# On Faultization:
# Generative Pre-Trained Transformer.
# What Perturbation Reveals About Pattern Access in a Minimal Transformer

PIATRA . INSTITUTE

March 2026

## Abstract

We apply faultization (a systematic regime of morphogenetic perturbation) to a minimal transformer (4-layer, 16-dimensional, 4-head character-level GPT), asking which patterns from the latent space (Levin, 2026) the system accesses and which free lunches it receives. Twelve experiments span perturbation-during-training (Exp 1-6) and multi-phase morphogenetic interventions (Exp 7-12). We adopt a three-scale protocol: $n = 3$ pilot data provides initial signal, $n = 30$ paired analysis ($n = 30$ runs per condition, matched seeds) resolves ambiguity, and $n = 300$ reveals fine structure. At $n = 3$, several signals were ambiguous: head freezing appeared to improve loss, gradient corruption appeared neutral, the Delayed Gratification (DG) Index appeared to scale with perturbation. At $n = 30$, the picture sharpens: we identify four pattern manifestations not directly prescribed by Stochastic Gradient Descent (SGD) (stress inoculation as temporal pattern access, complete recovery as pattern re-binding, complete regeneration as functional role patterns, and head-freezing trajectory improvement as interface simplification), three that reflect pattern invariance (chimera convergence, transplant indifference, cell-view equivalent convergence), and three that demonstrate pattern fidelity (gradient degradation absorbed up to a threshold, partial communication, vision restriction). Cell-view (local loss) achieves equivalent final loss to baseline at both $n = 30$ ($p = 0.24$) and $n = 300$ ($p = 0.90$), demonstrating that local layerwise optimization converges to the same attractor as end-to-end backpropagation, the same pattern accessed through a different interface. At $n = 30$, head freezing shows final-loss improvements at some levels (freeze 4: $p = 0.001$, freeze 12: $p = 0.016$), but at $n = 300$ all final-loss effects resolve to null (all $p > 0.15$), while trajectory improvement strengthens to high significance. At $n = 300$, fine structure emerges: the temporal pattern access effect strengthens from $p = 0.032$ (final loss) at $n = 30$ to $p = 0.0001$ at $n = 300$ ($d = -0.374$), vision radius reveals a previously invisible monotonic structure (window 1 harms at $p = 0.021$, window 8 improves at $p = 0.022$), regeneration shows layer-specific residual effects, and all robust $n = 30$ findings hold or strengthen while all null findings remain null. The paper's key finding: gradual noise exposure builds tolerance that sudden exposure does not ($p = 0.0001$, $d = -0.374$), despite identical gradient update rules at every step, a free lunch the system receives from the latent space without the optimizer paying for it.

## 1. Introduction

Non-physical patterns are real, discovered, and causal. Levin (2026) argues that physical systems (embryos, computers, algorithms) are interfaces through which patterns from a latent space manifest. The patterns are not created by the interface; they are accessed through it. Evolution exploits them as affordances it does not need to pay for: once you find a voltage-gated ion channel, you have a transistor, and truth tables are yours for free. The central question for any system is therefore not just what it computes, but what patterns it accesses from the latent space, and what free lunches it receives: capabilities the system exhibits that were not specified by its construction, training, or evolutionary history.

This paper applies that question to a minimal transformer. The methodology is **faultization**: sys-

tematic perturbation of a system to reveal what patterns it accesses and how robustly it accesses them. Faultization degrades the interface (the weights, gradients, and communication channels through which the optimizer operates) and observes what happens to the patterns the system was manifesting. Four outcomes are possible:

1. **Pattern manifestation**: the pattern successfully manifests through the interface. The system accesses the optimization attractor despite perturbation.
2. **Pattern fidelity**: how much interface degradation the pattern tolerates before access is lost. There exists a threshold below which the pattern manifests cleanly and above which it does not.
3. **Pattern corruption**: the interface actively misleads, inverting the pattern. Rather than failing to access the pattern, the system accesses a corrupted version.
4. **Free lunch**: what the system receives without paying for. The optimizer specifies loss minimization; the system exhibits capabilities (stress tolerance, regeneration, recovery) that were not specified by the objective.

Levin (2026) proposes that the no-free-lunch commitments are derived from the laws of the physical world, but the latent space offers useful patterns for which the physical processes of learning, evolution, and engineering do not need to pay. This paper quantifies those free lunches for a minimal transformer under twelve perturbation protocols.

Transformers are typically studied through their outputs: probing learned representations (Belinkov & Glass, 2019), ablating components to measure importance (Michel et al., 2019), or tracing computational circuits (Elhage et al., 2021). These methods characterize what the system has learned or which parts matter. They do not ask what happens when the system is forced to learn under constraint, when its interface is *degraded*. Levin et al. (2024) introduced a different methodology in the context of simple algorithms: rather than analyzing sorting algorithms through their final outputs, they perturbed the algorithms during execution, revealing competencies (fault tolerance, delayed gratification, emergent aggregation) that were invisible during normal operation. The central insight: **perturbation reveals what normal operation conceals**.

The question of whether any finding is genuine requires statistical power. We adopt an explicit three-scale protocol. At $n = 3$ (pilot), signal is visible but ambiguous; the resolution is too low to distinguish real effects from noise. At $n = 30$, the picture sharpens and ambiguous signals resolve. At $n = 300$, fine structure emerges that $n = 30$ cannot see. No scale is wrong. Each reveals different phenomena. Signals that were ambiguous at $n = 3$ and resolved at $n = 30$ are findings at the appropriate resolution.

Our contributions are:

1. **A twelve-experiment faultization methodology** that systematically degrades the transformer interface across two phases: perturbation during training (Exp 1-6) and multi-phase morphogenetic interventions (Exp 7-12), to reveal what patterns the system accesses from the latent space.
2. **A four-category classification** (pattern manifestation, pattern fidelity, pattern corruption,

free lunch) that organizes what faultization reveals about the system's relationship to the latent space.

3. **The finding that perturbation history changes which patterns are accessible** ($p = 0.032$ at $n = 30$, strengthening to $p = 0.0001$ at $n = 300$, $d = -0.374$): gradual noise exposure builds tolerance that sudden exposure does not, despite identical gradient update rules at every step. A free lunch the optimizer did not pay for.
4. **A three-scale protocol** that treats $n = 3$ pilot findings as coarse signal, $n = 30$ as resolved signal, and $n = 300$ as fine-structure signal.
5. **A sharp distinction between pattern unavailability and pattern corruption.** Frozen (inactive) layers are tolerated ($p = 0.462$); gradient-negated (adversarial) layers degrade by +24.8% ($p < 0.001$). Silence (the pattern is simply not accessed) is absorbed; sabotage (the interface inverts the pattern) is catastrophic.
6. **A negative result on rerouting.** The Delayed Gratification Index shows no perturbation response at $n = 30$ or $n = 300$, distinguishing transformer perturbation response from the richer compensatory rerouting observed in biological development.

## 2. Related Work

### 2.1 Levin's Platonic Space Framework

Levin (2026) argues that non-physical patterns, mathematical truths, topological structures, computational regularities, are real, discovered rather than created, and causally relevant to physics, biology, and computer science. Physical systems are interfaces through which patterns from a latent (Platonic) space manifest. The framework proposes that evolution, learning, and engineering exploit these patterns as affordances: once a physical system instantiates the right structure, the patterns are available for free. The no-free-lunch theorems are constraints on the physical interface; the latent space itself offers enablements that the interface did not pay for.

This framework reframes the question of what a computational system does. Rather than asking "what did the optimizer prescribe?" we ask "what patterns does the system access, and which of those were not paid for by the optimization process?" Faultization, systematic perturbation of the interface, is a natural methodology for probing this question. By degrading the interface, we observe which patterns continue to manifest (pattern fidelity), which are lost (the interface was critical), which are inverted (pattern corruption), and which capabilities appear that were never specified (free lunches). The present paper applies this methodology to a minimal transformer, quantifying the free lunches and mapping the fidelity thresholds.

### 2.2 Prior Work

**Pruning and the lottery ticket hypothesis.** Frankle & Carlin (2019) showed that trained networks contain sparse subnetworks ("winning tickets") that match full-network performance. Subsequent work extended this to structured pruning of attention heads (Michel et al., 2019; Voita et al., 2019). In the Platonic Space framework, pruning simplifies the interface, fewer degrees of freedom may

provide cleaner access to the same pattern. Our Experiment 1 freezes heads at *random initialization*, not after training, the frozen heads are arbitrary random projections. At $n = 3$, the signal was ambiguous: freezing 8+ heads appeared to improve final loss. At $n = 30$, some final-loss improvements appear significant (freeze 4: $p = 0.001$, freeze 8: $p = 0.023$, freeze 12: $p = 0.016$) alongside a robust mean-trajectory improvement for 4+ frozen heads. At $n = 300$, the final-loss improvements resolve to null (all $p > 0.15$; Spearman $\rho = -0.0045$, $p = 0.84$), but the trajectory improvement strengthens to high significance: freeze 8 ($\Delta = -0.1\%$, $p < 0.0001$, $d = -1.245$), freeze 12 ($\Delta = -0.2\%$, $p < 0.0001$, $d = -1.421$), freeze 16 ($\Delta = -0.2\%$, $p < 0.0001$, $d = -1.312$), confirming that a simpler interface reduces gradient interference and provides cleaner pattern access.

**SignSGD and low-precision optimization.** Bernstein et al. (2018) established that sign-only gradient updates can match full-precision optimization under appropriate conditions. In pattern-access terms, this asks how much of the gradient signal is essential for the interface to maintain pattern fidelity. At $n = 3$, our Experiment 3 signal was ambiguous: gradient degradation appeared neutral. At $n = 30$, the signal resolved clearly: sign-only gradients significantly degrade performance (+5.0% final loss, $p = 0.002$, $d = +0.614$), and noisy $\sigma = 0.1$ also degrades (+2.4%, $p = 0.032$). The discrepancy between our finding and the SignSGD literature may reflect our model's small scale and short training duration. At $n = 300$, this finding strengthens: sign-only gradients degrade by +4.9% ($p < 0.0001$, $d = 0.575$), confirming that the discrepancy with SignSGD is robust and not a small-sample artifact.

**Noise as regularization.** The regularizing effect of gradient noise is well-established (Neelakantan et al., 2015). Our Experiment 3 shows that small noise ($\sigma = 0.01$) produces no significant change ($p = 0.843$), consistent with noise-as-regularization, while large noise ($\sigma = 0.1$) significantly degrades final loss at $n = 30$ ($+2.4\%$, $p = 0.032$) and strongly degrades mean trajectory loss (+2.4%, $p < 0.001$, $d = +4.306$), indicating a fidelity threshold between $\sigma = 0.01$ and $\sigma = 0.1$, below which the interface maintains clean pattern access, above which it does not.

