-minimal* yet *fully transparent*. This scale breaks in both directions: smaller networks lack the capacity to reliably develop self-correction, while larger ones lose complete simulatability.

At 35 neurons (325 parameters), the system can be exhaustively dissected, placing this work squarely in the toy-model tradition of mechanistic interpretability (Olah et al., 2020; Elhage et al., 2022). The 50-parameter recurrent matrix can be arithmetically decomposed, hidden trajectories tracked in their native 10-dimensional space, and causal interventions—such as wrong-trajectory substitutions and per-neuron knockouts—performed exactly, without dimensionality-reduction artifacts.

We adopt the logic of Hodgkin & Huxley (1952)'s squid giant axon: the preparation's scale permits direct causal measurements that are infeasible in the opaque models where these phenomena are typically deployed. We do not claim biological relevance—the analogy is about methodology, not mechanism. What matters is whether the *phenomenon* observed here also appears at larger scales, not whether 35 neurons are themselves representative. We address this directly: scale verification shows that self-correction and the C1 fake-mirror effect persist up to hidden width 245 and on MNIST at 120k parameters, while the C2 effect weakens and can become non-harmful at larger scale (§ 3.7).

1.3 Prior Work

Our investigation lies at the intersection of three established literatures.

Emergent self-reference. Metacognitive capabilities can emerge spontaneously in recurrent architectures. Recent work shows tiny RNNs discover biological cognitive strategies (Li et al., 2025) and uncertainty monitoring (Ma et al., 2025) without explicit training, while evolutionary simulations yield self-referential metamemory (Yamato et al., 2022). Theoretically, self-referential pathways have been argued to be representationally necessary for true meta-learning (Kirsch & Schmidhuber, 2022). Our system is self-referential in a weaker sense—output feedback rather than parameter self-modification—but this suffices to test whether the recurrent pathway develops model-specific co-adaptation.

Iterative computation. Recurrent networks naturally support temporal integration and iterative refinement. Architectures like ACT (Graves, 2016) and PonderNet (Banino et al., 2021) train models to dynamically determine their computational depth. More broadly, recurrent systems accumulate evidence from noisy observations over time (Gold & Shadlen, 2007), a trajectory-refinement process often analyzed through fixed-point attractor dynamics (Sussillo & Barak, 2013). These works establish that recurrent loops can effectively refine predictions.

Solution degeneracy. Concurrently, a robust body of work demonstrates that independently trained neural networks develop wildly divergent internal geometries. Despite identical architectures and tasks, neural dynamics exhibit massive solution degeneracy across random initializations (Maheswaranathan et al., 2019; Huang et al., 2025). To quantify and bridge these structural incompatibilities, techniques like model stitching (Bansal et al., 2021) employ learned affine transformations to map representations between frozen layers of different networks.

Closed-loop co-adaptation. The general principle that a controller must be compatible with the system it regulates is well established in control theory, and analogous distribution-shift problems arise in reinforcement learning when policies are evaluated off-policy or when reward models are over-optimized beyond their calibration range (Gao et al., 2023). Our work can be understood as a mechanistic demonstration of this principle in the specific context of recurrent output feedback: the recurrent weights and the output geometry are jointly trained, and substituting a foreign output geometry into this loop creates a closed-loop mismatch analogous to controller-plant incompatibility. The term *feedback-contract specificity* refers to this co-adaptation as observed in our minimal recurrent setting.

1.4 The Gap

These two facts—that recurrent loops enable iterative refinement, and that independently trained models develop incompatible representations—raise an untested question: *does this incompatibility causally degrade self-correction when one model's output is substituted into another's feedback loop?*

While it may seem intuitive that foreign feedback should be less effective than self-feedback, the empirical structure is far from obvious: foreign feedback actively *harms* rather than merely failing to help; this harm decomposes into separable coordinate-level and geometric-level components; progressively stronger static alignment recovers most but not all of the damage, with recovery saturating despite increasing capacity; and the geometric component attenuates at larger scales while the coordinate component does not. These patterns—the active harm, the decomposability, the saturation, and the scale dependence—are not predicted by the simple intuition that "incompatible feedback should hurt." To our knowledge, no prior study has performed this causal substitution test in a minimal, fully transparent network where every neuron and connection can be directly inspected.

1.5 Contributions

1. **Self-correction requires recurrence.** A minimal recurrent network learns to improve its predictions across timesteps via output feedback. Ablating this recurrent loop eliminates all correction gain under both static and variable-noise conditions.

2. **Feedback utility is graded by geometric compatibility.** Replacing a model's self-feedback with a clone's output degrades performance below the no-feedback baseline at the tested minimal scale. This degradation follows a graded hierarchy based on the feedback's origin and attenuates as model capacity increases.

3. **The feedback contract is soft and continuous.** Interpolation between self-generated and foreign feedback reveals a smooth, monotonic degradation of correction gain, indicating the absence of a sharp collapse threshold.

4. **Clone-feedback harm stems from trajectory-level incompatibility.** Mechanistic decomposition shows that progressively stronger alignment models (30 to 17,925 parameters) recover most of the performance drop but saturate at $\sim$85% recovery, with the residual not closing despite a 3.6$\times$ capacity increase from MLP-medium to MLP-large. This saturation is consistent with geometric, trajectory-dependent incompatibility that open-loop static alignment cannot resolve.

5. **Self-correction is robust and architecturally specific.** Parameter-matched and compute-matched feedforward controls fail to produce correction gain, confirming that capacity and depth cannot substitute for recurrent self-reference. The phenomenon remains highly robust across extensive hyperparameter variations.

2 Methods

2.1 Architecture, Task, and Training

We constructed a recurrent MLP: Input(10) $\rightarrow$ H1(10) $\rightarrow$ H2(10) $\rightarrow$ Output(5), totaling 35 neurons (325 parameters). Hidden layers use ReLU; the output is linear. The previous output feeds back to H1 via feedback $= \tanh(y_{t-1}/\tau)$. Temperature scaling ($\tau = 2.0$) prevents tanh saturation and maintains gradient flow. The system was implemented in pure NumPy; numerical gradient checking verified correctness (details in Appendix A).

The task is 5-class pattern classification on 10-dimensional prototypes with inter-class ambiguity: class $k \in \{0..4\}$ activates dimensions $2k$ and $2k{+}1$ with amplitude 1.0, and adjacent class dimensions with amplitude 0.3. In the primary setting (variable noise, § 2.3), independently sampled Gaussian noise is added per timestep, enabling temporal evidence accumulation. A complementary static setting provides the *identical* noise-corrupted pattern at all $T = 3$ timesteps; because the input is then constant, the recurrent loop

cannot serve as a memory pathway—it can only carry information about past outputs, cleanly isolating self-correction from memory.

Models were trained via full-batch gradient descent (lr = 0.01, 500 epochs, 200 train / 200 test samples) using 3-step BPTT. The time-weighted cross-entropy loss is $L = w_1 \cdot L(t{=}1) + w_2 \cdot L(t{=}2) + 1.0 \cdot L(t{=}3)$. Setting $w_1 = 0.0$ removes the penalty on the initial output, creating optimization pressure to use the feedback loop for iterative refinement. Because all timesteps share the same weights, the unpenalized t=1 output still benefits from gradient flow through t=2 and t=3, achieving non-trivial accuracy even without direct loss. Correction gain ($\text{acc}_{t3} - \text{acc}_{t1}$) thus measures the improvement from an unpenalized intermediate guess to a fully penalized final output. This explicit time-weighting is sufficient but not necessary: under variable noise, self-correction emerges even with uniform weighting ($w_1 = w_2 = 1.0$; § 3.6).

2.2 Experimental Conditions

Ten evaluation conditions isolate the causal role of self-referential feedback. We organize them into four functional families (detailed architectures and hyperparameters for Groups D′, D″, and C2-mlp are provided in Appendix B):

1. Baseline and structural ablations. Standard controls testing structural dependence on the recurrent pathway:

- **Baseline:** The fully trained recurrent model, evaluated normally.
- **Group A (Recurrent ablation):** All recurrent weights zeroed post-training. Preserves feedforward capacity but severs the feedback signal.
- **Group B1 (Random ablation):** 50 randomly zeroed weights (30 repeats/model) to control for generic parameter loss.
- **Group B2 (Structured ablation):** The entire h2-to-output pathway zeroed as a catastrophic structural sanity check.

2. Feedforward controls. Stateless networks trained from scratch, verifying that capacity and depth cannot substitute for recurrence:

- **Group D:** Standard feedforward network.
- **Group D′ (Parameter-matched):** Adds a skip connection to exactly match the recurrent model's parameter count.
- **Group D″ (Compute-matched):** A deep 6-layer network matching the computational depth and multiply-add operations of the Baseline's 3-timestep unroll.

