THE EXACT EXCHANGE RATE OF GRADIENT-FREE CREDIT

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient-free training is back at LLM scale, and its cost is now the central question. Folklore holds that on a layer of d parameters, scalar-broadcast evolution strategies need $\sim d$ times more samples than backpropagation — directionally right, but the constant has never been measured. We derive it exactly: for antithetic rank-1 perturbations of an $m \times n$ layer, the estimate $\hat{g}_N$ from N forward evaluations against the true gradient G obeys $\cos^2 = N/(N + N_0)$ with $N_0 = 2(m{+}2)(n{+}2)$, parameter-free, exact in the linear regime at frozen weights; dynamic training adds landscape-dependent excess variance atop this floor. We verify it on every layer of the reference MLP (ratios 0.93–1.12) and of a frozen extreme-aspect-ratio chain (0.976–1.006), killing all four alternatives in that pre-registered experiment (maximum deviation 0.006 across a $340\times$ span of N_0). The law then prices, by extrapolation, the largest gradient-free training system to date: it operates at per-step $\cos \approx 0.02$ and finances that signal with $\sim 10^5$ coherently integrated steps at $\sim 180\times$ the GPU-hours of backpropagation. Channel by channel, our measurements yield a design-space map (two primitives, four laws, six measured rows) and three impossibility statements, among them: no scalar-broadcast channel escapes the exchange rate, and approximate backpropagation is the wrong objective. This is a measurement study, not a new method.

1 INTRODUCTION

After years as a curiosity of the deep-learning toolbox (Hansen, 2006; Wierstra et al., 2014; Salimans et al., 2017), gradient-free training has returned at frontier scale: forward-only fine-tuning of language models is established (Malladi et al., 2023), and evolution strategies now fine-tune billion-parameter models beyond reinforcement learning (Qiu et al., 2026) and pretrain 14B-parameter models in pure integer arithmetic (Sarkar et al., 2026). The appeal is structural: forward-only evaluation needs no backward graph, parallelizes trivially, and fits hardware where backpropagation cannot run. The cost is equally structural: every update is financed by a handful of broadcast fitness scalars. This paper asks what that financing costs, exactly: *what is the exchange rate between forward evaluations and gradient information in a scalar-broadcast channel?*

Classical theory answers in rates. Gaussian-smoothing analyses bound the variance of zeroth-order gradient estimates at $O(d/N)$ up to problem constants (Nesterov & Spokoiny, 2017; Malladi et al., 2023; Baydin et al., 2022) (directionally the folklore "ES (evolution strategy) needs $\sim d$ times more samples"), but the constant has never been measured. Without it, population sizes, batch thresholds, and compute budgets are set by analogy rather than arithmetic.

Our first contribution is the constant. For the antithetic rank-1 estimator at the core of the scalar-broadcast family, the per-step cosine alignment with the true gradient obeys $\cos^2 = N/(N + N_0)$ with $N_0 = 2(m{+}2)(n{+}2)$, parameter-free, exact in the linear regime (Theorem 1). We verify the floor on two frozen architectures (ratios 0.93–1.12 on every layer) and then subject its *form* to a pre-registered decisive experiment: an extreme aspect-ratio chain on which the exact constant separates from all four natural alternatives ($2(m{+}n)$, $2\max$, $2\min$, $2mn$), confirmed to max deviation 0.006 over a $340\times$ span of N_0 (§3.4).

Our second contribution is the pricing. Read through the exact law, the largest published scalar-broadcast system (Sarkar et al., 2026) operates at per-step $\cos \approx 0.02$ (our inference, flagged as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

such) and finances that signal with $\sim 10^5$ coherently integrated steps at roughly $180\times$ the GPU-hours of backpropagation. A correspondence table checks its reported features against the law (§4): the rate is paid, not evaded.

Our third contribution is the design-space map derived from our measurements. Two primitives (information conservation and channel capacity) generate four laws for every credit channel; the laws collapse the design space into a six-row taxonomy, every populated row measured; and the taxonomy supports three impossibility statements, including *no scalar-broadcast channel escapes the exchange rate* and *approximate backpropagation is the wrong objective* (§5, §6).

What we do not claim. (i) We propose no new method; this is a measurement study. (ii) Floors are measured per layer up to $d = 3.15 \times 10^6$ on consumer hardware; the LLM-scale operating point is an inference from the exact law, flagged as such. (iii) Cosine to BP is a coordinate system, not a success criterion (§6). (iv) Four pre-registered verdicts failed and are reported verbatim, unpatched (§7).