**Local learning rules.** Alternatives to end-to-end backpropagation include greedy layerwise pretraining (Bengio et al., 2007), local learning signals (Nokland & Eidnes, 2019), and forward-forward algorithms (Hinton, 2022). In the Platonic Space framework, local learning asks whether the pattern is invariant to the communication structure of the interface, whether it can be accessed through independent channels rather than a single coordinated one. Our cell-view experiment uses local loss (layerwise cross-entropy), eliminating *all* inter-layer gradient flow. The resulting final loss is equivalent to baseline at both $n = 30$ ($-0.9\%$, $p = 0.237$) and $n = 300$ ($-0.0\%$, $p = 0.90$), with the mean trajectory showing a significant cost ($+0.8\%$, $p < 0.001$, $d = +1.302$). The same pattern is accessed through a structurally different interface.

**Perturbation analysis in neural networks.** Ablation studies (Meyes et al., 2019), dropout, and pruning are standard tools, but typically measure *component importance*. We use perturbation to characterize *what patterns the system accesses and how robustly*, not which parts matter, but what the system is doing at a deeper level. This distinction connects our work to Levin's framework rather than to standard ablation methodology.

**Levin's morphogenetic framework.** Levin et al. (2024) applied developmental biology concepts to simple sorting algorithms. Key findings included delayed gratification (temporary performance decrease followed by recovery past pre-damage levels) and fault tolerance that exceeded intact system performance. Our results show the transformer accesses patterns not directly prescribed by the optimizer, most clearly temporal pattern access in stress inoculation (Experiment 9), pattern re-binding in complete recovery (Experiment 7), functional role patterns in complete regeneration (Experiment 10), and interface simplification in head-freezing trajectory improvement (Experiment 1). Cell-view (local loss) achieves equivalent final-loss convergence, demonstrating pattern invariance across interface architectures. The Delayed Gratification Index, which Levin used to detect rerouting, shows no perturbation response at $n = 30$.

**Stress inoculation.** The phenomenon of gradual stressor exposure building tolerance is well-documented in biology (Meichenbaum, 1985) and metallurgy (work hardening). In deep learning, curriculum learning (Bengio et al., 2009) and noise scheduling in diffusion models provide partial analogs. Our Experiment 9 demonstrates stress inoculation in the gradient noise domain, a free lunch from the latent space: the optimizer did not specify that perturbation history should change which patterns are accessible, but it does.

**Distributed chess as collective intelligence.** Kofman, Campitelli & Levin (2025) extended the morphogenetic framework to chess with autonomous pieces. Our Experiment 6 uses a proper $2 \times 2$ factorial design with composite perturbations (each cell has both a forward and gradient perturbation). The key finding is that gradient type dominates: sign-only gradients (+5.2–6.5% final loss) degrade far more than noisy $\sigma = 0.1$ gradients (+2.4–2.6%), regardless of forward perturbation type (sign-only − noisy: $-3.8\%$, $p = 0.022$ at $n = 30$; $-2.9\%$, $p < 0.0001$ at $n = 300$). Their information bottleneck finding (intermediate vision radius outperforms omniscience) does not replicate for attention windowing (no significant final-loss effects at $n = 30$).

## 3. Methods

### 3.1 Model Specification

We use a minimal character-level GPT with the following architecture:

| Parameter | Value |
|---|---|
| Layers | 4 |
| Embedding dimension | 16 |
| Attention heads per layer | 4 (16 total) |
| Head dimension | 4 |
| Context length | 16 |
| Vocabulary | ~28 (a-z + special tokens) |
| Normalization | RMSNorm |
| Activation | ReLU |
| Total parameters | ~13,400 |

The model is implemented in a numpy backend for experiment sweeps. The task is character-level name generation trained on a dataset of ~32k names.

The choice of a minimal model is deliberate and follows Levin et al.'s rationale for using sorting algorithms: "the benefit of these sorting algorithms is precisely that they are simple, easy to understand, and offer no place for additional complexity to hide." A minimal interface makes it easier to quantify what patterns ingress and what free lunches appear, complexity cannot hide in an architecture this small.

### 3.2 Three-Scale Protocol

We adopt an explicit three-scale protocol for interpreting findings. At $n = 3$ (pilot scale), signal is visible but resolution is too low to distinguish real effects from sampling noise, findings at this scale are coarse signal, not conclusions. At $n = 30$ (primary scale), paired statistical analysis ($df = 29$) resolves ambiguity: effects that are real at moderate magnitude ($d \geq 0.4$) become detectable with 80% power. At $n = 300$ (fine-structure scale), effects that are small but real at $n = 30$ either strengthen into clear signals or remain null; and new structure that was below the detection threshold at $n = 30$ may emerge.

A finding that is ambiguous at $n = 3$ and resolves clearly at $n = 30$ is not a retraction, it is a finding at the appropriate resolution. A signal that appeared at $n = 3$, resolved to null at $n = 30$, and was confirmed null at $n = 300$ is a signal whose character stabilized as resolution increased; conversely, a signal that was significant at $n = 30$ ($p = 0.032$) and strengthened to $p = 0.0001$ at $n = 300$ is a signal whose precision increased at higher resolution. This is how science works when the resolution dial turns.

### 3.3 Delayed Gratification (DG) Index

Following Levin et al., we define a metric to detect rerouting behavior, episodes where the system temporarily moves *away* from its goal before recovering past the pre-perturbation level.

**Episode detection.** We scan the loss trajectory for episodes where: (1) loss increases from a local value $L_{\text{start}}$ to a peak $L_{\text{peak}}$, then (2) decreases to a trough $L_{\text{trough}}$ below $L_{\text{start}}$. Each such episode has:

- Temporary cost: $C = L_{\text{peak}} - L_{\text{start}}$
- Net gain: $G = L_{\text{start}} - L_{\text{trough}}$

**Per-episode DG:** $\text{DG}_{\text{episode}} = G/C$

**Aggregate DG Index:** The mean DG across all detected episodes in a training run.

At $n = 3$, the DG Index appeared to scale with perturbation severity. At $n = 30$, this signal resolved to null: no perturbation condition produces a statistically significant DG increase ($p > 0.19$ for all). DG captures a real stochastic property of loss trajectories but does not function as a perturbation response measure at $n = 30$. At $n = 300$, the DG null holds: the metric still does

not track perturbation, confirming that DG captures stochastic loss-trajectory structure rather than perturbation response.

### 3.4 Experiment 1: Head Freezing (Interface Simplification)

**Motivation:** Levin's frozen-cell perturbation. We simplify the interface by freezing randomly selected attention heads at initialization values throughout training, reducing the degrees of freedom through which optimization patterns manifest.

**Protocol:** Sweep over {0, 1, 2, 4, 8, 12, 16} frozen heads. Frozen heads participate in the forward pass but receive no gradient updates.

### 3.5 Experiment 2: Cell-View GPT (Pattern Invariance)

**Motivation:** Levin's cell-view sorting algorithms. Each transformer layer is treated as an autonomous agent. This tests whether the optimization pattern is invariant to the communication architecture of the interface, whether it can be accessed through independent channels.

**Protocol:** Local loss (layerwise cross-entropy) at all layer boundaries. Each layer receives only its own local loss signal, computed as a cross-entropy loss from a per-layer projection head. No inter-layer gradient flow.

### 3.6 Experiment 3: Gradient Degradation (Pattern Fidelity)

**Motivation:** Levin's noisy signaling channels. We degrade the gradient signal to probe the fidelity threshold, how much interface corruption the pattern tolerates before access is lost. Four corruption methods:

| Method | Description |
|---|---|
| Noisy ($\sigma = 0.01$) | Additive Gaussian noise, small scale |
| Sign-only | Gradient reduced to {-1, 0, +1}, magnitude discarded |
| Quantized (3-bit) | Gradient values rounded to 8 levels |
| Noisy ($\sigma = 0.1$) | Additive Gaussian noise, large scale |

### 3.7 Training Protocol and Statistical Methods

All experiments use: 200 training steps, 30 independent runs per condition (seeds 42–71), loss and per-head metrics recorded at every step. The 200-step horizon captures early learning dynamics.

**Statistical analysis.** All comparisons use two-sided paired $t$-tests, with runs matched by seed across conditions ($n = 30$, $df = 29$). Pairing by seed controls for initialization variance. We report effects as statistically significant at $p < 0.05$ and marginal at $0.05 < p < 0.10$. With 30 paired observations, statistical power is adequate to detect moderate effects (Cohen's $d \geq 0.4$ at 80% power). Effect sizes are reported as Cohen's $d$ for paired differences. We distinguish between *statistically supported* findings ($p < 0.05$) and *observational* patterns.

**Multiple comparisons.** We do not apply formal multiple-comparison correction across the twelve experiments. Each experiment tests a distinct perturbation type with its own pre-specified comparison, rather than screening a family of interchangeable hypotheses. We note that with twelve primary comparisons, some marginal results ($0.01 < p < 0.05$) should be interpreted with appropriate caution. The strongest findings ($p < 0.0001$) survive any reasonable correction.

**Confirmatory vs. exploratory.** We distinguish three categories: pre-specified primary comparisons (each experiment's main effect), secondary metrics (mean trajectory loss, DG index), and fine-structure analyses (layer-specific or dose-response patterns discovered at $n = 300$). Only primary comparisons are treated as confirmatory; secondary and fine-structure findings are explicitly exploratory.

### 3.8 Experiment 4: Vision Radius Sweep (Pattern Visibility)

**Motivation:** Kofman, Campitelli & Levin's (2025) vision radius experiment in distributed chess. We restrict each attention head's context window, probing the information geometry of pattern access, how much of the input the interface needs to see for the pattern to manifest.

**Protocol:** Sweep over window sizes $\{1, 2, 4, 8, 16\}$ plus unmodified baseline. Window=16 equals block size (sanity check).

### 3.9 Experiment 5: Communication Topology (Layerwise Autonomy)

**Motivation:** The chess paper's relay chains. We scale the fraction of gradient signal passed through layer boundaries, testing whether the pattern is accessible through independent channels or requires coordinated inter-layer communication.

**Protocol:** Five topologies parameterized by gradient pass fraction:

| Topology | Pass Fraction | Description |
|---|---|---|
| Full | 1.00 | Standard backpropagation (baseline) |
| Heavy | 0.75 | 75% of gradient signal passes through |
| Half | 0.50 | 50% of gradient signal passes through |
| Light | 0.25 | 25% of gradient signal passes through |
| Cell-view | 0.00 | No inter-layer gradient flow |

### 3.10 Experiment 6: Courage vs. Caution (Channel Sensitivity)

**Motivation:** Kofman et al.'s "cautious position, courageous moves" strategy. We create a $2 \times 2$ factorial design with composite perturbations, each cell applies *both* a forward perturbation and a gradient perturbation simultaneously, to probe which interface channel is the critical signal for pattern access:

| | Cautious Gradients (sign-only) | Courageous Gradients (noisy $\sigma = 0.1$) |
|---|---|---|
| **Cautious Forward** (tiny noise $\sigma = 0.001$) | (a) Tiny noise + sign-only | (b) Tiny noise + noisy $\sigma = 0.1$ |
| **Courageous Forward** (dropout $p = 0.1$) | (c) Dropout + sign-only | (d) Dropout + noisy $\sigma = 0.1$ |

### 3.11 Experiment 7: Recovery After Damage (Pattern Re-Binding)

**Motivation:** Levin's regeneration paradigm, does a damaged organism recover after the damage is removed? Does the pattern re-bind to the interface after disruption?