3. Feedback identity manipulation (the "fake mirror"). These novel conditions preserve the recurrent architecture but alter the feedback signal to test for model-specific co-adaptation:

- **Group C1 (Shuffled feedback):** The feedback vector is independently re-permuted at each timestep (per-model random seed). This maintains the exact marginal statistics (distribution, norm) while destroying coordinate identity.
- **Group C2 (Clone feedback):** Self-feedback is replaced with the output of an independently trained "clone" (identical architecture and procedure, independent data; target seeds 0–19, donor seeds 100–119). Because the clone's output is a mathematically valid, in-distribution classification vector, any degradation strictly isolates trajectory-level geometric incompatibility from generic marginal-distribution artifacts.

4. Alignment controls (quantifying C2 mismatch). To decompose the C2 performance drop, we fitted progressively stronger transformations from donor logits to target logits on held-out self-feedback calibration data. Aligned donor outputs were injected during evaluation:

- **Group C2-aligned (Affine):** A fitted linear transform (least-squares, 30 parameters) to disentangle linear representation misalignment.
- **Group C2-mlp (Non-linear family):** Fitted MLPs of increasing capacity ($5 \to 16 \to 5$ at 181 parameters up to $5 \to 128 \to 128 \to 5$ at 17,925 parameters) to test whether stronger non-linear mappings can close the residual gap.

Together, these quantify how much of the "fake-mirror" degradation is reducible representation mismatch versus persistent incompatibility that static open-loop alignment cannot resolve (details in Appendix B).

2.3 Variable-Noise Task

To enable temporal evidence accumulation, we introduce a variable-noise variant: $x_t = \text{prototype}_k + \varepsilon_t$, with independently sampled $\varepsilon_t \sim \mathcal{N}(0, \sigma^2 I)$ per timestep. Target models and 20 donor models (seeds 100–119) were retrained under variable noise using the baseline hyperparameters (w_1=0.0, w_2=0.2, τ=2.0, 500 epochs). During evaluation, paired target and donor networks received identical per-timestep noise sequences, ensuring output differences strictly reflect model identity rather than input variation. A noise-level sweep ($\sigma \in \{0.0, 0.1, \ldots, 1.0\}$) probed the dependence of correction gain on task difficulty.

2.4 Interpolation and Mechanistic Analyses

The following exploratory analyses used the static-input cohort (N=10, noise = 0.5):

Feedback interpolation. We swept $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$ in $\text{feedback}_t = \alpha \cdot y_{t-1}^{\text{self}} + (1-\alpha) \cdot y_{t-1}^{\text{other}}$, testing zero, shuffle, and clone contaminations.

Mechanistic tracking. (a) Internal activations were logged to classify trials by correction outcome. (b) Hidden Layer 1 (H1) trajectories were projected into a 2D PCA space fitted on Baseline t=1. (c) We computed the per-trial cosine divergence between self and clone recurrent (W_{rec}) contributions.

Neuron importance. To quantify single-neuron contributions without baseline-collapse confounds, we performed decoupled knockouts (seed = 0). For H1, *intelligence importance* was measured by zeroing feedforward inputs (observing Δacc_{t1}), and *correction importance* by zeroing recurrent inputs (observing Δgain). Hidden Layer 2 neurons underwent full knockouts.

2.5 Supplementary Experiments

Seven additional experiments probe specificity and generalization:

(a) *Wrong-trajectory control:* to disentangle foreign-model identity from state-conditional mismatch, each model's feedback was replaced with its own output from a different trial (matched on predicted class and t=1 correctness; when multiple candidates matched, one was selected uniformly at random using a fixed seed), continuously substituted at every $t \geq 2$.

(b) *Cross-pairing:* 400 donor–target pairings (20×20) under VN to assess donor variance.

(c) *Divergence null baselines:* isotropic random vectors (upper bound) and same-class resampled self-feedback (lower bound) to contextualize clone divergence.

(d) *Training dynamics:* correction gain tracked at 11 epoch checkpoints (N=10) with fully trained donor.

(e) *Timestep extension:* evaluation at $T \in \{3, 5, \ldots, 20\}$ without retraining.

(f) *Scale verification:* four hidden widths (10, 20, 45, 245) under static and VN, 20 models each.

(g) *MNIST extension:* replicated on MNIST (∼120k parameters) to test generalizability (architecture details in Appendix B).

2.6 Statistical Design and Robustness

Primary experiments used N=20 independently initialized models (seeds 0–19); exploratory analyses (§ 2.4) used N=10. The inferential unit is the model, not the repeat. The primary measure is **correction gain** ($\mathrm{acc}_{t3} - \mathrm{acc}_{t1}$).

Significance was assessed with exact Wilcoxon signed-rank tests with Holm–Bonferroni correction. Pre-specified families: static m=4 (Baseline vs. {A, B1, C1, C2}), VN m=3 (Baseline vs. {A, C1, C2}), secondary m=2 (C1-vs-A, C2-vs-A). 95% bootstrap CIs (10,000 resamples, resampling model-level paired differences as a unit to preserve pairing structure) are reported as descriptive summaries. Groups B2, D, D′, D″ were excluded from the Holm families *a priori* as structural controls.

Robustness was assessed via hyperparameter sweeps: 80 static hyperparameter configurations ($w_1 \times w_2 \times \tau$ grid, 10 models each) and 40 VN hyperparameter configurations (10 models each; 400 model-runs total), including truly uniform time weighting ($w_1 = w_2 = 1.0$). Grid details are provided in Appendix C.

3 Results

3.1 The Baseline: Self-Correction Requires Recurrence

Under variable noise, the recurrent Baseline improved from an initial t=1 accuracy of 0.672 to a final t=3 accuracy of 0.791. This yields a substantial correction gain of $+\mathbf{0.119} \pm \mathbf{0.082}$ (N=20, 95% CI [+0.083, +0.154]). Under static input (constant noise across timesteps), the Baseline achieves a smaller but significant correction gain of $+0.042 \pm 0.034$ (95% CI: [+0.027, +0.057], N=20). By providing identical inputs across timesteps, this static setting definitively isolates self-correction from memory, establishing a clean foundation for the mechanistic analyses in § 3.5 (Figure 1).

Removing recurrence kills the gain. Ablating the recurrent weights (Group A) entirely eliminated this improvement under both settings (VN: $+0.006 \pm 0.038$, Static: 0.000), leaving t=1 and t=3 accuracies effectively identical. Because stateless models processing independent noisy inputs have an expected gain of zero by symmetry, this contrast shows that the Baseline actively accumulates evidence across timesteps strictly via its feedback loop. A noise-level sweep (§ 3.4) confirms correction peaks at intermediate difficulty, declining at extremes where signals are either unrecoverable or feedforward-sufficient.

3.2 Feedforward and Ablation Controls

Ablation controls confirm structural specificity. Group A's initial t=1 accuracy (0.658) matches the Baseline perfectly, confirming that recurrent ablation selectively neutralizes the correction mechanism without impairing feedforward recognition. Conversely, random parameter loss (B1) and severing the hidden-to-output pathway (B2) collapse t=1 accuracy to 0.500 and a chance-level 0.205, respectively. This demonstrates that generic network damage destroys basic pattern recognition rather than isolating the self-correction loop.

Feedforward controls rule out capacity and depth. Three stateless architectures—D (standard), D′ (parameter-matched), and D″ (compute-matched 6-layer)—yielded zero correction gain under both static and variable-noise conditions. Neither additional parameters nor computational depth enable iterative evidence accumulation without recurrent self-reference.

By design, the parameter-matched D′ model achieves a higher single-pass accuracy (0.820 at t=1) than the recurrent Baseline's refined accuracy (0.700 at t=3). We deliberately operate in a capacity-constrained regime to force the network to utilize temporal recursion, isolating the iterative refinement mechanism from raw parameter scaling.