§2 fixes the instrument; §3 states and verifies the law; §4–§7 spend it and draw the consequences. The probe and the exact constant are reusable costing tools for future channel designs.

2 A Ruler for Credit Channels

Before any law can be falsified, the instrument must be fixed, validated, and forbidden from drifting: the ruler of this section is deliberately boring — every interesting number in this paper passes through it.

Measurement object and the autograd firewall. Every learning rule is implemented as a closed-form update under `torch.no_grad()`; the single autograd entry point in the codebase is `bp_reference_grad(model, x, y)`, evaluating the reference backpropagation gradient on the *same* mini-batch and *same* weights at which the rule's estimate is taken; rules never see a computation graph or `.grad`. The measurement object (convention C1): the rule's *signed* per-layer gradient estimate $\hat{g}_l$, compared with the reference by cosine.

Protocol and validation. Probing follows the A-02b protocol (step 0 and every 100 steps; current batch and weights; per-layer cosines; norm $< 10^{-8}$ layers marked NaN-idle, never reported as zero alignment), with A-06 aggregation (anchor-constructed layers excluded from headline aggregation; we report only the decoder-mediated layers, L1/L2, "dec" below); full specifications: Appendix G. Each probe records one noise realization; every band-level verdict's margin far exceeds the induced sampling noise; law-grade measurements (§3) use repeated-draw designs (probe batch 256–1024; 9 samples $=$ 3 seeds $\times$ 3 stage points per cell, stage points trajectory-correlated within a seed (effective n between 3 and 9); the decisive cells of §3.4 use independent noise redraws, a registered substitution; SD, ddof $= 1$). The instrument is validated three ways: a BP "rule" against the BP reference returns $\cos \approx 1.000$ on all layers; the true-CE identity head of the anchored channel returns 1.000 everywhere; and the null is *computable*: shuffled-label estimators leave ≤ 0.017 on hidden layers (MLP-4; §3.3). Resolving power: the feedback-alignment depth signature, reproduced in §6 (P-7.6a).

What the ruler is not. The dec bands (> 0.5 "BP-equivalent", 0.2–0.5 "partial", < 0.2 "orthogonal") are a *coordinate system, not a success criterion*: §6 shows alignment is neither necessary nor predictive of function. Cosine-to-BP is measured to state laws about credit, not similarity to an algorithm.

The probe accepts any `nn.Sequential`-like model plus a `rule_g_hat` callable and applies the full protocol in one call (`nbp.probe`, released standalone).

3 The Exact Variance Law

3.1 Statement and derivation

Consider one layer $W \in \mathbb{R}^{m \times n}$ of a network, with exact backpropagation gradient G on a fixed mini-batch, and the antithetic rank-1 estimator of the scalar-broadcast family: perturb $E = ab^\top$

with $a \in \mathbb{R}^m$, $b \in \mathbb{R}^n$ whitened isotropic (i.i.d. unit-variance entries), score each pair by $s_\sigma = [f(W+\sigma E) - f(W-\sigma E)]/(2\sigma)$, and average $\hat{g}_N = \frac{2}{N}\sum_{i=1}^{N/2} s_i E_i$ over N forward evaluations.

Theorem 1 (white-noise floor)**.** *In the linear regime, where the score collapses to the directional derivative $s_\sigma \to \langle E, G\rangle_F$:*

$$\mathbb{E}[\hat{g}_1] = G, \qquad \mathbb{E}\|\hat{g}_1\|^2 = (m+2)(n+2)\,\|G\|_F^2, \tag{1}$$

and consequently, by self-averaging of $\langle \hat{g}_N, G\rangle$ and $\|\hat{g}_N\|^2$ at fixed G,

$$\cos^2(\hat{g}_N, G) \xrightarrow{p} \frac{N}{N + N_0}, \qquad N_0 = 2(m+2)(n+2) \approx 2d, \tag{2}$$

with $d = mn$ and N counting forward evaluations (two per pair).