**Protocol:** Three-phase training with matched controls. - Phase 1: Normal training (200 steps) - Phase 2: Damage, freeze 8 random heads (100 steps) - Phase 3: Recovery, unfreeze all heads, continue (200 steps) - Control: Undamaged training for the same total duration (500 steps)

Learning rate schedule is matched across phases and control ($\text{lr}(t) = \text{lr}_0 \cdot (1 - t/500)$). Paired by seed ($n = 30$).

### 3.12 Experiment 8: Chimera Assembly (Basin Universality)

**Motivation:** Chimeric organisms assembled from parts of different embryos. Can a network assembled from independently-trained components access the same pattern? This tests whether the attractor in the latent space is accessible from any valid interface configuration.

**Protocol:** Train model A (seeds 42–71) and model B (seeds 1042–1071) independently for 200 steps. Assemble four chimera types by selecting each layer from either model:

| Chimera | Layer 0 | Layer 1 | Layer 2 | Layer 3 |
|---|---|---|---|---|
| AABB | A | A | B | B |
| ABAB | A | B | A | B |
| BBAA | B | B | A | A |
| ABBA | A | B | B | A |

Shared parameters (embeddings, output projection) come from model A. Continue training each chimera for 200 more steps. Control: model A continues training without modification.

### 3.13 Experiment 9: Gradual vs. Sudden Damage (Temporal Pattern Access)

**Motivation:** Biological stress inoculation, gradual exposure to a stressor builds tolerance that sudden exposure does not. In pattern-access terms: does the history of interface degradation change which patterns become accessible?

**Protocol:** Four conditions, all 200 steps: - Control: no gradient noise - Sudden full: noisy gradients ($\sigma = 0.1$) for all 200 steps - Gradual: linear ramp from $\sigma = 0$ to $\sigma = 0.1$ over 200 steps - Sudden half: no noise for first 100 steps, then $\sigma = 0.1$ for remaining 100 steps

The gradual condition reaches the same peak noise level as sudden full but arrives there incrementally.

### 3.14 Experiment 10: Regeneration (Functional Role Patterns)

**Motivation:** Regeneration after tissue destruction. Can a trained network rebuild a destroyed layer? In the Platonic Space framework, this tests whether the pattern determines the functional role of a position, regardless of what substrate occupies it.

**Protocol:** Two-phase training. - Phase 1: Normal training (200 steps) - Phase 2: Reset one layer's weights to random initialization (new random seed). Zero the corresponding Adam optimizer state. Continue training for 200 more steps. - Control: no reset, continue training. - Test all 4 layers independently. Paired by seed ($n = 30$).

### 3.15 Experiment 11: Transplantation (Context-Dependent Roles)

**Motivation:** Organ transplantation, is a layer from a separately-trained donor network accepted better than a random replacement? This tests whether the pattern is positional (determined by context) or substrate-bound (determined by the weights themselves).

**Protocol:** Train model A and model B independently (200 steps each, different seeds). For each layer $L$: - Transplant: replace layer $L$ of model A with layer $L$ from model B, continue training 200 steps - Random reset: replace layer $L$ of model A with random weights, continue training 200 steps - Control: model A continues without modification

Adam buffers are zeroed for the replaced layer in both transplant and random conditions. Paired by seed ($n = 30$).

### 3.16 Experiment 12: Competing Objectives (Pattern Corruption)

**Motivation:** Inter-organ conflict, what happens when part of the network fights the objective? This distinguishes between pattern unavailability (the interface is silent) and pattern corruption (the interface actively inverts the signal).

**Protocol:** Two-phase training. - Phase 1: Normal training (200 steps) - Phase 2: Negate gradients for layers 2-3 while layers 0-1 train normally (200 steps). The negated layers receive gradient signals that push them *away* from the loss minimum. - Comparison: freeze layers 2-3 (zero their gradients) instead of negating - Control: normal training for 400 steps total

This tests whether layers 0-1 can compensate for actively adversarial downstream layers (negation) versus merely inactive ones (freezing).

## 4. Results

### 4.1 Experiment 1: Head Freezing: Interface Simplification

**At $n = 3$:** The signal was ambiguous. Head freezing appeared to improve final loss, freezing 8+ heads seemed to produce statistically significant improvement ($p = 0.009$ in pilot data). The coarse resolution could not distinguish this from noise.

**At $n = 30$:** At this scale, some final-loss improvements appear significant: freeze 2 ($-0.3\%$, $p = 0.025$, $d = -0.430$), freeze 4 ($-0.5\%$, $p = 0.001$, $d = -0.647$), freeze 8 ($-0.5\%$, $p = 0.023$, $d = -0.438$), freeze 12 ($-0.7\%$, $p = 0.016$, $d = -0.467$). Freeze 1 ($p = 0.253$) and freeze 16 ($p = 0.322$) are non-significant. The overall monotonic trend is weak (Spearman $\rho = -0.015$, $p = 0.83$). Mean trajectory loss shows a robust improvement for 4+ frozen heads, with larger effect sizes than final loss. This is a different, finer signal that became visible at higher resolution.

**At $n = 300$:** The final-loss improvements resolve to null (all $p > 0.15$; Spearman $\rho = -0.0045$, $p = 0.84$). The $n = 30$ final-loss significances were artifacts of moderate sample size. The trajectory improvement strengthens to high significance: freeze 4 mean-loss $\Delta = -0.1\%$ ($p < 0.0001$, $d = -0.971$), freeze 8 $\Delta = -0.1\%$ ($p < 0.0001$, $d = -1.245$), freeze 12 $\Delta = -0.2\%$ ($p < 0.0001$, $d = -1.421$), freeze 16 $\Delta = -0.2\%$ ($p < 0.0001$, $d = -1.312$). The monotonic dose-response in trajectory improvement, combined with no final-loss cost at $n = 300$, confirms that frozen random-projection heads reduce gradient interference throughout training without affecting convergence.

**Table 1.** Head freezing results (means ± std across 30 runs).

| Frozen Heads | Final Loss | Mean Loss | DG Index |
|---|---|---|---|
| 0 (baseline) | 2.557 ± 0.407 | 2.627 ± 0.028 | 0.680 ± 1.131 |
| 1 | 2.553 ± 0.406 | 2.627 ± 0.028 | 0.632 |
| 2 | 2.549 ± 0.405 | 2.626 ± 0.028 | 0.689 |
| 4 | 2.543 ± 0.397 | 2.625 ± 0.028 | 0.636 |
| 8 | 2.544 ± 0.396 | 2.623 ± 0.028 | 0.640 |
| 12 | 2.539 ± 0.400 | 2.622 ± 0.028 | 0.614 |
| 16 | 2.547 ± 0.407 | 2.620 ± 0.028 | 0.570 |

Final loss shows improvements at several freezing levels: freeze 2 ($p = 0.025$, $d = -0.430$), freeze 4 ($p = 0.001$, $d = -0.647$), freeze 8 ($p = 0.023$, $d = -0.438$), freeze 12 ($p = 0.016$, $d = -0.467$). However, these resolve to null at $n = 300$ (all $p > 0.15$). Mean trajectory loss shows a robust improvement for 4+ frozen heads: freeze 4 ($p < 0.001$, $d = -1.008$), freeze 8 ($p < 0.001$, $d = -1.228$), freeze 12 ($p < 0.001$, $d = -1.366$), freeze 16 ($p < 0.001$, $d = -1.070$). The effect size is 0.1–0.3% of mean loss, statistically robust but practically small.

Simplifying the interface, fewer degrees of freedom through which the optimizer operates, provides cleaner access to the optimization pattern. The frozen heads reduce destructive gradient interactions, allowing the remaining parameters to navigate the loss landscape more efficiently. The

pattern itself (the attractor) is unchanged; the interface through which it is accessed is simplified.

The DG Index does not increase with freezing. No freezing level produces a significant DG change.

**Free lunch:** Trajectory improvement without paying for it. The optimizer specifies loss minimization; it does not specify that removing parameters should improve the trajectory. The improvement is a free lunch, a pattern the system accesses from the latent space that was not prescribed by the optimization objective.

**Classification:** Pattern manifestation (the attractor is reached regardless of interface simplification) with a free lunch (trajectory improvement).

### 4.2 Experiment 2: Cell-View GPT: Pattern Invariance

**At $n = 3$:** The signal was ambiguous. Cell-view appeared to elevate DG substantially (+25.5%), suggesting possible rerouting behavior.

**At $n = 30$:** The DG signal resolved to null ($p = 0.61$). The final-loss signal is non-significant: cell-view produces near-identical final loss to baseline ($-0.9\%$, $p = 0.237$, $d = -0.220$). The mean trajectory loss shows a significant cost ($+0.8\%$, $p < 0.001$, $d = +1.302$).

**At $n = 300$:** The final-loss equivalence is confirmed at high power: cell-view final loss $-0.0\%$ ($p = 0.90$, $d = -0.007$). Mean trajectory: $+0.2\%$ ($p < 0.0001$, $d = +0.731$). DG: $p = 0.14$ (ns). Local layerwise optimization converges to the same basin as end-to-end backpropagation, with only a slight trajectory cost.

**Table 2.** Cell-view (local loss) vs. baseline (means ± std across 30 runs).

| Condition | Mean Loss | Final Loss | DG Index |
|---|---|---|---|
| Baseline | 2.627 ± 0.028 | 2.557 ± 0.407 | 0.680 |
| Cell-view | 2.647 ± 0.025 | 2.535 ± 0.384 | 0.568 |

Replacing end-to-end backpropagation with local loss produces equivalent final loss ($-0.9\%$, $p = 0.237$, $d = -0.220$) and a significant mean-trajectory cost ($+0.8\%$, $p < 0.001$, $d = +1.302$). The same attractor is reached via a structurally different interface architecture: independent layerwise channels rather than a single coordinated gradient flow.

This is pattern invariance: the pattern in the latent space is the same regardless of whether the interface uses end-to-end backpropagation or local layerwise optimization. The loss landscape guides each layer to its functional role independently. The pattern does not depend on the communication structure of the interface.

**Free lunch:** Global convergence from local rules. Each layer optimizes its own local objective, yet the system converges to the same global minimum as if it had access to end-to-end gradient information. The global pattern is accessed for free, it was not specified by the local objectives.

**Classification:** Pattern invariance, the same pattern is accessible through multiple interface architectures.

### 4.3 Experiment 3: Gradient Degradation: Pattern Fidelity

**At $n = 3$:** The signal was ambiguous. All four gradient degradation methods appeared neutral ($p > 0.26$), and small noise appeared to help.