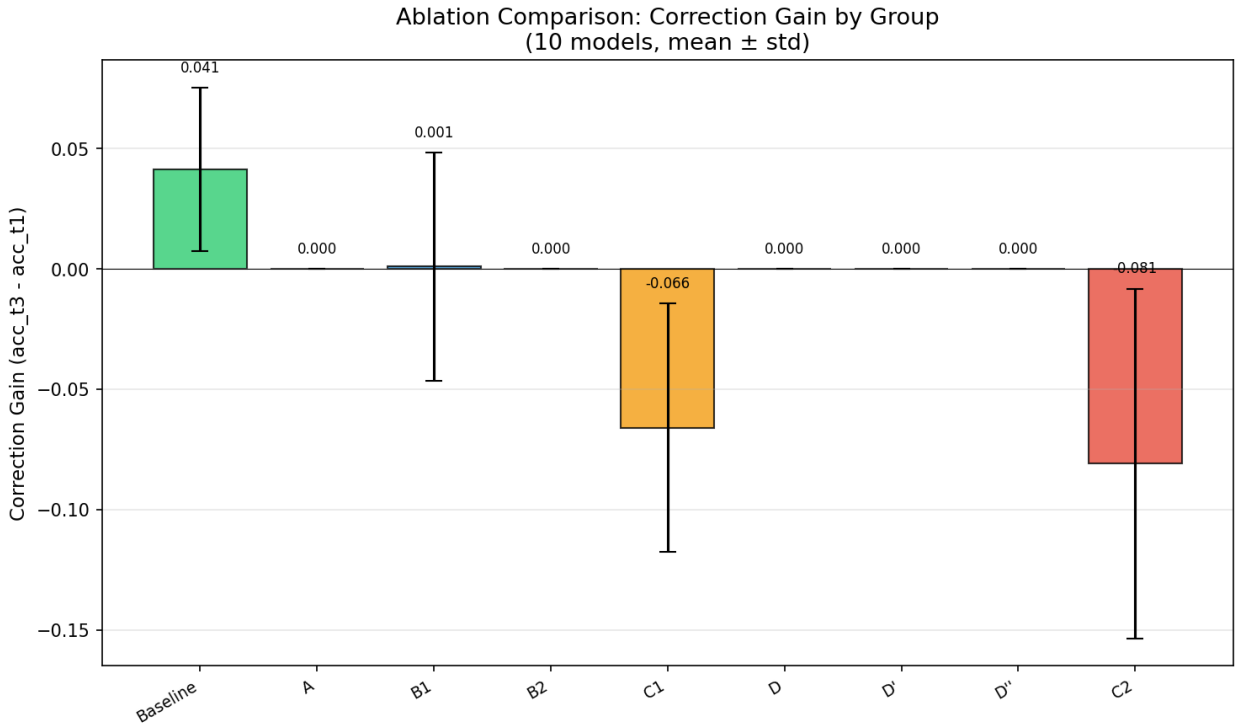

Figure 1: Correction gain across experimental conditions (static input, noise = 0.5, N=20). Baseline achieves positive gain; Group A (recurrent ablation) yields exactly zero; C1 and C2 produce negative gains worse than no feedback. Feedforward controls (D, D′, D″) show zero gain regardless of capacity.

3.3 The Fake Mirror Effect

Foreign feedback is worse than nothing. Providing non-veridical feedback does not merely fail to help; it actively degrades performance *below* the initial prediction across both settings. While removing recurrence yields a structurally guaranteed zero gain, foreign feedback actively drives the gain negative.

Replacing a model's self-generated feedback with non-veridical signals produced strictly negative correction gain—performance at t=3 consistently dropped below the initial t=1 prediction:

- **Shuffled feedback (Group C1):** gain $= -0.079 \pm 0.078$ (VN) and -0.066 ± 0.045 (Static CI: $[-0.086, -0.046]$). Permuting the feedback vector destroys coordinate identity while preserving its exact marginal statistics.

- **Clone feedback (Group C2):** gain $= -0.065 \pm 0.097$ (VN) and -0.081 ± 0.073 (Static CI: $[-0.112, -0.049]$). Substituting the mathematically valid, in-distribution output of an independently trained clone produced comparable degradation.

Critically, both substitutions performed significantly worse than severing the feedback loop entirely (Group A, no feedback, gain = 0.000 static / +0.006 VN). **Statistical summary:** All primary comparisons (Baseline vs. A, C1, C2) are highly significant (Holm–Bonferroni $p < 0.00002$ for VN). Secondary comparisons (C1-vs-A and C2-vs-A, separately corrected family m=2) confirm that foreign feedback is significantly worse than no feedback (Static C1 vs. A corrected $p = 0.000038$, Static C2 vs. A corrected $p = 0.000097$), with bootstrap 95% CIs for all paired differences excluding zero (Table 1). Processing structurally valid but foreign feedback misdirects the network more severely than receiving no feedback at all.

Table 1: Primary results (N=20). Correction gain ($\text{acc}_{t3} - \text{acc}_{t1}$) with bootstrap 95% CIs and exact Wilcoxon signed-rank p-values (uncorrected; all remain significant after Holm–Bonferroni correction with static m=4, VN m=3). Paired differences are Baseline minus each condition.

Condition	**Gain (mean ± SD)**	**95% CI**	**BL−X [CI]**	p
Panel A: Static Input				
Baseline	$+0.042 \pm 0.034$	$[+0.027, +0.057]$	—	—
Group A	$+0.000 \pm 0.000$	$[+0.000, +0.000]$	$+0.042$ $[+0.026, +0.056]$	1.6×10^{-4}
Group B1	$+0.001 \pm 0.017$	$[-0.006, +0.008]$	$+0.040$ $[+0.028, +0.052]$	3.6×10^{-5}
Group C1	-0.066 ± 0.045	$[-0.086, -0.046]$	$+0.107$ $[+0.088, +0.128]$	1.9×10^{-6}
Group C2	-0.081 ± 0.073	$[-0.112, -0.049]$	$+0.122$ $[+0.091, +0.156]$	1.9×10^{-6}
Panel B: Variable Noise				
Baseline	$+0.119 \pm 0.082$	$[+0.083, +0.154]$	—	—
Group A	$+0.006 \pm 0.038$	$[-0.011, +0.022]$	$+0.112$ $[+0.084, +0.142]$	1.9×10^{-6}
Group C1	-0.079 ± 0.078	$[-0.112, -0.044]$	$+0.198$ $[+0.155, +0.238]$	3.8×10^{-6}
Group C2	-0.065 ± 0.097	$[-0.106, -0.021]$	$+0.183$ $[+0.147, +0.221]$	1.9×10^{-6}

We refer to this pattern as *feedback-contract specificity*: the recurrent loop is tuned to the geometry of its own outputs, so even in-distribution foreign feedback can be disruptive. We examine the boundary between this observed specificity and standard closed-loop co-adaptation in § 4.2.

A cross-pairing analysis (400 donor–target pairings under variable noise) confirmed the C2 degradation generalizes robustly across initializations.

3.4 Continuous Dynamics: Noise and Feedback Interpolation

Feedback interpolation. To test whether the learned feedback contract is binary or graded, we swept the self-feedback fraction α from 1.0 (pure self) to 0.0 (pure contamination). Correction gain degrades smoothly and monotonically across substitution types (Figure 2), revealing a "soft" contract with no sharp collapse threshold. The zero-crossing occurs at $\alpha \approx 0.3$ for shuffled and $\alpha \approx 0.5$ for clone contamination, indicating that structured misdirection requires more veridical self-signal to overcome than incoherent noise. Performance plateaus near $\alpha \approx 0.7$, demonstrating that ∼70% self-feedback suffices for near-maximum correction (full tabular data in Appendix D).

Noise dependence. Correction gain exhibits an inverted-U relationship with task difficulty, peaking at intermediate noise ($\sigma = 0.5$, gain $= +0.119$) and declining at both extremes. At low noise, single-pass feedforward processing suffices; at high noise, the signal is unrecoverable. Self-correction thus provides the greatest advantage in an intermediate regime where initial predictions are frequently incorrect but the underlying signal remains recoverable through iterative refinement.

3.5 Mechanistic Dissection and Alignment Decomposition

Having shown that foreign feedback degrades performance smoothly rather than catastrophically (§ 3.4), we next ask what exactly is mismatched when that feedback enters the recurrent loop.