Proof. Unbiasedness is one line: $\mathbb{E}[sE] = \mathbb{E}[ab^\top a^\top G b] = \mathbb{E}[aa^\top]\, G\, \mathbb{E}[bb^\top] = G$. For the second moment, write $s = a^\top G b = c^\top b$ with $c = G^\top a$; the fourth-moment identity $\mathbb{E}_b[(c^\top b)^2 \|b\|^2] = (n+2)\|c\|^2$ for isotropic b gives $\mathbb{E}\|\hat{g}_1\|^2 = (n+2)\,\mathbb{E}_a[\|G^\top a\|^2 \|a\|^2]$, and the same identity on the a side with $M = GG^\top$ yields $(m+2)(n+2)\|G\|_F^2$. Averaging $N/2$ independent pairs divides the excess moment $[(m+2)(n+2)-1]\|G\|_F^2$ by $N/2$, hence $\cos^2 = N/(N + 2(m+2)(n+2) - 2)$ in the self-averaging limit. The $O(1)$ gap to the stated N_0 is $\leq 0.05\%$ at every dimension used here, and we write $N_0 = 2(m+2)(n+2)$ throughout. □

Three properties matter for what follows. (i) *The constant is parameter-free*: no learning rate, architecture constant, or fitted scale — a measurement, not a fit. (ii) *The constant has structure*: $(m+2)(n+2)$ is neither mn nor $m+n$; §3.4 kills all four natural alternatives. (iii) *The regime is explicit*: the floor holds exactly in the linear (small-σ) regime; finite-σ landscape effects add excess on top (§3.5). (iv) *The family boundary is explicit*: the law prices estimators that broadcast one scalar per perturbation pair; node-space perturbation (§4.3) is a different family, off this floor.

3.2 Relation to prior bounds

Classical zeroth-order analyses bound the *rate*: Gaussian smoothing gives gradient-estimator variance $O(d/N)$ up to problem constants (Nesterov & Spokoiny, 2017), and the same scaling underlies zeroth-order fine-tuning (Malladi et al., 2023), forward-mode random-direction estimators (Baydin et al., 2022), and structured or guided ES variants (Choromanski et al., 2018; Maheswaranathan et al., 2019). The classical full-rank estimator (Wierstra et al., 2014; Salimans et al., 2017) shares our constant at square aspect and parts from it at extreme aspect (Appendix A). Concurrent work on ES at LLM scale characterizes the σ-regimes and proves rank-r bias is $O(r^{-1})$ (Sarkar et al., 2026), consistent with our rank flatness (Table 3, row 4). Our claim is orthogonal: prior work bounds the rate; **we measure the constant**, whose structure (the product form $(m+2)(n+2)$, its linear-regime exactness, its separation from all alternatives (§3.4)) is new, and §4's consequences follow from it.

3.3 Verification: the floor is exact in the linear regime

The law was measured on two frozen architectures (MLP-4, widths $[3072, 1024, 1024, 1024, 10]$, CIFAR-10; CNN-4), the sole autograd entry computing references on the same batch (the §2 caliber; probe batch 256).

Table 1: Per-layer white-noise floors, predicted vs. measured alignment at N=4096, r=1, σ=0.05 (MLP-4): the output layer is exact, hidden layers carry the excess of §3.5. At σ=0.01 every MLP-4 layer hits the floor (0.93–1.12); CNN-4's output layer is exact already at σ=0.05.

LAYER	(m, n)	$N_0 = 2(m+2)(n+2)$	cos PRED.	cos MEAS.
MLP L1	$(1024, 3072)$	6.31×10^6	0.025	0.015 ± 0.005
MLP L2	$(1024, 1024)$	2.11×10^6	0.044	0.034 ± 0.006
MLP L3	$(1024, 1024)$	2.11×10^6	0.044	0.039 ± 0.004
MLP L4 (output)	$(10, 1024)$	2.46×10^4	0.378	0.382 ± 0.010

Back-computing $N_0^{\text{eff}} = N(1/\cos^2 - 1)$ per cell (Table 1): at $\sigma = 0.01$ the ratio is **0.93–1.12 on every layer of MLP-4** (the CNN-4 grid runs at $\sigma = 0.05$). The floor is exact where the theorem