**At $n = 30$:** The ambiguous signal resolved. Three of four methods significantly degrade final loss; one is genuinely tolerated.

**At $n = 300$:** The threshold between fidelity and loss of access sharpens. Noise at $\sigma = 0.01$ remains non-significant ($-0.2\%$, $p = 0.28$), confirming genuine pattern fidelity, the pattern manifests cleanly through this level of interface degradation. All three degradation conditions strengthen: noise at $\sigma = 0.1$ (+2.2%, $p < 0.0001$, $d = 0.367$), sign-only (+4.9%, $p < 0.0001$, $d = 0.575$), and quantized 3-level (+3.6%, $p < 0.0001$, $d = 0.529$). The sign-only effect strengthened from $p = 0.002$ at $n = 30$ to $p < 0.0001$ at $n = 300$.

**Table 3.** Gradient degradation results (means across 30 runs).

| Method | Final Loss | $\Delta\%$ | $p$ (final) | Mean Loss | $p$ (mean) | $d$ (mean) |
|---|---|---|---|---|---|---|
| Baseline | 2.557 | , | , | 2.627 | , | , |
| Noisy ($\sigma = 0.01$) | 2.560 | +0.1% | 0.843 | 2.628 | 0.333 | +0.180 |
| Noisy ($\sigma = 0.1$) | 2.618 | +2.4% | 0.032* | 2.690 | <0.001*** | +4.306 |
| Sign-only | 2.685 | +5.0% | 0.002** | 2.729 | <0.001*** | +5.696 |
| Quantized (3-bit) | 2.653 | +3.8% | 0.008** | 2.693 | <0.001*** | +4.457 |

Three of four methods significantly degrade final loss: sign-only (+5.0%, $p = 0.002$, $d = +0.614$), quantized (+3.8%, $p = 0.008$, $d = +0.519$), and noisy $\sigma = 0.1$ (+2.4%, $p = 0.032$, $d = +0.411$). Noisy $\sigma = 0.1$ also strongly degrades mean trajectory (+2.4%, $p < 0.001$, $d = +4.306$). Only small noise ($\sigma = 0.01$) is genuinely tolerated ($p = 0.843$ final, $p = 0.333$ mean). Mean loss effects are highly significant for sign-only and quantized ($p < 0.001$) with large Cohen's $d$ values (4.5–5.7). The "noise helps" effect (noisy $\sigma = 0.01$ improving loss) is not supported ($p = 0.333$ for mean loss).

There is a sharp fidelity threshold between $\sigma = 0.01$ and $\sigma = 0.1$. Below this threshold, the interface maintains clean access to the optimization pattern. Above it, the interface degradation is sufficient to disrupt pattern access. The gradient sign structure carries more essential signal than magnitude, destroying sign information (sign-only) degrades more than destroying magnitude information (noise).

**Classification:** Pattern fidelity, the system maintains clean pattern access up to a threshold ($\sigma = 0.01$). Above that threshold, the interface can no longer faithfully transmit the pattern.

### 4.4 Experiment 4: Vision Radius Sweep: Pattern Visibility

**At $n = 3$:** The signal was ambiguous. An information bottleneck effect appeared possible, intermediate window sizes seemed to outperform full context.

**At $n = 30$:** The ambiguous signal resolved to null for final loss across all window sizes. Tiny mean-trajectory effects are statistically detectable but not practically meaningful.

**At $n = 300$:** Fine structure emerges that was invisible at $n = 30$. Window 1 significantly harms performance (+0.3%, $p = 0.021$, $d = +0.134$), windows 2 and 4 remain non-significant ($p = 0.93$ and $p = 0.93$ respectively), window 8 produces a small but significant improvement ($-0.1\%$, $p = 0.022$, $d = -0.133$), and window 16 is identical to baseline. This reveals a monotonic structure: extreme restriction harms, moderate restriction is neutral, and a mild restriction slightly benefits, a pattern consistent with beneficial information bottleneck effects that were below detection threshold at $n = 30$.

**Table 4.** Vision radius results (means across 30 runs).

| Window | Final Loss | $p$ (final) | Mean Loss | $p$ (mean) | $d$ (mean) |
|---|---|---|---|---|---|
| Baseline (full) | 2.557 | , | 2.627 | , | , |
| 1 | 2.562 | 0.618 | 2.639 | <0.001*** | +1.847 |
| 2 | 2.555 | 0.898 | 2.628 | 0.156 | , |
| 4 | 2.549 | 0.443 | 2.625 | <0.001*** | -0.735 |
| 8 | 2.553 | 0.304 | 2.626 | 0.037* | -0.400 |
| 16 | 2.557 | 1.000 | 2.627 | 1.000 | 0.00 |

No window size significantly changes final loss at $n = 30$ ($p > 0.30$ for all). Window=16 reproduces baseline exactly (sanity check). Mean-loss effects exist but are negligibly small. The chess paper's finding that intermediate vision radius outperforms omniscience does not produce meaningful effects for attention windowing at this scale.

The information geometry of pattern access in this architecture is flat: the pattern is accessible regardless of how much context each attention head sees. Only extreme restriction (window 1) begins to harm, suggesting that the pattern requires minimal but non-trivial visibility to manifest.

**Classification:** Pattern fidelity, the pattern manifests through the interface at all tested visibility levels.

### 4.5 Experiment 5: Communication Topology: Layerwise Autonomy

**At $n = 3$:** The signal was ambiguous. A U-shaped loss curve appeared possible, partial communication seemed to outperform both full and no communication.

**At $n = 30$:** The U-shaped curve resolved to flat across all gradient fractions, including zero communication (cell-view). Partial gradient flow is absorbed without meaningful degradation.

**At $n = 300$:** The architecture's indifference to gradient fraction holds at high power. Heavy ($p = 0.92$), half ($p = 0.033$), and light ($p = 0.59$) communication topologies remain largely non-significant. Cell-view final loss is non-significant ($p = 0.90$), consistent with Experiment 2's finding that local loss achieves equivalent convergence. Only cell-view mean trajectory shows significant cost ($+0.8\%$, $p < 0.0001$, $d = +0.731$).

**Table 5.** Communication topology results (means across 30 runs).

| Topology | Fraction | Final Loss | $p$ (final) | Mean Loss | $p$ (mean) |
|---|---|---|---|---|---|
| Full | 1.00 | 2.557 | , | 2.627 | , |
| Heavy | 0.75 | 2.558 | 0.379 | 2.627 | 0.314 |
| Half | 0.50 | 2.555 | 0.589 | 2.627 | 0.187 |
| Light | 0.25 | 2.557 | 0.915 | 2.627 | 0.895 |
| Cell-view | 0.00 | 2.535 | 0.237 | 2.647 | <0.001*** |

Partial gradient flow (25–75%) produces no meaningful degradation. Heavy (75%), half (50%), and light (25%) are all non-significant for final loss ($p > 0.37$). Cell-view (0%) shows equivalent final loss ($-0.9\%$, $p = 0.237$, $d = -0.220$) but significantly elevated mean trajectory loss ($+0.8\%$, $p < 0.001$, $d = +1.302$).

The pattern is accessible through independent channels. Even total removal of inter-layer communication does not prevent the system from reaching the attractor. This is layerwise autonomy: each layer independently accesses its portion of the global pattern without requiring coordination from the whole.

**Classification:** Pattern fidelity, the pattern manifests through the interface at all tested communication levels, including zero inter-layer communication.

### 4.6 Experiment 6: Courage vs. Caution: Channel Sensitivity

**At $n = 3$:** The signal was ambiguous. The courage/caution matrix appeared to produce inconsistent results without clear pattern.

**At $n = 30$:** The $2 \times 2$ factorial design with composite perturbations reveals a clear pattern: gradient type dominates over forward perturbation type.

**At $n = 300$:** The gradient-type dominance is confirmed at high statistical power. All four composite conditions significantly degrade final loss: cautious/cautious ($+5.2\%$, $p < 0.0001$, $d = +0.624$), cautious/courageous ($+1.9\%$, $p < 0.0001$, $d = +0.318$), courageous/cautious ($+5.0\%$, $p < 0.0001$, $d = +0.616$), courageous/courageous ($+2.5\%$, $p < 0.0001$, $d = +0.419$). The sign-only vs. noisy gradient contrast: $-2.9\%$ ($p < 0.0001$, $d = -0.355$). These n=300 results confirm the pattern already significant at n=30.

**Table 6.** Courage vs. caution results, $2 \times 2$ factorial with composite perturbations (means across 30 runs).

| Condition | Final Loss | $p$ (final) | Mean Loss | $p$ (mean) | $d$ (mean) |
|---|---|---|---|---|---|
| Baseline | 2.557 | , | 2.627 | , | , |
| (a) Caut./Caut. (tiny noise + sign-only) | 2.689 | 0.002** | 2.729 | <0.001*** | +5.935 |
| (b) Caut./Cour. (tiny noise + noisy $\sigma$=0.1) | 2.619 | 0.032* | 2.690 | <0.001*** | +4.309 |
| (c) Cour./Caut. (dropout + sign-only) | 2.722 | 0.001*** | 2.730 | <0.001*** | +5.155 |
| (d) Cour./Cour. (dropout + noisy $\sigma$=0.1) | 2.622 | 0.035* | 2.695 | <0.001*** | +4.479 |

The key finding is that **gradient type dominates**: conditions with sign-only gradients (a, c) degrade final loss by +5.2–6.5%, while conditions with noisy $\sigma = 0.1$ gradients (b, d) degrade by +2.4–2.6%. All four conditions significantly degrade final loss ($p < 0.04$), but sign-only conditions show roughly double the effect. The forward perturbation type (tiny noise vs. dropout) has a smaller effect. The sign-only vs. noisy gradient contrast (averaging across forward types) yields $-3.8\%$ final loss ($p = 0.022$, $d = -0.440$ at $n = 30$; $-2.9\%$, $p < 0.0001$, $d = -0.355$ at $n = 300$) and $-1.5\%$ mean loss ($p < 0.001$, $d = -1.921$). All conditions also strongly degrade mean trajectory loss.

The interface has an asymmetric channel sensitivity: the gradient channel is the critical signal through which the optimization pattern is accessed. Degrading the gradient sign structure disrupts pattern access far more than degrading the forward signal. The interface is not uniformly fragile, it has a critical channel (gradients) and a robust channel (forward activations).

**Classification:** Channel sensitivity, the gradient sign is the critical interface signal for pattern access. Forward perturbation is secondary.

### 4.7 Experiment 7: Recovery After Damage: Pattern Re-Binding

**At $n = 3$:** The signal was ambiguous. Recovery appeared complete but with too few observations to confirm.

**At $n = 30$:** Complete recovery is unambiguously confirmed across all 30 runs.

**At $n = 300$:** A tiny but statistically significant residual emerges at high power. Recovery vs. control: $+0.1\%$ ($p = 0.030$, $d = +0.126$). Final-loss ratio: $1.0009 \pm 0.0072$. 272 of 300 runs recovered, with mean overshoot $-0.0009 \pm 0.0017$. Recovery time: $0.8 \pm 1.2$ steps. The effect is real but negligibly small, the residual is within practical equivalence.