Hidden-state trajectories and wrong-trajectory control. Trials were classified by correction outcome: *corrected* (wrong at t=1, correct at t=3; 11.1%), *stable correct* (59.4%), *stable incorrect* (23.1%), and *over-corrected* (correct at t=1, wrong at t=3; 6.4%). PCA projection of internal activations (Figure 3) shows that self-feedback drives corrected trials toward class-appropriate regions. The over-corrected minority moves away, indicating that self-feedback is not infallible. Clone (C2) feedback induces comparable displacement magnitude but in misaligned directions. To test whether this misdirection stems from foreign-model identity or merely state-conditional mismatch, we evaluated a wrong-trajectory control (N=20): replacing each model's feedback with its own output from a different, class-matched trial. Same-model wrong-trial feedback preserved positive gain (static: +0.031, VN: +0.072), whereas clone feedback degraded it (static: −0.069,

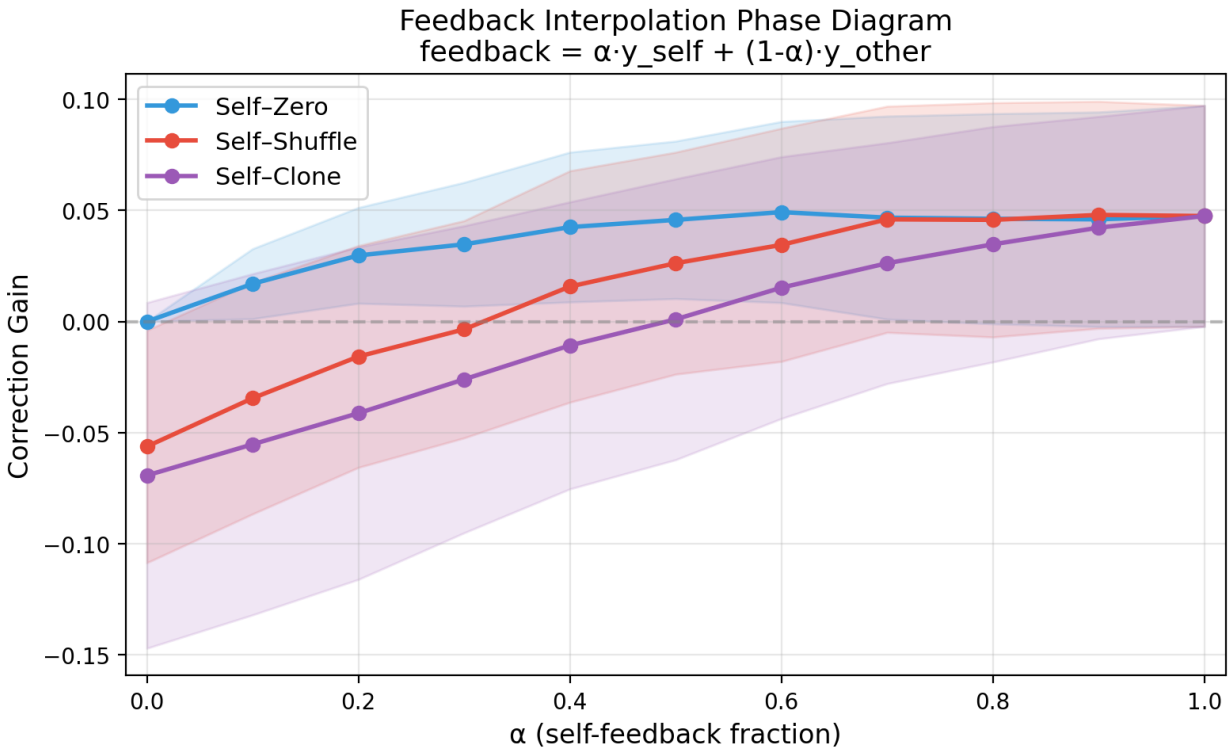

Figure 2: Feedback interpolation: correction gain as a function of self-feedback fraction α for three contamination types (zero, shuffle, clone). Degradation is smooth and monotonic with no sharp threshold. Zero-crossings differ by contamination type ($\alpha \approx 0.3$ for shuffle, ≈ 0.5 for clone).

VN: -0.057; note: clone gains here are evaluated on the matched-trial subset and may differ slightly from the full-sample C2 values in § 3.3). The network tolerates its own output from an incorrect trial but rejects a foreign model's output.

Clone feedback geometry. Although two models may produce similar discrete predictions, their continuous logit vectors can drive recurrent weights in conflicting directions. Across all test trials, the cosine divergence between self and clone recurrent (W_{rec}) contributions averages 0.701 (Figure 4)—falling between same-class resampled self-feedback (0.386) and isotropic random vectors (1.001), confirming that clone feedback occupies an intermediate position: more aligned than noise but substantially less aligned than self-feedback. At the seed level, however, divergence does not significantly predict C2 degradation (Spearman $\rho = -0.17$, $p = 0.46$, N=20), suggesting that the aggregate geometric mismatch drives the group-level effect without a strong per-model dose-response relationship.

Alignment decomposition (N=20). To quantify how much of this geometric incompatibility is simply static representation mismatch, we fitted learned transformations mapping donor to target logits. Recovering the Baseline-to-C2 gap yielded a distinct three-part decomposition:

- **Linear mismatch (66–76%):** An affine transform (30 params) recovered the majority of the lost gain (static: 76% recovery, VN: 66%).
- **Non-linear mismatch (additional 11–14%):** A learned MLP (181 params) recovered further gain (cumulative static: 87%, VN: 80%).
- **Scaling gains (additional ∼5%):** A larger MLP (4,869 params) further narrowed the gap (cumulative static: 87%, VN: 85%).

Table 2: Alignment saturation under variable noise (N=20). Recovery percentage indicates the fraction of the Baseline–C2 gap closed by each alignment level.

Alignment	**Params**	**VN Gain**	**Recovery**
C2-raw (no alignment)	0	−0.065	0%
Affine	30	+0.057	66%
MLP-small (5-16-5)	181	+0.081	80%
MLP-medium (5-64-64-5)	4,869	+0.091	85%
MLP-large (5-128-128-5)	17,925	+0.087	83%
Baseline (self-feedback)	—	+0.119	100%

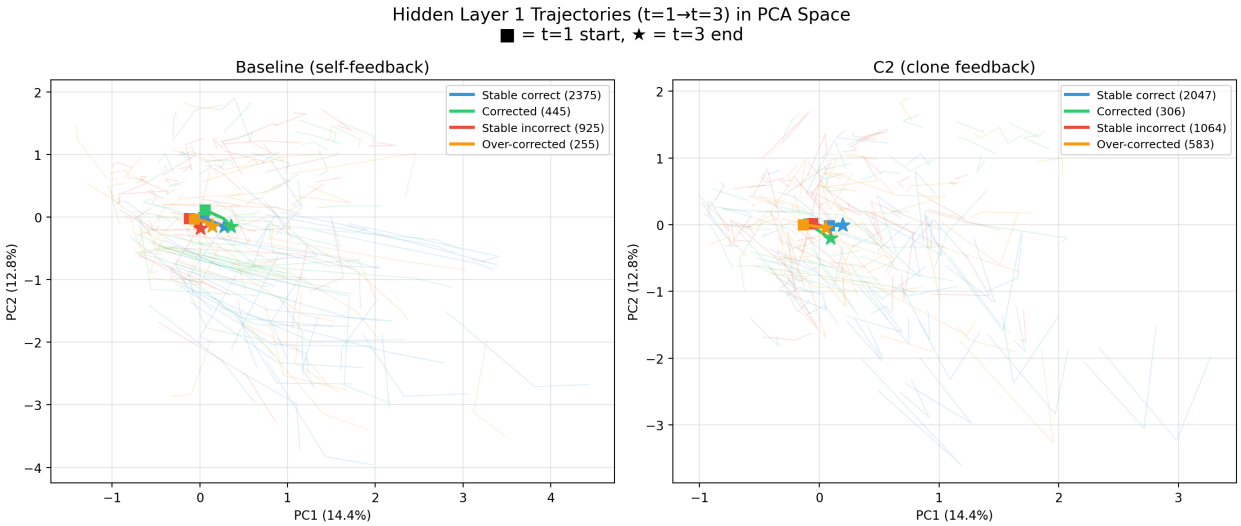

Figure 3: PCA projection of Hidden Layer 1 trajectories (seed = 0). Self-feedback (Baseline) drives corrected trials toward class-appropriate regions; clone feedback (C2) induces comparable displacement magnitude but in misaligned directions.

- **Persistent residual (∼15–17%):** A 17,925-parameter model provided no additional improvement—recovery saturated despite a 3.6× capacity increase over the previous level. Even at saturation, performance remained significantly below Baseline ($p = 0.008$ under VN; targeted residual test, not part of the pre-specified Holm-corrected families). All alignment models were trained with vanilla SGD using a 20% sample-level validation split with best-epoch model selection to control overfitting; the saturation pattern persists under this regularized regime.

This saturation pattern is a key finding (Table 2): progressively stronger alignment models show diminishing returns, and increasing the alignment model from 4,869 to 17,925 parameters yields no further improvement. This suggests that the remaining ∼15–17% residual reflects a limitation of static open-loop alignment for a closed-loop feedback task, rather than merely insufficient alignment capacity within the tested model family (full alignment family details in Appendix B).