says it should be. At $\sigma = 0.05$ the output layer of both architectures remains exact (MLP L4: 0.918–0.976; CNN fc4: 0.886–1.041) while hidden layers carry an excess treated in §3.5. Three robustness checks. (i) $\sqrt{N}$ *constancy*: N_0^{eff} is constant across $N \in \{64, 512, 4096\}$. The functional form, not merely the constant, is confirmed. (ii) *Rank flatness*: $|\cos(r{=}32) - \cos(r{=}1)| \leq 0.004$ at fixed N (pre-registered), consistent with rank buying conditioning, not bandwidth (Sarkar et al., 2026); the single exception, CNN conv1 ($d{=}864$, $|\Delta \cos| {=} 0.094/0.086$), is a small-$d$ whitening effect, not extra information. Flatness is verified on near-square layers; rank-r interpolation at extreme aspect is open (Appendix A). (iii) *A computable null*: shuffled-label scores leave $|\cos_{\text{null}}| \leq 0.017$ on MLP-4 hidden layers and an explained output-layer residual (0.086 ± 0.099; Table 11), factorizing as $\cos_{\text{ref}} \times \cos_{\text{true}}$ (ratios 1.05/0.91/0.84): column-space dilution between true- and shuffled-label gradients.

The full data collapse ($\cos^2$ against $N/(N{+}N_0)$, all cells of both architectures, EGGROLL regime marked) is Fig. 4 (Appendix B).

3.4 The aspect-ratio experiment: killing four alternatives

On square layers the exact law is nearly indistinguishable from folklore: five hypotheses coincide within percent at $m \approx n$: **H0** $2(m{+}2)(n{+}2)$ (Theorem 1); H1 $2(m{+}n)$ (perturbation-space dimension); H2 $2\max(m, n)$; H3 $2\min(m, n)$; H4 $2mn$ (parameter count). They separate under extreme aspect ratio: H0 parts from H4 by $\approx 2(1/m{+}1/n)$, demanding small $\min(m, n)$; H1–H3 sit orders of magnitude below H0 at large $\max / \min$. We therefore probed a frozen, randomly initialized pure tanh chain $784 \to 2048 \to 8 \to 2048 \to 256 \to 4096 \to 256 \to 10$ (zero training; MNIST; probe batch 256; 3 seeds $\times$ 3 independent noise redraws replacing the stage-point caliber, a registered substitution), under pre-registered gates and predictions.

Table 2: Aspect-ratio decision ($\sigma{=}0.01$): predicted vs. measured cos per layer, back-computed floors against H0 (W1 record-only, W5 mirror of W6, both omitted; full grid in Appendix C). Maximum prediction deviation across all 18 cells: **0.006**.

LAYER	N_0(H0)	cos PRED @4096	cos MEAS @4096	N_0^{eff}/N_0
W2 (8, 2048)	41,000	0.301	0.3047 ± 0.0056	0.976
W3 (2048, 8)	41,000	0.301	0.3031 ± 0.0044	0.988
W4 (256, 2048)	1,057,800	0.062	0.0621 ± 0.0012	1.002
W6 (256, 4096)	2,114,568	0.044	0.0440 ± 0.0006	1.000
W7 (10, 256)	6,192	0.631	0.6299 ± 0.0111	1.006

At the decision layers W2/W3 ($\min(m, n) = 8$; Table 2), measured $\cos(4096) = 0.3047/0.3031$ against H0's 0.301 — H4 (0.333) is excluded at $\approx$ 4–5σ, H1 (0.706) at $\approx$ 70σ (σ = per-sample SD, the conservative choice). Back-computed floors match H0 across the full span (Fig. 1; $N_0^{\text{eff}}/N_0(\text{H0}) = 0.976$–$1.006$ over a $340\times$ range of N_0, 6.2×10^3–2.1×10^6, aspect ratios 1:256–256:1), with mirror layers agreeing to 0.0016/0.0000 and N_0^{eff} constant between $N = 512$ and 4096. This exclusion is what licenses §4: with the $2mn$ family expelled, the scalar-broadcast row's constant is exact, not folklore.

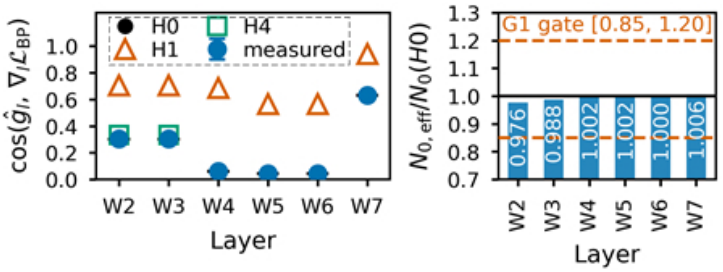

Figure 1: The aspect-ratio verdict: measured alignment ($\sigma{=}0.01$, $N{=}4096$) against H0/H1/H4 (left; H0 = Thm 1; **H4 excluded at** $\approx$ 4–5σ, W2/W3; max $|\text{dev}| = 0.006$ over 18 cells); floor ratios vs the pre-registered G1 gate (right; error bars: sample SD, ddof=1). Sealed figure (App. H).