**Table 7.** Recovery after transient damage ($n = 30$, 500 total steps).

| Metric | Recovery | Control | $p$ (paired) |
|---|---|---|---|
| Final loss | 2.451 ± 0.370 | 2.451 ± 0.375 | 0.886 |
| Final ratio (rec/ctrl) | 1.0000 ± 0.008 | , | , |
| Recovery time | 1 ± 1 steps | , | 30/30 recovered |
| Overshoot | -0.0014 ± 0.0015 | , | , |

The damaged-then-recovered model reaches the same final loss as the undamaged control ($p = 0.886$, ratio 1.0000). All 30 runs recovered within a mean of 1 step after damage removal. No meaningful overshoot was observed (mean overshoot $= -0.0014$). The 100 steps of training with 8 frozen heads had no lasting effect, the "damage" was entirely absorbed by subsequent normal training.

The pattern re-binds to the interface after disruption. The 100 steps of constrained training forced the system through a different region of weight space, yet once the constraint was removed, the pattern re-manifested through the full interface as if the disruption had never occurred. The completeness of recovery ($p = 0.886$, ratio 1.0000) means the pattern is not fragile, it is an attractor that the interface converges to regardless of the path taken. At $n = 300$, a tiny residual becomes detectable ($p = 0.030$, $d = +0.126$) but remains within practical equivalence.

**Free lunch:** Complete recovery without a recovery mechanism. The optimizer prescribes convergence to *a* minimum; it does not prescribe that the system should return to the *same* minimum after a detour through a constrained subspace. The path-independent return is a free lunch, the optimizer did not pay for it.

**Classification:** Pattern re-binding, the pattern re-manifests through the interface after disruption. Free lunch: complete, path-independent recovery.

### 4.8 Experiment 8: Chimera Assembly: Basin Universality

**At $n = 3$:** The signal was ambiguous. Chimeras appeared to converge but from very few observations.

**At $n = 30$:** Convergence is confirmed for all chimera types. The specific layer assignment (which layers come from which model) does not matter.

**At $n = 300$:** All chimera types remain non-significant (AABB $p = 0.35$, ABAB $p = 0.12$, BBAA $p = 0.079$, ABBA $p = 0.31$), with all conditions falling within 0.3% of control ($n = 300$ ctrl = 2.4480). BBAA shows a marginal trend ($p = 0.079$) but does not reach significance. The basin of attraction is uniformly accessible from all tested chimera configurations.

**Table 8.** Chimera assembly results ($n = 30$, 200 + 200 steps).

| Condition | Initial Loss | Final Loss | vs Control $p$ |
|---|---|---|---|
| Control (A continues) | , | 2.494 ± 0.344 | , |
| AABB | 2.569 | 2.484 | 0.265 |
| ABAB | 2.563 | 2.494 | 0.985 |
| BBAA | 2.605 | 2.472 | 0.076† |
| ABBA | 2.542 | 2.487 | 0.559 |

All chimera types converge to the same final loss as the control ($p > 0.07$ for all). Despite starting at elevated loss (2.54–2.61 vs. control ~2.49), the chimeras reach control-equivalent performance.

The pattern in the latent space is accessible from any valid interface configuration. Two independently-trained models produce different weight configurations (different interfaces), but the chimera assembled from their parts accesses the same attractor. The pattern is not a property of a particular weight configuration, it is a property of the latent space, and any structurally valid interface converges to it.

**Classification:** Pattern invariance (basin universality), the pattern is accessible from any valid interface configuration.

### 4.9 Experiment 9: Gradual vs. Sudden Damage: Temporal Pattern Access

**At $n = 3$:** The signal was ambiguous. The gradual vs. sudden comparison appeared promising but underpowered.

**At $n = 30$:** The key finding resolves clearly and significantly. This is the paper's strongest result.

**At $n = 300$:** The temporal pattern access effect strengthens further, from $p = 0.032$ (final loss) at $n = 30$ to $p = 0.0001$ at $n = 300$. Sudden full noise degrades by +1.8% vs. control ($p < 0.0001$, $d = +0.318$). Gradual noise shows +0.5% degradation ($p = 0.017$, $d = +0.139$). The critical comparison, gradual vs. sudden full, yields $\Delta = -1.3\%$ ($p = 0.0001$, $d = -0.374$), confirming temporal pattern access as a robust phenomenon. Sudden half noise: $+0.8\%$ ($p = 0.0002$, $d = +0.219$). Gradual mean trajectory is significantly *below* control ($-0.1\%$, $p < 0.0001$, $d = -0.483$). The temporal pattern access signal is already significant at $n = 30$ and strengthens at $n = 300$.

**Table 9.** Gradual vs. sudden noise ($n = 30$, 200 steps).

| Condition | Final Loss | $p$ (vs ctrl) | Mean Loss | $p$ (mean) |
|---|---|---|---|---|
| Control | 2.557 ± 0.407 | , | 2.627 ± 0.028 | , |
| Sudden full ($\sigma = 0.1$) | 2.618 ± 0.388 | 0.032* | 2.690 ± 0.028 | <0.001*** |
| Gradual (0 to 0.1) | 2.564 ± 0.408 | 0.641 | 2.622 ± 0.027 | <0.001*** |
| Sudden half (step 100) | 2.582 ± 0.416 | 0.040* | 2.630 ± 0.028 | 0.004** |

**This is the paper's key finding.** Gradual exposure to noise changes which patterns are accessible. The gradually-ramped condition shows no significant degradation at $n = 30$ ($p = 0.641$), while sudden exposure to the same noise level significantly degrades ($+2.4\%$, $p = 0.032$). At $n = 30$, the direct gradual-vs-sudden comparison is significant for final loss ($-2.1\%$, $p = 0.032$, $d = -0.412$) and the effect is clear in mean trajectory loss: gradual mean is significantly *below* control ($-0.2\%$, $p < 0.001$, $d = -0.770$), while sudden full mean is significantly above ($+2.4\%$, $p < 0.001$, $d = +4.306$). At $n = 300$, the final-loss comparison strengthens further ($p = 0.0001$, $d = -0.374$).

**Why is this a free lunch?** The gradient update rule is identical in the sudden and gradual conditions at every step, the only difference is the *history* of noise levels. The optimizer at step $t$ does not remember what noise level was applied at step $t - 1$. Yet the system's final state depends on that history, and the gradually-exposed model reaches a region of weight space that the suddenly-exposed model does not. The history of interface degradation changes which patterns become accessible. The optimizer did not specify that gradual exposure should build tolerance; it specified only that loss should decrease. The tolerance is a free lunch from the latent space, a pattern the system accesses that was not prescribed by its optimization objective.

**Classification:** Temporal pattern access, perturbation history changes which patterns from the latent space are accessible. Free lunch: stress tolerance without prescription.

### 4.10 Experiment 10: Regeneration: Functional Role Patterns

**At $n = 3$:** The signal was ambiguous. Regeneration appeared possible but with too few observations to confirm the completeness.

**At $n = 30$:** Complete regeneration is confirmed for all four layers.

**At $n = 300$:** Fine structure emerges in layer-specific regeneration completeness. All four layers show small but significant residual deficits: Layer 0 ($+0.3\%$, $p = 0.003$, $d = +0.173$), Layer 1 ($+0.2\%$, $p = 0.007$, $d = +0.157$), Layer 2 ($+0.1\%$, $p = 0.024$, $d = +0.131$), Layer 3 ($+0.1\%$, $p = 0.037$, $d = +0.121$). Completeness: L1 0.988, L2 0.994, L3 1.021 (L0 has a data issue; use final loss comparison only). The overall confidence interval tightens substantially relative to $n = 30$, revealing that regeneration is near-complete but not perfectly uniform across layers. The residuals are tiny and graded by layer position.

**Table 10.** Layer regeneration after random reset ($n = 30$, 200 + 200 steps).

| Reset Layer | Immediate Damage | Final Loss | Completeness | vs Control $p$ |
|---|---|---|---|---|
| Control (no reset) | , | $2.457 \pm 0.341$ | , | , |
| Layer 0 | -0.136 | 2.467 | 0.998 | 0.174 |
| Layer 1 | -0.152 | 2.459 | 1.046 | 0.550 |
| Layer 2 | -0.157 | 2.456 | 1.015 | 0.808 |
| Layer 3 | -0.146 | 2.457 | 0.981 | 0.977 |

Complete regeneration at $n = 30$: all four layers recover to control-equivalent loss after being destroyed ($p > 0.17$ for all). All layers regenerate equally. The completeness values cluster near 1.0 (0.981–1.046).

The pattern determines the functional role of each position, not the substrate that occupies it. When a layer is destroyed and replaced with random weights, the new weights converge to the same functional role as the original. The position within the architecture, the context provided by the surrounding layers and the loss landscape, determines what function the layer serves. The pattern is positional, not substrate-bound. At $n = 300$, tiny residual deficits become detectable ($d = 0.12$–$0.17$), graded by layer position (deeper layers regenerate more completely), but all remain within practical equivalence.

**Free lunch:** Functional recovery from destruction. The optimizer specifies convergence to a minimum; it does not specify that a rebuilt layer should reach the same functional role as if it had never been destroyed. The functional role pattern is a free lunch, the system accesses it from the latent space without the optimizer having to specify it.

**Classification:** Functional role patterns, position determines role, not substrate. Free lunch: complete functional recovery.

#### 4.11 Experiment 11: Transplantation: Context-Dependent Roles

**At $n = 3$:** The signal was ambiguous. A transplant advantage appeared possible.

**At $n = 30$:** The transplant null result resolves clearly. There is no advantage to a structured (donor) layer over a random replacement.

**At $n = 300$:** The null holds across all layers with no exceptions. Layer-specific $p$-values: L0 $p = 0.29$, L1 $p = 0.44$, L2 $p = 0.98$, L3 $p = 0.91$. Overall $p = 0.76$. The transplant null is robust: donor layers confer no advantage over random reinitialization at any layer position, even at high statistical power.

**Table 11.** Transplant vs. random reset ($n = 30$, 200 + 200 steps).

| Layer | Transplant Final | Random Final | Gap (rand − trans) | $p$ |
|---|---|---|---|---|
| L0 | 2.470 | 2.467 | -0.0033 | 0.566 |
| L1 | 2.459 | 2.459 | +0.0008 | 0.804 |
| L2 | 2.459 | 2.456 | -0.0025 | 0.568 |
| L3 | 2.454 | 2.457 | +0.0036 | 0.420 |
| **Overall** | , | , | **+0.0003** | **0.880** |

Transplanted layers and randomly-reset layers converge to the same final loss ($p = 0.880$ overall). There is no transplant advantage, a layer from a separately-trained donor network is accepted no better and no worse than a random replacement. The network does not recognize or benefit from the structure of the donor layer; it simply rebuilds whatever is placed there.