Structural and single-neuron diversity (exploratory). Supporting this model-specific interpretation, W_{rec} sign consistency across 20 independently trained models averaged 0.559. This intermediate consistency suggests that multiple distinct weight configurations can implement effective self-correction, rather than a single universal algorithm. (Ablation controls confirmed this recurrent dependence is not a ReLU dead-zone artifact; activation shifts ≤ 12.3 pp.) Furthermore, decoupled single-neuron knockouts (seed = 0) revealed diverse internal wiring: Hidden Layer 1 partitioned into intelligence-specialized, correction-specialized, and shared functional groups. Not all recurrent channels are constructive; ablating h1_9's recurrent input

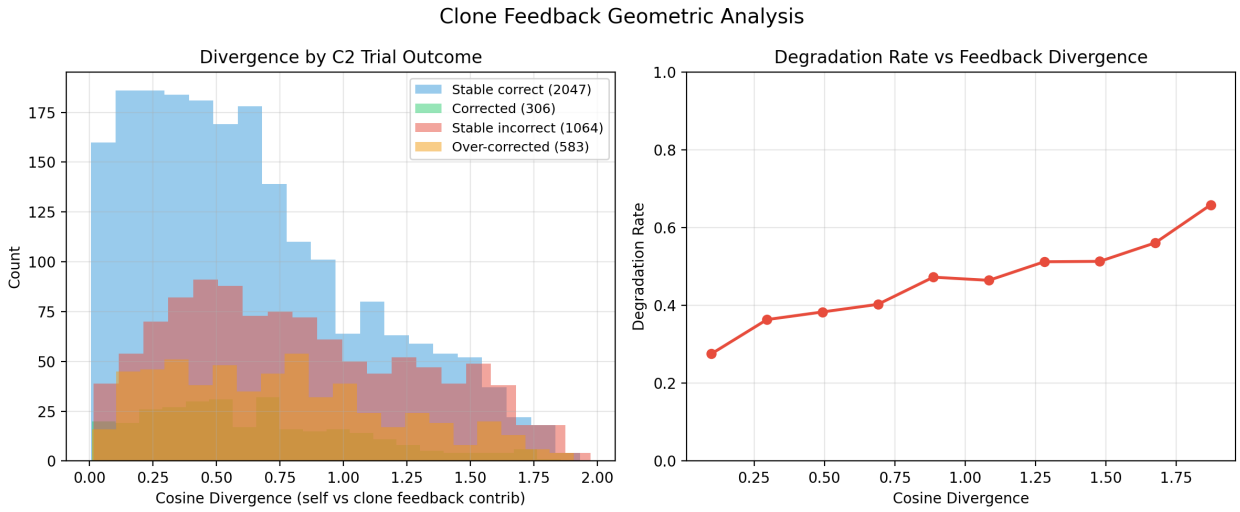

Figure 4: Cosine divergence between self and clone W_{rec} contributions. Clone divergence (0.701) falls between resampled self-feedback (0.386) and isotropic random vectors (1.001), confirming partial but insufficient geometric alignment.

actively *improved* correction gain (-0.035 importance), underscoring the idiosyncratic nature of the learned feedback contract. Full single-neuron heatmaps and Layer 2 analyses are deferred to Appendix E.

3.6 Robustness and Temporal Stability

Robustness and spontaneous emergence. Self-correction is not an artifact of a specific loss weighting. Under variable noise (VN), the phenomenon is highly robust (emerging in 100% of tested configurations; Figure 5) and arises spontaneously even under **truly uniform time weighting** ($w_1 = w_2 = 1.0$, all 40 models positive across four tau configurations of 10 seeds each). The explicitly time-weighted loss is therefore sufficient but not necessary: iterative refinement emerges naturally from the combination of a recurrent architecture and per-timestep noise variation. (Under static input, emergence was broadly robust at 68%; full grid details in Appendix C.)

Training dynamics. The feedback contract forms gradually over ∼200 epochs (N=10). Vulnerability to shuffled feedback (C1) develops in parallel with self-correction. Clone feedback (C2) harm emerges later (∼epoch 100). A caveat: because the donor is evaluated at its final trained state, this delayed C2 emergence may partly reflect progressive target-donor geometric divergence rather than purely intrinsic differentiation.

Timestep extension. Evaluating beyond the T=3 training horizon (up to T=20, no retraining) confirms stable convergence. Baseline performance plateaus safely (static: 0.712 → 0.692; VN: 0.779 → 0.770), whereas non-veridical feedback (C1, C2) compounds slightly rather than self-correcting at higher T.

3.7 Scale Verification and MNIST Extension

Scale verification. To confirm our findings are not toy-model artifacts, we evaluated hidden widths from 10 to 245 (∼325 to 65k parameters). Self-correction and the C1 fake mirror effect (harmful shuffled feedback) persisted universally (Figure 6). Crucially, C2 (clone feedback) exhibited a distinct **scale-dependent pattern**: actively harmful at small capacities but neutral or slightly positive at larger scales. This is consistent with independently trained representations converging at larger scales, rendering foreign feedback more geometrically compatible (full per-width data in Appendix F).

MNIST extension (∼120k parameters). Core phenomena successfully replicated on a 10-class MNIST task (N=20): Baseline gain = $+0.020 \pm 0.005$ (all 20 models positive, $p < 0.0001$ vs. Group A), C1 gain = -0.020 ± 0.008, confirming that coordinate disruption damages iterative refinement on real-world datasets.

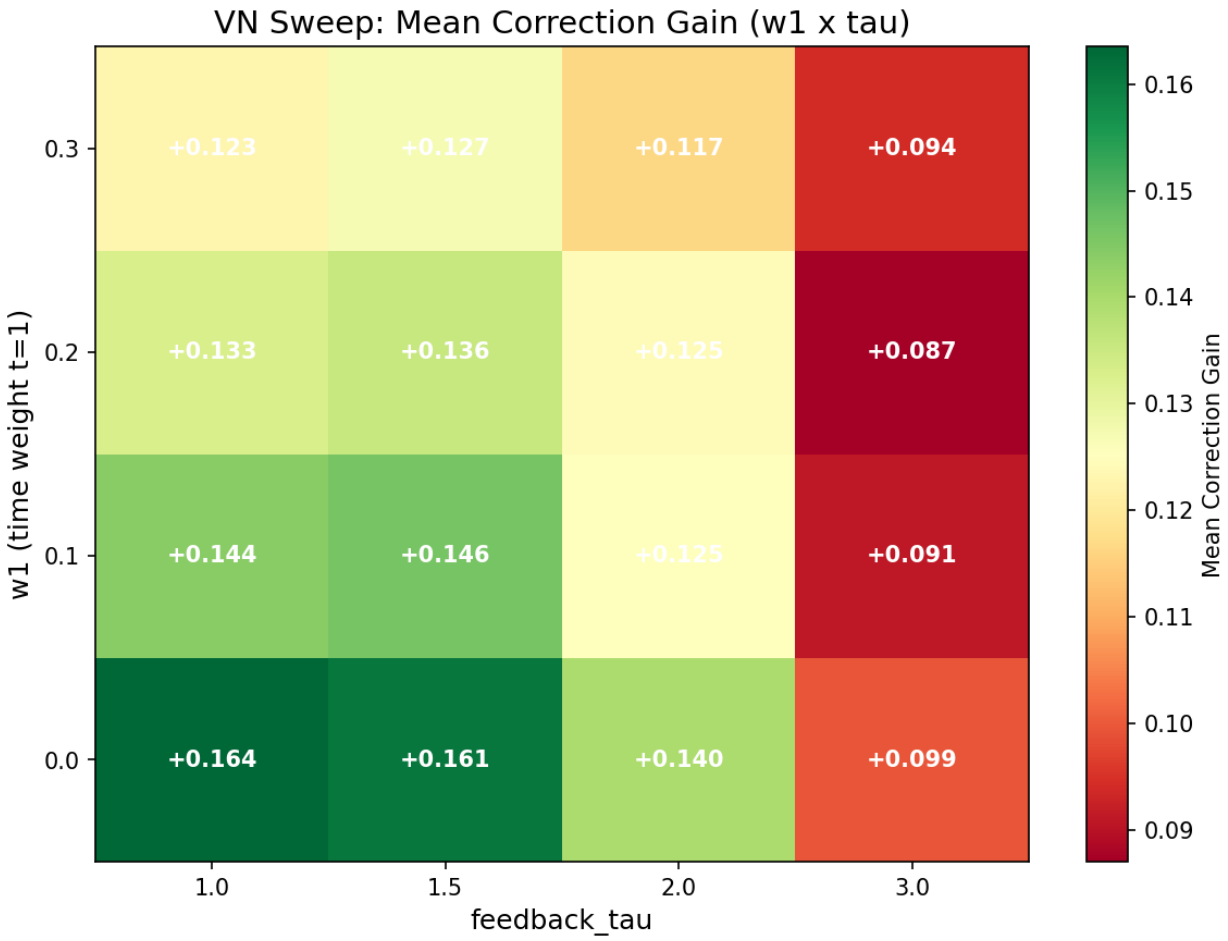

Figure 5: VN hyperparameter sweep (targeted grid): correction gain across w_1 and τ configurations (w_2=0.2 fixed, 10 seeds each). Self-correction emerges in 100% of these 16 configurations. An extended grid (24 additional configurations with $w_1 \in \{0.5, 0.7, 1.0\}$ and $w_2 \in \{0.2, 1.0\}$; Appendix C) also shows 100% emergence, yielding 40/40 total.