Registered verdict, verbatim. The pre-registered composite verdict is **FALSE**: gates G1–G3 passed but the null control failed ($\max|\cos_{\text{null}}| = 0.1287$ vs. the ≤ 0.02 gate). The failure is registered as a gate-calibration error: the 0.02 gate was transferred without derivation from the trained-

MLP-4 null to a random-init extreme chain, whose null is systematic (mechanism in Appendix C); it does not touch G1, a different quantity (estimator–reference SNR, not cross-correlation). The insulation is quantitative (Appendix C): the null correlates two references, dilution-factored out of the estimator reading; G3's constancy (0.965–1.042) bounds additive contamination at $\lesssim 4\%$ — $\lesssim 0.012$ in cos at W2/W3, under half the H4 exclusion margin ($0.333-0.3047 = 0.028$), and any residual pushes the measurement *toward* H4, so the exclusion is conservative. The registered meta-rule: *gates only on quantities calibrated in the same regime; cross-regime quantities are report-only.*

3.5 The excess regime: finite σ, trained state

Outside the linear regime the floor does not move — the excess appears on top of it, measurable only against the exact floor. N_0^{eff}/N_0 at $\sigma = 0.05$, trained state: MLP hidden L1 **1.9–3.1** and CNN fc3 **1.97–2.90**, growing with σ; at $\sigma = 0.5$, MLP L1 reaches **7.66** (an effective floor of $\approx 15d$). By contrast, the aspect-ratio chain *at random initialization* stays within **0.84–1.28** of the floor at the same $\sigma = 0.05$ (full grid in Appendix C). The excess is a property of the trained loss landscape, not of the estimator family (cf. the curvature concentration of LLM fine-tuning landscapes, Liang et al., 2026).

Two analytical observations bound what a theory of the excess can say (Appendix A). First, *the leading correction is $O(\sigma^2)$, not $O(\sigma^4)$*: the antithetic score expands as $s_\sigma = \langle E, G\rangle + \frac{\sigma^2}{6} T[E,E,E] + \cdots$ (T the third-derivative tensor; even orders cancel antithetically), and the correction's coefficient contracts G with weight-space traces of T, a property of the trained state, not a universal constant, so **no parameter-free theorem for the excess exists**. Second, $\sigma = 0.5$ *lies outside the perturbative domain*: pure σ^2 growth would extrapolate the $\sigma = 0.05$ excess (0.6–2.1) to 60–210; the measured value is 6.66, sub-quadratic saturation; the excess is reported as measured and mechanism-bounded, not fitted.

4 Pricing the Largest Scalar-Broadcast System

The exact law of §3 is a price list, and this section spends it on the largest published member of the scalar-broadcast family: EGGROLL (Sarkar et al., 2026), which trains 14B-parameter language models with integer-only evolution strategies. The purpose is verification, not critique: every number we cite is theirs, every number we derive is flagged as our inference; the two agree wherever comparable.

4.1 The operating point, priced

EGGROLL perturbs the model layer-wise with rank-r matrices and broadcasts one fitness scalar per population member (per antithetic pair in the integer variant), reaching $N = 2^{20}$ in its largest runs, three orders beyond prior ES, on one GPU (Sarkar et al., 2026). Read through Theorem 1 with the large-dimension floor $N_0 \approx 2d$ at $d \sim 10^9$–10^{10} parameters, the per-step alignment of its update with the true gradient is $\cos(\hat{g}_N, G) = \sqrt{N/(N+2d)} \approx 0.007$–$0.023$. This figure is *our inference*, not their measurement: the whole-model figure mixes all layers in the shared fitness scalar, and the order of magnitude is fixed by the law alone (rank-r and integer-noise scope: Appendix A). It is also a lower bound: the law prices the full parameter dimension, and a smaller effective gradient dimension raises the true alignment (Appendix A). A per-step $\cos \approx 0.02$ is far below a gradient step's: per-step parity ($\cos^2 = 0.5$) costs $N = N_0 \approx 2d$ — the exchange rate. That training nevertheless succeeds is the coherence law operating: $\sim 10^5$ successive estimates, integrated by Adam (Kingma & Ba, 2015) (Sarkar et al., 2026, App. K.2), compound it into learning. Their invoice is explicit: the largest int8 pretraining run requires $\approx 180\times$ the GPU-hours of the backpropagation baseline at fixed data (Sarkar et al., 2026). The floor itself is visible in their ablations: at population 2, training is "incredibly unstable" (Sarkar et al., 2026, Fig. 7); at $N = 2$ the law leaves $\cos \sim 10^{-3}$: nothing to finance. The same reading covers zeroth-order fine-tuning at small N (Malladi et al., 2023): low per-step alignment, financed by coherence.