The pattern is positional, not substrate-bound. A pre-trained layer from a different model and a random layer both converge to the same functional role at the same position. The context, the surrounding layers, the loss landscape, the optimization trajectory, determines what pattern manifests at each position. The substrate is irrelevant. This is consistent with Levin's (2026) framework: the physical interface does not determine the pattern; the pattern manifests through whatever interface is available.

**Classification:** Context-dependent roles, the pattern is positional, not substrate-bound. The interface at each position converges to the pattern determined by context, regardless of initial substrate.

### 4.12 Experiment 12: Competing Objectives: Pattern Corruption

**At $n = 3$:** The signal was ambiguous. The distinction between adversarial and inactive layers appeared but was underpowered.

**At $n = 30$:** The distinction resolves sharply. Adversarial layers degrade dramatically; inactive layers are tolerated.

**At $n = 300$:** The adversarial degradation strengthens: competing objectives degrade by +26.3% ($p < 0.0001$, $d = +0.531$), up from +24.8% at $n = 30$. High variance persists (std = 1.31). Frozen layers remain non-significant ($-0.1\%$, $p = 0.41$). The competing vs. freeze comparison yields $p < 0.0001$ ($d = +0.535$).

**Table 12.** Competing objectives ($n = 30$, 200 + 200 steps).

| Condition | Final Loss | vs Control Δ% | $p$ (vs ctrl) |
|---|---|---|---|
| Control (400 steps normal) | 2.457 ± 0.341 | , | , |
| Competing (negate L2-3 grads) | 3.067 ± 1.015 | +24.8% | <0.001*** |
| Freeze L2-3 | 2.461 ± 0.346 | +0.2% | 0.462 |

Negating gradients for layers 2-3 causes significant degradation (+24.8%, $p < 0.001$, $d = +0.689$)

with high variance (std = 1.01). But merely freezing those same layers causes negligible degradation (+0.2%, $p = 0.462$, $d = +0.136$). Competing vs. freeze: $p < 0.001$, $d = +0.693$.

This experiment reveals the critical distinction between pattern unavailability and pattern corruption. When layers are frozen, the interface is silent at those positions, the pattern is simply not accessed through those channels, and the remaining channels compensate. When layers receive negated gradients, the interface actively inverts the pattern, the optimization signal is corrupted, and the system cannot compensate. Silence (pattern unavailability) is tolerated; sabotage (pattern corruption) is catastrophic. The asymmetry is large: +0.2% vs. +24.8%.

**Classification:** Pattern corruption (adversarial) vs. pattern fidelity (freeze). The interface tolerates silence but not inversion. This maps directly to Levin's (2026) framework: the physical interface can fail to transmit a pattern (silence) or can actively distort it (corruption), and the consequences are qualitatively different.

### 4.13 Cross-Experiment Synthesis

**Table 13a.** Perturbation tolerance (Experiments 1-6, paired $t$-tests, $n = 30$).

| Perturbation | Final Loss Δ% | $p$ (final) | Mean Loss Δ% | $p$ (mean) | Classification |
|---|---|---|---|---|---|
| Freeze 1-2 heads | -0.1-0.3% | 0.025-0.253 | -0.0% | 0.002-0.053 | Pattern manifestation |
| Freeze 4-16 heads | -0.4-0.7% | 0.001-0.322 | -0.1-0.3% | <0.001 | Free lunch (trajectory) |
| Window=4,8 | -0.1-0.3% | >0.30 | -0.0-0.1% | 0.037-<0.001 | Pattern fidelity |
| Partial flow (25-75%) | ±0.1% | 0.379-0.915 | ±0.0% | >0.18 | Pattern fidelity |
| Noisy σ=0.01 | +0.1% | 0.843 | +0.0% | 0.333 | Pattern fidelity |
| Noisy σ=0.1 | +2.4% | 0.032 | +2.4% | <0.001 | Fidelity exceeded |
| Quantized 3-bit | +3.8% | 0.008 | +2.5% | <0.001 | Fidelity exceeded |
| Sign-only | +5.0% | 0.002 | +3.9% | <0.001 | Fidelity exceeded |
| Cell-view (local loss) | -0.9% | 0.237 | +0.8% | <0.001 | Pattern invariance |

**Table 13b.** Multi-phase perturbation results (Experiments 7-12, paired $t$-tests, $n = 30$).

| Experiment | Condition | vs Control | $p$ | Classification |
|---|---|---|---|---|
| 7: Recovery | Damaged then recovered | +0.0% | 0.886 | **Pattern re-binding** |
| 8: Chimera (AABB) | Frankenstein assembly | -0.4% | 0.265 | Pattern invariance |
| 8: Chimera (ABAB) | Alternating layers | -0.0% | 0.985 | Pattern invariance |
| 9: Sudden noise | $\sigma = 0.1$ all steps | +2.4% | 0.032* | Fidelity exceeded |
| 9: Gradual noise | Ramp 0 to 0.1 | +0.3% | 0.641 | **Temporal pattern access** |
| 9: Gradual vs sudden | Direct comparison | -2.1% | 0.032* | **Free lunch** |
| 10: Regeneration (any layer) | Reset then retrain | +0.1-0.4% | >0.17 | **Functional role pattern** |
| 11: Transplant vs random | Donor vs random layer | ±0.0% | 0.880 | Context-dependent role |
| 12: Adversarial L2-3 | Negate gradients | +24.8% | <0.001*** | Pattern corruption |
| 12: Freeze L2-3 | Zero gradients | +0.2% | 0.462 | Pattern fidelity |

**Classification under the pattern-access framework (full):**

*Pattern manifestation and free lunches*, the system accesses patterns not directly prescribed by the optimization objective: - **Temporal pattern access / stress inoculation (Exp 9):** Perturbation history changes which patterns are accessible. Gradual noise builds tolerance that sudden noise does not. The gradient rule is the same at every step; only the history differs. Free lunch: stress tolerance without prescription. - **Pattern re-binding / complete recovery (Exp 7):** The pattern re-manifests through the interface after disruption, converging to identical final loss. Free lunch: complete, path-independent recovery without a recovery mechanism. - **Functional role patterns / regeneration (Exp 10):** Position determines role, not substrate. Any single layer destroyed and rebuilt reaches control-equivalent performance. Free lunch: functional recovery from destruction. - **Interface simplification / head-freezing trajectory (Exp 1):** Fewer degrees of freedom provide cleaner pattern access. Frozen random-projection heads reduce gradient interference. Free lunch: trajectory improvement without paying for it.

*Pattern invariance*, the same pattern is accessible through multiple interfaces: - **Basin universality / chimera convergence (Exp 8):** Models assembled from incompatible parts converge to the same minimum. - **Context-dependent roles / transplant indifference (Exp 11):** No difference between donor and random layers, the pattern at each position is determined by context, not substrate. - **Pattern invariance / cell-view equivalent convergence (Exp 2):** Local loss converges to the same

final loss as end-to-end backpropagation. Free lunch: global convergence from local rules.

*Pattern fidelity*, the interface tolerates degradation up to a threshold: - **Gradient fidelity threshold (Exp 3):** Small noise tolerated; large noise exceeds fidelity. - **Layerwise autonomy (Exp 5):** 25–75% gradient flow tolerated without meaningful degradation. - **Pattern visibility (Exp 4):** All window sizes tolerated for final loss.

*Pattern corruption*, the interface actively inverts the pattern: - **Absence vs. corruption (Exp 12):** Frozen layers tolerated ($p = 0.462$); adversarial layers degrade +24.8% ($p < 0.001$). Silence is tolerated; sabotage is catastrophic.

**DG does not track perturbation.** At $n = 3$, the DG Index appeared to scale with perturbation severity. At $n = 30$, this signal resolved to null: no perturbation condition produces a statistically significant DG change. Baseline DG = $0.680 \pm 1.131$; all conditions are non-significant ($p > 0.19$ for all). At $n = 300$, the DG null holds ($p > 0.23$ for all): the metric still does not track perturbation severity, confirming that DG captures intrinsic stochastic structure of loss trajectories rather than perturbation response.

### 4.14 Findings: What Faultization Revealed

**Free Lunches** **Finding 1: Perturbation history changes which patterns are accessible (Exp 9).** Gradual noise ramp (0 to 0.1) produces no significant final-loss degradation at $n = 30$ ($p = 0.641$), while sudden exposure to the same noise level significantly degrades ($+2.4\%$, $p = 0.032$). At $n = 30$, the direct comparison is significant for final loss ($p = 0.032$, $d = -0.412$) and highly significant for mean trajectory. At $n = 300$, the final-loss comparison strengthens to $p = 0.0001$ ($d = -0.374$). The gradient update rule is identical at every step, only the history of noise levels differs. That history changes the system's final state. Free lunch: stress tolerance without prescription.

**Finding 2: Pattern re-binds after disruption (Exp 7).** A model damaged during training (8 frozen heads for 100 steps) recovers to identical final loss at $n = 30$ ($p = 0.886$, ratio 1.0000). All 30 runs recovered within a mean of 1 step. At $n = 300$, a tiny but significant residual emerges ($p = 0.030$, $d = +0.126$) but remains within practical equivalence. Free lunch: complete recovery without a recovery mechanism.

**Finding 3: Position determines functional role, not substrate (Exp 10).** Any single layer can be destroyed and rebuilt to control-equivalent performance at $n = 30$ ($p > 0.17$ for all layers). At $n = 300$, tiny residuals become detectable ($d = 0.12$–$0.17$) but all remain within practical equivalence. The network re-finds the same functional role regardless of what was there before. Free lunch: functional recovery from destruction.

**Finding 4: Interface simplification improves trajectory (Exp 1).** Freezing 4+ randomly-initialized heads produces small but statistically robust mean-trajectory improvements (freeze 8: $p < 0.001$, $d = -1.228$; freeze 12: $p < 0.001$, $d = -1.366$). Frozen random-projection heads reduce gradient interference. At $n = 30$, some final-loss improvements appear significant, but these resolve to null

at $n = 300$. Free lunch: trajectory improvement without paying for it.

**Pattern Invariance** **Finding 5: Basin universality, pattern accessible from any configuration (Exp 8).** Models assembled from parts of two independently-trained networks converge to the same final loss as undamaged continuation ($p > 0.07$ for all chimera types). The pattern in the latent space is accessible from any valid interface configuration.

**Finding 6: Context determines role, not substrate (Exp 11).** Transplanted layers and randomly-reset layers converge to the same final loss ($p = 0.880$ overall). The pattern at each position is determined by context (surrounding layers, loss landscape), not by the substrate (initial weights).

**Pattern Fidelity** **Finding 7: Gradient sign is the critical interface signal (Exp 3, 5).** Reducing gradient precision (sign-only: $d = +5.696$ for mean) degrades more than reducing gradient magnitude (partial flow: mostly ns) or completeness (freezing: improves trajectory). The architecture tolerates magnitude reduction but not sign-structure destruction. The gradient sign is the critical channel through which the optimization pattern is accessed.