C2 clone feedback yielded a positive gain ($+0.011 \pm 0.005$)—unlike in the 35-neuron model—but remained significantly below self-feedback (Baseline − C2 = +0.009, $p < 0.0001$). This C1/C2 dissociation confirms the fake mirror effect is driven by coordinate-level disruption (C1), not merely foreign-model identity (C2) (full per-seed data in Appendix F).

Training dynamics. Tracking MNIST training revealed an **inverted C2 trajectory**: C2 clone feedback is negative at epoch 1 but recovers to a positive gain by epoch 3, while C1 remains negative throughout. This pattern is consistent with representation divergence—foreign feedback may be more disruptive early in training when models are geometrically idiosyncratic—though a floor effect (low baseline accuracy at early epochs limits observable harm) and scale mismatch between the fully-trained donor and immature target may also contribute. The C1/C2 dissociation across epochs, however, is not explained by these confounds alone.

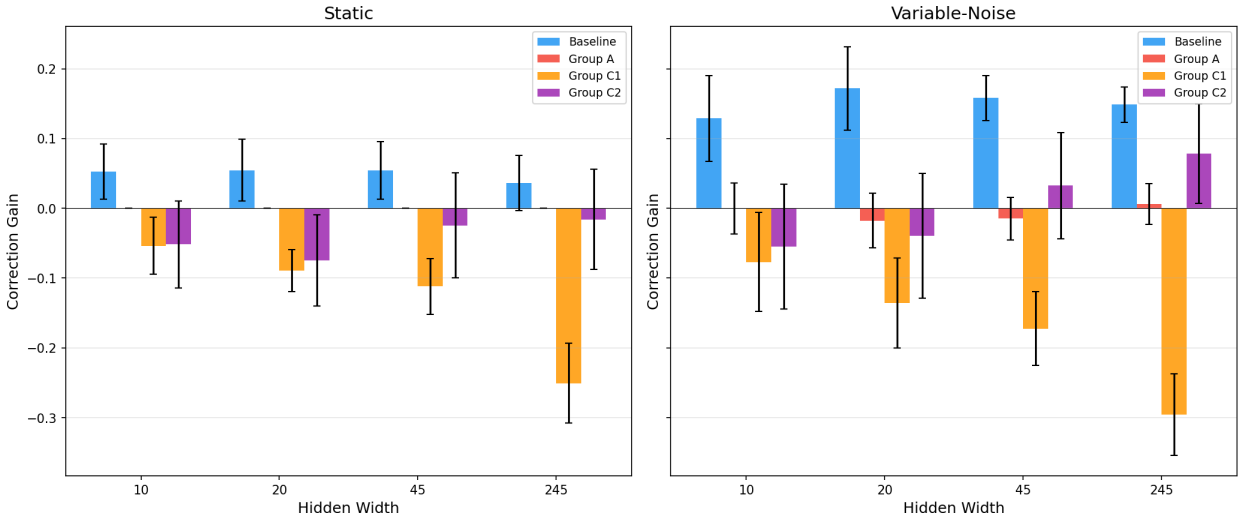

Figure 6: Scale verification across hidden widths 10–245. Self-correction (Baseline) and the C1 fake mirror effect persist at all scales. C2 (clone feedback) shows scale-dependent harm: negative at small scales, neutral or positive at larger scales under variable noise.

4 Discussion

4.1 Induction Depends on Conditions, Not Merely Scale

Self-correction developed in a network of 35 neurons—orders of magnitude smaller than systems where iterative refinement is typically studied. Under static input, temporal freedom ($w_1 < 0.3$) and gradient flow through the feedback pathway were important for reliable induction; static gain diminishes as w_1 increases and approaches zero at $w_1 = 0.3$, suggesting that static self-correction partly reflects deferred computation enabled by reduced early-timestep loss pressure. Under variable noise, however, self-correction emerged even under uniform time weighting ($w_1 = w_2 = 1.0$, all 40 models positive; § 3.6), demonstrating that the phenomenon arises spontaneously from the combination of recurrent architecture and per-timestep noise variation, without requiring any special loss design. Under VN, Group A gain ≈ 0 is the expected statistical baseline for a stateless model under i.i.d. inputs; the informative result is the Baseline's gain of +0.119, demonstrating evidence accumulation through recurrent feedback.

One might argue that VN self-correction is simply Bayesian evidence accumulation. The VN setting does provide independent observations that a temporal integrator can exploit, and the larger VN gain (+0.119 vs. +0.042 static) is consistent with an evidence-accumulation component. The key point is that this accumulation requires a mechanism for temporal integration: stateless models (Group A, D, D′, D″) have expected gain of zero under i.i.d. inputs and show gain ≈ 0 empirically. VN self-correction likely involves both evidence accumulation and error revision, both of which depend on recurrent self-reference in our setting.

4.2 The Fake Mirror Is Worse Than No Mirror

The fake mirror effect reveals two distinct layers of feedback-contract specificity. At the **coordinate level** (C1), permuting the feedback vector destroys dimensional identity while preserving marginal statistics—and this is universally harmful across all tested scales, from 35 neurons to MNIST at 120k parameters. At the **geometric level** (C2), the clone's output preserves coordinate structure but carries a misaligned continuous geometry—and this harm is scale-dependent: actively disruptive at small scales where independently trained models develop divergent representations, but tolerable at larger scales, consistent with representation convergence (Huh et al., 2024). The C1/C2 dissociation is itself informative: it demonstrates that the fake mirror effect is driven primarily by coordinate-level disruption, not merely by foreign-model identity. Even at scales where C2 is no longer harmful, self-feedback retains a significant advantage (§ 3.7), suggesting that recurrent

feedback loops are particularly sensitive to geometric compatibility beyond what representation convergence alone can resolve. A related problem arises in multi-agent self-play: independently trained agents develop incompatible conventions that break during cross-play (Hu et al., 2020). Our setting is the single-agent, closed-loop version of this—the recurrent weights overfit to their own model's output conventions.

The alignment decomposition (§ 3.5) provides the strongest evidence for why this sensitivity persists. Progressively stronger static alignment models (30 to 17,925 parameters) recover most of the C2 degradation but saturate at ∼85% recovery under variable noise. The key insight is not the residual size (∼15–17%) but the saturation pattern: increasing alignment capacity by 3.6× from MLP-medium to MLP-large yields no additional improvement. This suggests that the remaining gap reflects a limitation of static open-loop alignment for a closed-loop feedback task, rather than merely insufficient alignment capacity within the tested model family. The alignment is calibrated in open-loop (both models running independently), but evaluated in closed-loop (aligned output injected into the target's recurrent pathway). Consequently, this closed-loop distribution shift should be understood as part of the phenomenon rather than merely a nuisance confound: static open-loop similarity cannot guarantee dynamic closed-loop compatibility during iterative refinement.

4.3 Broader Connections and Implications

Our time-weighted loss is analogous in one respect to Chain-of-Thought prompting (Wei et al., 2022): both reduce pressure on intermediate computation steps, granting the model additional processing without penalizing intermediate exploration. Unlike PonderNet or ACT, our approach requires no explicit halting mechanism—only a loss reweighting.