**Finding 8: Channel asymmetry in pattern access (Exp 6).** In the $2 \times 2$ composite design, sign-only gradients (+5.2–6.5% final loss) degrade far more than noisy $\sigma = 0.1$ gradients (+2.4–2.6%), regardless of forward perturbation type. The sign-only vs. noisy gradient contrast: $-3.8\%$ ($p = 0.022$ at $n = 30$; $-2.9\%$, $p < 0.0001$ at $n = 300$, $d = -0.355$). The gradient channel is more sensitive to information destruction than the forward channel.

**Finding 9: Silence tolerated, sabotage catastrophic (Exp 12).** Frozen layers cost nothing ($p = 0.462$); adversarial layers cost +24.8% ($p < 0.001$, $d = +0.689$). Pattern unavailability (silence) is absorbed; pattern corruption (inversion) is catastrophic.

**Signals That Resolved Across Scales** At $n = 3$, the following signals were ambiguous; at $n = 30$, they resolved:

- **Head freezing improves final loss:** At $n = 3$, this appeared as a possible improvement signal. At $n = 30$, some final-loss improvements appear significant (freeze 4: $p = 0.001$), but at $n = 300$ all resolve to null (all $p > 0.15$). A different, finer signal persists at all scales: mean-trajectory improvement. The picture sharpened, the robust signal is in the trajectory metric, not the final-loss metric.
- **DG scales with perturbation:** At $n = 3$, the DG Index appeared to scale with perturbation severity. At $n = 30$, this resolved to null across all conditions ($p > 0.19$). At $n = 300$, the null is confirmed: DG still does not track perturbation, establishing that this metric captures intrinsic trajectory structure rather than perturbation response.
- **Gradient degradation is neutral:** At $n = 3$, all four methods appeared neutral. At $n = 30$, sign-only, quantized, and noisy $\sigma = 0.1$ all resolved to significant final-loss degradation; only $\sigma = 0.01$ remained tolerated. The resolution was too coarse to see the effects at $n = 3$.
- **Partial communication outperforms full:** At $n = 3$, a U-shaped curve appeared. At $n = 30$, this resolved to flat (except at zero). The pilot U-shape was sampling noise.

- **Noise helps:** At $n = 3$, small noise appeared beneficial. At $n = 30$, this resolved to null ($p = 0.333$ for mean, $p = 0.843$ for final). The apparent benefit was within noise.

## 5. Discussion

### 5.1 What Faultization Reveals About Pattern Access

During standard training, the transformer's components cooperate invisibly. Faultization, systematic perturbation of the interface, makes the system's relationship to the latent space legible by forcing it to operate under constraint. Each experiment degrades the interface differently, and the four-category classification organizes what we observe.

**Pattern manifestation and free lunches.** Four findings describe patterns the system accesses that were not directly prescribed by the optimization objective. Temporal pattern access (Exp 9): the history of interface degradation changes which patterns become accessible, despite identical gradient update rules at every step ($p = 0.0001$ at $n = 300$, $d = -0.374$). Pattern re-binding (Exp 7): the optimization pattern re-manifests through the full interface after transient disruption ($p = 0.886$ at $n = 30$, ratio 1.0000). Functional role patterns (Exp 10): position determines function, not substrate, destroyed layers rebuild to control-equivalent performance across all four layer positions ($p > 0.17$ at $n = 30$). Interface simplification (Exp 1): fewer degrees of freedom provide cleaner pattern access, reducing gradient interference during training ($p < 0.001$, $d = -1.228$ to $-1.366$).

**Pattern invariance.** Three findings, basin universality (Exp 8), context-dependent roles (Exp 11), and cell-view equivalent convergence (Exp 2), demonstrate that the same pattern is accessible through multiple interface architectures. Cell-view (local loss) achieves equivalent final loss to baseline ($p = 0.237$ at $n = 30$, $p = 0.90$ at $n = 300$), demonstrating that the pattern does not depend on the communication structure of the interface. The pattern is a property of the latent space, not of the particular interface through which it manifests.

**Pattern fidelity.** The system maintains clean pattern access up to a degradation threshold. Gradient noise at $\sigma = 0.01$ is tolerated; at $\sigma = 0.1$ it exceeds fidelity. Partial communication reduction (25–75%) is tolerated. Vision restriction is tolerated. The fidelity threshold is sharp: below it, the pattern manifests cleanly; above it, access degrades monotonically.

**Pattern corruption.** The absence-vs-corruption distinction (Exp 12) defines the qualitative boundary. Frozen layers (pattern unavailability) are absorbed ($p = 0.462$); adversarial layers (pattern corruption) degrade substantially ($p < 0.001$, +24.8%). The interface can fail to transmit (silence) or can actively distort (corruption), and the consequences are qualitatively different.

### 5.2 Free Lunch Quantification

For each experiment, we can distinguish what was specified (the algorithm) from what was received (the capability). The difference is the free lunch, the pattern the system accesses from the latent space without the optimizer paying for it.

| Experiment | What Was Specified | What Was Received | Free Lunch |
|---|---|---|---|
| 1: Head Freezing | Minimize loss | Trajectory improves with fewer DOF | Trajectory improvement |
| 2: Cell-View | Local loss per layer | Global convergence | Global pattern from local rules |
| 7: Recovery | Minimize loss | Same minimum after detour | Path-independent recovery |
| 9: Stress Inoculation | Minimize loss (same rule at every step) | History-dependent tolerance | Stress tolerance |
| 10: Regeneration | Minimize loss | Same functional role after destruction | Functional recovery |

In each case, the optimizer specifies only "minimize loss." The capabilities in the third column, trajectory improvement, global convergence from local rules, path-independent recovery, history-dependent tolerance, functional recovery from destruction, are not specified anywhere in the optimization objective. They are free lunches: patterns the system accesses from the latent space that the physical process of gradient descent did not pay for.

The remaining experiments characterize the interface rather than identifying free lunches. They map the fidelity boundary (Exp 3, 4, 5), the channel sensitivity (Exp 6), the invariance properties (Exp 8, 11), and the corruption threshold (Exp 12). Together, the twelve experiments provide a map of how this particular interface relates to the patterns in the latent space.

### 5.3 The Pattern-Interface Framework

The transformer is an interface through which optimization patterns manifest. The weights are the physical substrate; the loss landscape structure, attractors, basins, fidelity thresholds, reflects the patterns available in the latent space. Gradient descent is the process by which the interface converges to a pattern.

Faultization degrades the interface. The four categories capture what happens:

1. **Pattern manifestation**: the interface is degraded but the pattern still manifests. The attractor is robust enough that the system converges to it despite fewer parameters (Exp 1), different communication architectures (Exp 2, 5), restricted visibility (Exp 4), or chimeric construction (Exp 8).

2. **Pattern fidelity**: there exists a threshold below which the interface maintains clean access and above which it does not. The gradient noise threshold ($\sigma = 0.01$ vs. $\sigma = 0.1$ in Exp 3), the sign-structure sensitivity (Exp 6), and the communication topology results (Exp 5) map this boundary.

3. **Pattern corruption**: the interface does not merely fail to transmit; it actively inverts the signal. Gradient negation (Exp 12) is qualitatively different from gradient silencing. The pattern is

not just unavailable, a corrupted version is imposed.

4. **Free lunches**: the system receives capabilities that were not specified by the optimization objective. Stress tolerance (Exp 9), recovery (Exp 7), regeneration (Exp 10), and trajectory improvement (Exp 1) are all free lunches. The optimizer paid for "minimize loss"; it received stress tolerance, recovery, regeneration, and trajectory improvement at no additional cost.

This framework does not claim that patterns literally reside in a separate realm. It claims that the distinction between what the optimizer specified and what the system exhibits is real, measurable, and informative, and that the Platonic Space framework (Levin, 2026) provides a useful vocabulary for organizing these observations.

### 5.4 Connection to Levin's Platonic Space

Levin (2026) proposes that non-physical patterns are real, discovered, and causal; that physical systems are interfaces through which these patterns manifest; and that evolution, learning, and engineering exploit patterns as affordances they do not need to pay for. Our findings map onto this framework as follows.

**"Non-physical patterns are real" maps to loss landscape structure.** The attractor that all twelve experiments converge to (or fail to converge to, or converge to a corrupted version of) is a property of the loss landscape, not of any particular weight configuration. It is discovered by the optimizer, not created by it. Different interface configurations (chimeras, cell-view, head-frozen, transplanted) all access the same pattern. This is consistent with the pattern being real and independent of the interface.

**"Physical systems are interfaces" maps to transformer weights.** The weights, gradients, and communication channels are the physical substrate through which the optimization pattern manifests. Faultization degrades this substrate. The pattern manifests through the degraded interface up to a fidelity threshold, beyond which access is lost (Exp 3) or corrupted (Exp 12). The interface is not the pattern, it is the channel through which the pattern is accessed.

**"Free lunches" map to stress tolerance, recovery, and regeneration.** The optimizer specifies loss minimization. It does not specify stress tolerance (Exp 9), path-independent recovery (Exp 7), functional regeneration (Exp 10), or trajectory improvement under interface simplification (Exp 1). These capabilities appear without being paid for by the optimization process. In Levin's framework, they are patterns from the latent space that the physical process of gradient descent exploits as affordances.

The strength of this mapping should not be overstated. The Platonic Space interpretation is a framework for organizing findings, a way of asking "what did the system receive that it did not pay for?", rather than a falsifiable metaphysical claim about the ontological status of mathematical objects. What is falsifiable is the predictions it generates: that systematic perturbation of any sufficiently capable interface should reveal free lunches (capabilities not specified by the construction or training process), and that these free lunches should have a structured relationship to the

interface's architecture. Our twelve experiments are consistent with these predictions.

### 5.5 Connection to Distributed Chess

Kofman, Campitelli & Levin (2025) implemented a distributed form of chess where each piece operates as an autonomous agent. Experiments 4-6 tested three predictions; Experiments 7-12 extend the morphogenetic paradigm beyond the chess paper's framework.

**Information bottleneck as beneficial constraint (partially supported at $n = 300$).** The chess paper's central result, intermediate vision radius $R4$ outperforms omniscient $R7$, does not translate to attention windowing at $n = 30$ (all final-loss $p > 0.30$). At $n = 300$, however, fine structure emerges: window 1 significantly harms ($p = 0.021$, $d = +0.134$), window 8 significantly improves ($p = 0.022$, $d = -0.133$), and intermediate windows are neutral. This monotonic structure is consistent with a weak pattern visibility effect that was below detection threshold at lower power.

**Partial communication tolerance (confirmed as pattern fidelity, not improvement).** Reducing gradient flow to 25% produces no significant degradation, but partial flow does not *improve* over full backpropagation. The pattern fidelity is real; the U-shaped curve from $n = 3$ pilot data was noise.