Extrapolation from toy models to billion-parameter systems requires caution, but the MNIST extension (§ 3.7) provides a partial bridge: self-correction and the C1 fake mirror effect replicate at 120k parameters on a real-world dataset. The C2 effect is scale-dependent—harmful at small scales, suboptimal but not harmful at MNIST scale—suggesting that feedback-contract specificity is strongest when model representations are most individualized. The core finding that recurrent self-reference is necessary for correction (Group A ≈ 0 in all settings) is consistent with theoretical arguments that self-referential structure is representationally necessary for certain forms of meta-learning (Kirsch & Schmidhuber, 2022). The C1 result—that misaligned feedback is worse than no feedback—is a cautionary observation for any system that trains on external signals as if they were self-generated: feedback that deviates from the model's own output geometry may actively degrade rather than merely fail to improve. Precedents in adjacent domains support this direction: in RLHF, optimizing a proxy reward model beyond a certain point degrades ground-truth performance (Gao et al., 2023), and in knowledge distillation, stronger teachers can paradoxically produce worse students when the capacity gap is large (Mirzadeh et al., 2020). Whether the specific feedback-contract mechanism we identify applies to RLHF or preference-based training at scale is an open empirical question that this minimal study cannot answer, but the direction of the effect is consistent across all scales we tested.

4.4 Limitations

1. **Scale and mechanistic specificity.** Thirty-five neurons performing pattern classification cannot capture the complexity of language or reasoning. The mechanistic analyses (trajectory visualization, W_{rec} dissection, cosine divergence) are specific to this architecture; whether the same geometric signatures appear in larger recurrent systems is untested. The MNIST extension (§ 3.7) provides a partial bridge to 120k parameters, but a substantial gap remains to the scales at which iterative refinement is practically deployed.

2. **Effect size in context.** Under variable noise, correction gain corresponds to a ∼18% relative improvement in accuracy. Under static input, the gain is smaller but the primary claim—that feedback utility is graded by geometric compatibility—rests on the relative ordering across conditions, not on absolute gain magnitude. As noted in § 3.2, the parameter-matched feedforward model D′ achieves higher single-pass accuracy than the recurrent Baseline's final t=3 accuracy, confirming that recurrent refinement complements rather than replaces feedforward capacity.

3. **Alignment recovery structure.** All alignment models tested in this study are static, open-loop mappers: they are trained on independently collected logit pairs and applied without adaptation during the closed-loop evaluation. We did not test dynamic alignment methods (e.g., training the alignment module via BPTT through the recurrent steps) that could potentially close the ∼15–17% residual by adapting to the closed-loop distribution shift. The saturation pattern (no improvement from 4,869 to 17,925 parameters) argues against simple capacity limitations, but whether a fundamentally different alignment paradigm—one that operates within the loop rather than outside it—could fully spoof the feedback contract remains an open question.

5 Conclusion

In a 35-neuron recurrent network, we have identified and characterized feedback-contract specificity—the co-adaptation of recurrent weights to a model's own output geometry. Four converging lines of evidence support this conclusion:

1. **Self-correction requires recurrence, not merely capacity or depth.** The recurrent Baseline achieves a correction gain of +0.119 under variable noise, while ablating the recurrent loop, adding parameters (D′), or increasing computational depth (D″) all yield zero gain. This holds across 68% of 80 static and 100% of 40 variable-noise hyperparameter configurations, and self-correction emerges spontaneously even under uniform time weighting ($w_1 = w_2 = 1.0$).

2. **Feedback utility is graded by geometric compatibility, not binary.** Feedback interpolation reveals smooth monotonic degradation from self-feedback to foreign contamination, with the zero-crossing at $\alpha \approx 0.3$–0.5. The wrong-trajectory control confirms this is model-specific: same-model wrong-trial feedback preserves correction, while foreign-model feedback degrades it.

3. **Static alignment recovers most but not all of the damage.** Progressively stronger alignment models (30 to 17,925 parameters) recover up to ∼85% of the clone-feedback degradation, but recovery saturates—increasing capacity by 3.6× yields no further improvement. This saturation suggests a limitation of open-loop static alignment for closed-loop feedback, rather than insufficient alignment capacity.

4. **The core findings generalize beyond the toy model.** Scale verification across hidden widths 10–245 confirms that self-correction and the C1 fake mirror effect persist at all scales. C2 clone feedback shows scale-dependent harm, becoming tolerable at larger scales and on MNIST (120k parameters), consistent with representation convergence. The C1/C2 dissociation—coordinate corruption is universally harmful, geometric misalignment is scale-dependent—reveals two distinct layers of feedback-contract specificity.

The central finding is that feedback utility is graded by geometric compatibility: self-feedback produces the largest correction, clone feedback is intermediate (with the gap modulated by representation convergence across scale), and shuffled feedback actively degrades performance. We offer these results as a case study from a fully transparent minimal model. In this recurrent setting, a fake mirror is not a harmless substitute: iterative refinement depends on *feedback-contract specificity*—whether the feedback matches the geometry the recurrent loop has learned to read.

Reproducibility Statement

All code, data, and analysis scripts will be publicly released upon acceptance. The 35-neuron experiments are implemented in pure NumPy and can be reproduced on a standard multi-core desktop computer; the MNIST extension is implemented in a separate companion repository using PyTorch and benefits from GPU acceleration. A master pipeline script (`run_all.py`) executes all 35-neuron experimental phases in sequence. Raw experimental data for the 35-neuron experiments are included as CSV files; MNIST results are provided in the companion repository.

Ethics Statement

This work studies fundamental mechanisms of iterative refinement in a minimal artificial neural network. The system is a 35-neuron toy model performing synthetic pattern classification; it has no direct real-world application and poses no foreseeable societal risk. The research may contribute to longer-term understanding of self-referential processing in AI systems, which is relevant to AI safety and alignment research. No human subjects were involved. All experiments are fully reproducible from the publicly available code and data.

References

Andrea Banino, Jan Balaguer, and Charles Blundell. Pondernet: Learning to ponder. *arXiv preprint arXiv:2107.05407*, 2021.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Stephen M Fleming and Raymond J Dolan. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1338–1349, 2012.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Joshua I Gold and Michael N Shadlen. The neural basis of decision making. *Annual Review of Neuroscience*, 30:535–574, 2007.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, 1952.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Ann Huang, Satpreet Harcharan Singh, Flavio Martinelli, and Kanaka Rajan. Measuring and controlling solution degeneracy across task-trained recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Louis Kirsch and Jürgen Schmidhuber. Eliminating meta optimization through self-referential meta learning. *arXiv preprint arXiv:2212.14392*, 2022.

Ji-An Li, Marcus K Benna, and Marcelo G Mattar. Discovering cognitive strategies with tiny recurrent neural networks. *Nature*, 644(8078):993–1001, 2025.

Haiping Ma, Wenlian Lu, and Jianfeng Feng. Spontaneous emergence of metacognition in neuronal computation. *Physical Review Research*, 7(3):033188, 2025.

Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5191–5198, 2020.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. doi: 10.23915/distill.00024.001.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yusuke Yamato, Reiji Suzuki, and Takaya Arita. Evolution of metamemory based on self-reference to own memory in artificial neural network with neuromodulation. *Scientific Reports*, 12(1):6233, 2022.

A Implementation and Verification

A.1 Pure NumPy Implementation

The entire system was implemented in pure NumPy without deep learning frameworks, ensuring that every computation is directly inspectable and that no hidden optimizer behavior, automatic differentiation artifact, or framework-level abstraction influences the results.

A.2 Gradient Verification

Backpropagation through time was implemented as a 3-step unroll, with numerical gradient checking (relative error threshold 10^{-4}) to verify correctness. Numerical gradient checking on a representative model (seed=0) yielded maximum relative errors of 6.57×10^{-8} (static) and 1.48×10^{-7} (variable noise) across all weight matrices, well below the 10^{-4} acceptance threshold.

A.3 Overfitting Assessment

The mean train-test accuracy gap at t=3 was +0.040 (static) and +0.050 (variable noise), indicating modest overfitting consistent with the small sample size (200 training samples).

A.4 Evaluation Noise and Reproducibility

Under variable noise, each evaluation generates independently sampled noise sequences per timestep. Consequently, per-seed correction gains may vary across runs even for identical trained models, while aggregate statistics (means, CIs) converge. All reported VN results use a single consistent evaluation run; re-running evaluation with a different random state will produce different per-seed values but statistically equivalent group-level summaries.

A.5 Mechanistic Transparency

Three classes of analysis were employed: (1) direct W_{rec} visualization and arithmetic decomposition, (2) per-trial native-space trajectory tracking without dimensionality reduction, (3) exact causal interventions (per-neuron knockout, wrong-trajectory substitution, cosine divergence null baselines across 4,000 trials).

A.6 PyTorch Cross-Validation

An independent PyTorch reimplementation of the 35-neuron architecture was used to verify that the core phenomena are not artifacts of the NumPy implementation. Under He normal initialization (matching the

NumPy codebase), all qualitative patterns replicate (N=20): Baseline gain is positive (static: +0.042, VN: +0.095), Group A yields zero gain, and both C1 and C2 produce negative gains. Effect sizes differ from the NumPy results because PyTorch's random number generator produces different weight initializations for the same seed; the qualitative ordering is preserved. Under PyTorch's default Kaiming uniform initialization, which produces ∼6× smaller initial weight variance, self-correction fails to emerge entirely—all conditions yield gain ≈ 0. This initialization dependence is consistent with the finding in § 3.6 that emergence depends on training conditions: smaller initial weights reduce the feedback signal below the threshold needed for the recurrent pathway to develop during training (Table 3).

Table 3: PyTorch cross-validation (N=20). Correction gain under two initialization schemes.

Init	**Baseline**	**Group A**	**C1**	**C2**
Static				
He normal	+0.042	0.000	−0.053	−0.049
Kaiming uniform	+0.002	0.000	−0.002	+0.001
Variable Noise				
He normal	+0.095	+0.000	−0.072	−0.017
Kaiming uniform	−0.000	−0.001	−0.003	−0.003

B Control Architectures and Alignment Details

B.1 Group D′: Parameter-Matched Feedforward

A feedforward network with an additional skip connection from input to output (50 weights), matching the total parameter count of the recurrent model (325 parameters).

B.2 Group D″: Compute-Matched Feedforward

A deep feedforward network with 6 hidden layers of 10 neurons each (Input(10) → H1(10) → . . . → H6(10) → Output(5)), matching the computational depth of the recurrent model's 3-timestep unroll. 715 parameters, approximately 650 multiply-add operations (vs. Baseline ∼900 MACs across 3 timesteps).

B.3 Affine Alignment

Both models were run with self-feedback on calibration data (200 samples × 3 timesteps = 600 logit pairs per pair); a least-squares regression $\hat{y}_{\text{target}} \approx W \cdot y_{\text{donor}} + b$ (30 parameters: 5×5 weight matrix + 5 bias) was fitted and applied during evaluation.

B.4 MLP Alignment

A small MLP (5 → 16 → 5 with ReLU; 181 parameters) from donor logits to target logits, trained with MSE loss and gradient-clipped SGD (500 epochs, lr=0.01, max gradient norm 1.0).

B.5 Stronger Alignment Family

Table 4 lists the alignment model configurations.

Calibration set doubled to 400 samples (1,200 logit pairs). R^2 alignment quality: 0.449 ± 0.053 (static), 0.465 ± 0.104 (VN).

Table 4: Stronger alignment model configurations.

Alignment Model	Parameters	Epochs	Learning Rate
Affine	30	500	0.01
MLP-small (5→16→5)	181	500	0.01
MLP-medium (5→64→64→5)	4,869	2,000	0.005
MLP-large (5→128→128→5)	17,925	2,000	0.003

B.6 MNIST Configuration

Input(784) → H1(128) → H2(128) → Output(10) with output→hidden feedback (119,562 parameters), implemented in PyTorch. 20 target models (seeds 0–19) and 20 donors (seeds 100–119) trained for 30 epochs with SGD (lr=0.01) under variable noise ($\sigma = 0.5$, $T = 3$) with time-weighted loss ($w_1 = 0.0$, $w_2 = 0.2$).

C Statistical Details and Robustness

C.1 Within-Model Pairing

Groups A, B1, C1, and C2 are post-hoc interventions on the same trained model (within-model pairing); Groups D, D′, and D″ are separately trained seed-matched controls.

C.2 B2 Exclusion

B2 was excluded from the primary correction family a priori because its catastrophic feedforward degradation renders correction gain comparison uninformative.

C.3 Emergence Heuristic

Mean gain > 0 with at least 60% of models showing positive gain. This criterion is deliberately liberal and intended only for exploratory grid characterization.

C.4 Static Hyperparameter Grid

$w_1 \in \{0.0, 0.1, 0.2, 0.3\}$, $w_2 \in \{0.1, 0.2, 0.3, 0.5\}$, $\tau \in \{1.0, 1.5, 2.0, 3.0, 5.0\}$, 10 models each (800 total). Reported configuration ranked 8th/80 by mean gain.

Static gain depends substantially on w_1: mean gain = +0.035 at w_1=0.0 (80% positive), +0.021 at w_1=0.1, +0.007 at w_1=0.2, and −0.002 at w_1=0.3, indicating that static self-correction partly reflects deferred computation from reduced early-timestep loss. In contrast, VN gain remains robust across all w_1 values (+0.104 to +0.141, 100% positive at every setting including w_1=w_2=1.0).

C.5 Variable-Noise Grids

Targeted: $w_1 \in \{0.0, 0.1, 0.2, 0.3\}$, $\tau \in \{1.0, 1.5, 2.0, 3.0\}$, $w_2 = 0.2$ fixed (160 runs). Extended: $w_1 \in \{0.5, 0.7, 1.0\}$, $w_2 \in \{0.2, 1.0\}$ (240 runs).

C.6 Static Noise Sweep

Under static input, correction gain peaked at noise=0.2 (gain∼+0.099) and declined at both extremes.

C.7 Pseudoreplication Control

Each model used per-model data seeds (seed = model seed for training, seed + 500 for testing).

D Interpolation Data

Table 5 reports the full interpolation data.

Table 5: Feedback interpolation results (N=20 models). α=1.0 is pure self-feedback; α=0.0 is pure contaminated feedback.

α	**Zero**	**Shuffle**	**Clone**
0.0	+0.000	−0.056	−0.069
0.1	+0.017	−0.035	−0.055
0.2	+0.030	−0.016	−0.041
0.3	+0.035	−0.004	−0.026
0.4	+0.043	+0.016	−0.011
0.5	+0.046	+0.026	+0.001
0.6	+0.049	+0.035	+0.015
0.7	+0.047	+0.046	+0.026
0.8	+0.046	+0.046	+0.035
0.9	+0.046	+0.048	+0.042
1.0	+0.048	+0.048	+0.048

E Neuron Knockout Details

E.1 Decoupled Knockout Methodology

Full neuron knockout simultaneously eliminates both the neuron's feedforward contribution to classification and its recurrent contribution to self-correction, confounding intelligence and correction effects. To decouple these, we employed a baseline-collapse methodology: for each knocked-out neuron, we compare correction gain against a per-neuron baseline where that neuron is ablated but the recurrent pathway remains intact, isolating the recurrent-specific contribution.

E.2 Hidden Layer 2 Caveat

For Hidden Layer 2 neurons (which have no direct recurrent input), full knockout was used; the intelligence-correction confound should be noted.

E.3 Trial Classification Breakdown

Across all trials, the following classification breakdown was observed: corrected 11.1%, stable correct 59.4%, stable incorrect 23.1%, over-corrected 6.4%.

F Scale Verification and MNIST Data

F.1 Scale Verification

Table 6 reports correction gain across hidden widths.

F.2 MNIST Per-Seed Data

Full per-seed gains are available in the supplementary CSV files (`mnist_metrics.csv`).

F.3 MNIST Training Dynamics

Table 7 tracks MNIST correction gain across training epochs.

Table 6: Correction gain across hidden widths (mean, N=20 models). Baseline is self-feedback; A is recurrent ablation; C1 is shuffled feedback; C2 is clone feedback.

Condition	**Hidden Width**	**Baseline**	**Group A**	**C1**	**C2**
Static	10	+0.053	0.000	−0.054	−0.052
	20	+0.055	0.000	−0.089	−0.075
	45	+0.054	0.000	−0.112	−0.025
	245	+0.036	0.000	−0.251	−0.016
VN	10	+0.129	−0.000	−0.077	−0.055
	20	+0.172	−0.018	−0.136	−0.040
	45	+0.158	−0.015	−0.172	+0.033
	245	+0.149	+0.006	−0.296	+0.079

Table 7: MNIST correction gain across training epochs (N=20 models). Baseline is self-feedback; A is recurrent ablation; C1 is shuffled feedback; C2 is clone feedback.

Epoch	**Baseline**	**Group A**	**C1**	**C2**
0	−0.001	+0.000	−0.000	−0.000
1	+0.014	+0.001	−0.018	−0.008
2	+0.017	−0.001	−0.023	+0.002
3	+0.019	−0.001	−0.023	+0.006
5	+0.021	+0.001	−0.023	+0.011
7	+0.021	−0.000	−0.023	+0.012
10	+0.021	+0.001	−0.021	+0.013
15	+0.020	+0.000	−0.022	+0.011
20	+0.019	+0.001	−0.022	+0.010
25	+0.019	−0.001	−0.021	+0.011
30	+0.018	−0.000	−0.022	+0.010