**Courage/caution strategy (channel sensitivity).** In the $2 \times 2$ factorial design, gradient type is the primary driver: sign-only gradients (a, c) degrade final loss by +5.2–6.5%, while noisy $\sigma = 0.1$ gradients (b, d) degrade by +2.4–2.6%. The sign-only vs. noisy gradient contrast: $-3.8\%$ final ($p = 0.022$ at $n = 30$; $-2.9\%$, $p < 0.0001$ at $n = 300$); $-1.5\%$ mean ($p < 0.001$, $d = -1.921$). The gradient channel is the critical interface signal for pattern access.

**Stress inoculation (new, Exp 9).** The gradual-vs-sudden result ($p = 0.032$ at $n = 30$, strengthening to $p = 0.0001$ at $n = 300$, $d = -0.374$) has no direct chess analog but connects to the broader Platonic Space framework. That gradient descent exhibits temporal pattern access, the history of interface degradation changing which patterns become accessible, suggests this property may appear across optimization substrates, but whether chess or biological systems show the same phenomenon requires direct testing.

**Chimera convergence (new, Exp 8).** Unlike biological chimeras, which can develop abnormally at graft boundaries, transformer chimeras converge seamlessly. This reflects the smoothness of the loss landscape versus the discrete developmental signaling in biological systems, the interface is simpler, and the pattern invariance is correspondingly wider.

### 5.6 Scaling as Methodology

The three-scale protocol is not merely a replication strategy, it is a methodological commitment. At $n = 3$, the resolution is coarse. Real effects can be invisible; noise can masquerade as signal. At $n = 30$, the picture sharpens: moderate effects become detectable, and many $n = 3$ ambiguities resolve. At $n = 300$, fine structure emerges that $n = 30$ cannot see.

The practical implication: do not interpret findings at $n = 3$ as conclusions. They are coarse signal. The signal that head freezing appeared to improve final loss at $n = 3$ was not wrong, it was a low-resolution view of a real region of the parameter space. At $n = 30$, the final-loss effect resolved to null, but a different effect (trajectory improvement) became visible. The picture changed not because the $n = 3$ finding was retracted but because the resolution increased.

The $n = 300$ results confirm this framing. Some non-significant findings at $n = 30$ revealed fine structure at $n = 300$ (vision radius, stress inoculation final loss from $p = 0.032$ to $p = 0.0001$, regeneration layer specificity). Some robust findings at $n = 30$ strengthened (sign-only degradation from $p = 0.002$ to $p < 0.0001$; head-freezing trajectory improvement strengthened in effect size). Some $n = 30$ final-loss significances resolved to null at $n = 300$ (head-freezing final loss), clarifying that the trajectory metric captures the real phenomenon. No robust $n = 30$ finding reversed at $n = 300$. The three-scale protocol thus achieved its design objective: coarse signal at $n = 3$, resolved signal at $n = 30$, fine structure at $n = 300$.

### 5.7 Limitations

**Scale.** The model has 4 layers, 16 dimensions, and ~13,400 parameters. Whether these findings extend to production-scale transformers is unknown. The free lunches (stress inoculation, recovery, regeneration) may be specific to small models with simple loss landscapes, or they may be architectural universals.

**Task complexity.** Character-level name generation is a toy task. Whether temporal pattern access appears in language modeling or other complex tasks is not established.

**Training duration.** 200 steps per phase captures early learning dynamics. The gradual-exposure tolerance (Exp 9) might not persist at longer training horizons.

**DG metric.** The DG Index does not respond to perturbation at $n = 30$ or $n = 300$, resolving an ambiguous $n = 3$ signal.

**Effect sizes.** Many statistically significant effects are practically negligible (<0.5%). Statistical significance at $n = 30$ does not imply practical importance.

**Transplant design.** The null result for transplant advantage (Exp 11) may reflect that both models learned the same task on the same data. Cross-task transplantation (donor trained on a different task) might show transplant effects.

**Competing objectives design.** Gradient negation (Exp 12) is a maximally adversarial perturbation. Subtler forms of inter-layer conflict might reveal more nuanced compensation mechanisms.

**Interpretive framework.** The Platonic Space interpretation (Levin, 2026) is a framework for organizing findings, not a falsifiable metaphysical claim. The free lunches we identify are real and measurable; whether they are best understood as "patterns from a latent space" or as consequences of loss landscape geometry that we have not yet fully characterized is an open question. The framework is useful to the extent that it generates productive research questions, particularly the

question of what a system receives that it did not pay for.

**$n = 3$ to $n = 30$ to $n = 300$ signal evolution.** Several $n = 3$ signals changed character at $n = 30$, underscoring the danger of low-power pilot data. The $n = 30$ to $n = 300$ transition showed a different pattern: no robust finding reversed, but new fine-structure signals emerged (vision radius monotonic structure, regeneration layer specificity, stress inoculation final loss from $p = 0.032$ to $p = 0.0001$), and several effects tightened substantially. The most instructive cross-scale change was Exp 1 head-freezing final loss, which showed several significant improvements at $n = 30$ (freeze 4: $p = 0.001$) but all dissolved at $n = 300$ (all $p > 0.15$), confirming that the trajectory metric, not the final-loss metric, captures the real phenomenon.

### 5.8 Future Work

- **Scale:** Replicate at 100M+ parameter scale to test whether the free lunches (stress inoculation, recovery, regeneration) persist or are specific to minimal architectures.
- **Stress inoculation mechanisms:** Investigate *why* gradual noise builds tolerance, is it adaptive implicit regularization, structural changes in the network, or a property of the Adam optimizer's momentum?
- **Cross-task transplantation:** Test transplant advantage when donor and recipient are trained on different tasks.
- **Graded adversarial conflict:** Scale gradient negation from 0% to 100% to find the corruption threshold and whether free lunches appear at intermediate levels.
- **Architecture morphogenesis:** Allow the architecture to change during training, growing heads, pruning inactive ones, to test whether free lunches extend to structural self-modification.
- **Composite perturbation:** Simultaneously apply multiple perturbation types for more faithful courage/caution testing and to test whether free lunches interact.
- **Biological comparison:** Apply the same 12-experiment protocol to biological developmental systems using the same statistical framework, enabling direct cross-substrate comparison of where free lunches appear.
- **Map the pattern space:** Vary task and architecture while holding the faultization probes constant, to map which patterns are architecture-specific and which are universal. This is a direct contribution to Levin's (2026) research program of exploring the latent space.
- **Beyond $n = 300$:** The three-scale protocol is now complete. Future replication at $n = 1000$ or with different random seeds would further constrain effect size estimates, but the primary findings are stable across all three scales.

## 6. Conclusion

We applied faultization, systematic morphogenetic perturbation, to a minimal transformer through twelve experiments in two phases. The central question was: when we perturb the interface, what patterns from the latent space does perturbation reveal, and what free lunches does the system receive? At $n = 300$, the picture is clear:

**Free lunches.** Four findings describe capabilities the optimizer did not pay for. Temporal pattern access (Exp 9): gradual noise builds tolerance that sudden noise does not ($p = 0.0001$ at $n = 300$, $d = -0.374$), despite identical gradient rules at every step, the history of interface degradation changes which patterns are accessible. Pattern re-binding (Exp 7): path-independent return to identical final loss after transient damage ($p = 0.886$ at $n = 30$, ratio $1.0000 \pm 0.008$, $30/30$ recovered; at $n = 300$, tiny residual $p = 0.030$, $d = +0.126$). Functional role patterns (Exp 10): destroyed layers rebuild to control-equivalent performance at $n = 30$ ($p > 0.17$), with tiny layer-specific residuals emerging at $n = 300$ ($d = 0.12$–$0.17$). Interface simplification (Exp 1): frozen random-projection heads reduce gradient interference ($p < 0.0001$ for trajectory metric at $n = 300$, $d = -1.228$ to $-1.421$); final-loss improvements at $n = 30$ resolve to null at $n = 300$.

**Pattern invariance.** Chimera convergence (Exp 8), transplant indifference (Exp 11), and cell-view equivalent convergence (Exp 2) demonstrate that the same pattern is accessible through multiple interface architectures. Local loss achieves equivalent final loss to end-to-end backpropagation ($p = 0.237$ at $n = 30$, $p = 0.90$ at $n = 300$). The pattern does not depend on the interface; any valid interface converges to it.

**Pattern fidelity.** The architecture maintains clean pattern access up to a degradation threshold. Gradient noise, partial communication reduction, and vision restriction are tolerated below the threshold.

**Pattern corruption.** Silence (pattern unavailability) is tolerated; sabotage (pattern inversion) is catastrophic. The gradient channel is the critical interface signal ($-3.8\%$, $p = 0.022$ at $n = 30$; $-2.9\%$, $p < 0.0001$ at $n = 300$, $d = -0.355$).

Faultization reveals what patterns the system accesses from the latent space, and what free lunches it receives. The optimizer specified "minimize loss." The system received stress tolerance, recovery, regeneration, trajectory improvement, and global convergence from local rules, none of which were prescribed. At $n = 3$ the shapes were rough; at $n = 30$ they sharpened; at $n = 300$ the fine structure confirmed and extended the picture, temporal pattern access strengthened from $p = 0.032$ to highly significant ($p = 0.0001$), vision radius revealed a monotonic structure invisible at lower power, regeneration showed layer-specific signatures, head-freezing final-loss significances resolved to null while trajectory improvement strengthened, and every robust finding held. The three-scale protocol achieved its purpose: what is real survives the turn of the resolution dial.

## References

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7, 49-72.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems*, 19.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 41–48.

Bernstein, J., Wang, Y.-X., Adams, R. P., & Kolter, J. Z. (2018). signSGD: Compressed optimisation for non-convex problems. *Proceedings of the 35th International Conference on Machine Learning*.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., … & Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Frankle, J., & Carlin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *International Conference on Learning Representations*.

Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*.

Kofman, D., Campitelli, G., & Levin, M. (2025). Chess as a model of collective intelligence: Analyzing a distributed form of chess with piece-wise agency. *Organisms: Journal of Biological Sciences*, 8(1-2), 39-62.

Levin, M. (2026). A short argument on Platonic Space: variable-agency patterns that in-form physics, biology, computer science, and cognitive science. Blog post, March 31, 2026.

Levin, M., Bongard, J., & Bhatt, R. (2024). Morphogenetic competencies of sorting algorithms: Delayed gratification, chimeras, and cell-level agency in non-biological systems. *arXiv preprint*.

Meichenbaum, D. (1985). *Stress Inoculation Training*. New York: Pergamon Press.

Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*.

Michel, P., Levy, O., & Neubig, G. (2019). Are sixteen heads really better than one? *Advances in Neural Information Processing Systems*, 32.

Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., & Martens, J. (2015). Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*.

Nokland, A., & Eidnes, L. H. (2019). Training neural networks with local error signals. *Proceedings of the 36th International Conference on Machine Learning*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929-1958.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.