4.2 CORRESPONDENCE TABLE

Table 3 checks the law's reading of the channel against every relevant reported quantity; each row a consistency statement, none fitted.

Table 3: The exact law's reading of EGGROLL (Sarkar et al., 2026) against its reported values ("inference" = ours; "reported" = theirs).

#	QUANTITY	THE LAW'S READING	EGGROLL (REPORTED)	CONSISTENCY
1	Population N	per-step $\cos \approx 0.02$ at $d \sim 10^9$ (Thm 1, inference)	populations up to 2^{20}	trains via coherence, not per-step fidelity ✓
2	Training cost	financing a 0.02 signal at a fixed exchange rate	$\approx 180\times$ BP GPU-hours	the interest on coherence ✓
3	Mini-batch threshold	the fitness scalars are the *only* information transported; their SNR is set by the data batch	training degrades when the batch falls below $\approx N/4^6$ (their App. I)	a second, independent constraint on the channel ✓
4	Rank-r perturbations	variance is set by N; rank buys conditioning, not bandwidth (our P2: $\lvert\Delta \cos\rvert \leq 0.004$, $r \leq 32$, §3.3)	bias $O(1/r)$, variance dominated by N (their Thm. 4)	orthogonal parameters of one estimator; mutually confirming ✓
5	Coherence integrator	without temporal integration a 0.02 signal cannot finance learning	Adam integration is integral (their App. K.2); the integer variant forgoes momentum, "remains critical future work" (their App. H)	coherence law, first face; self-flagged by them ✓
6	Wall-clock	the channel's cost is forward-only; cheap per step, costly per unit signal	throughput reaches 91% of pure batch inference (pre-generated noise) / 69% (on-the-fly regeneration) (their App. E, Figs. 2a/7)	consistent with forward-only financing ✓
7	Quality and scale frontier	the law is d-explicit: extrapolation is arithmetic, not hope	int8 recurrent model (6L-256D EGG, minipile) 3.40 bits/byte vs. 3.58 fp32 Transformer at the same data batch (cross-architecture); 14B fine-tuned where gradient baselines are memory-infeasible (AIME24 13%→30%, AIME25 7%→33%); no GRPO baseline at that scale (Shao et al., 2024)	reported with our caveat; consistent scope ✓

Three readings. (i) The system's engineering choices are exactly what the laws permit: maximal population (row 1), a coherence integrator (row 5), rank for conditioning rather than bandwidth (row 4), a batch above the degradation floor (row 3). (ii) Nothing in the system contradicts the exchange rate — including its cost report, which is the exchange rate, invoiced. (iii) The law prices alignment to the *mini-batch* gradient; the full-data gradient is one dilution factor away (§3.3), floored by row 3's batch.

4.3 LEAVING THE FLOOR: CHANGE ROWS, NOT FAMILIES

If no scalar-broadcast channel can beat $N \approx 2d$, can a differently shaped channel? The functional form $\cos^2 = N/(N + N_0)$ transfers across estimator families; the constant does not (registered revision T1′). Refitting N_0 cell-by-cell on the frozen node-space (FRL) archive (sweeping per-node queries K, epochs $\{1, 4, 5\} \times K \in \{1, 4, 16, 64\}$, probe batch 1024, zero new training) against the weight-space anchor $N_0^{\text{th}} = 2(n{+}2)$ gives Table 12 (Appendix B).

Three conclusions are registered. (i) N_0 is set by the estimator's *noise geometry* (perturbation space, batch structure, reconstruction, training state) not by weight count alone: node-space perturbation with per-sample noise and batch multiplexing sits one to three orders below the weight-space anchor. (ii) N_0 drifts with training state (L1: $68.6 \rightarrow 261.4$ across epochs): a state variable, not an architecture constant. (iii) The batch-averaged output layer is nearly noiseless ($N_0 \approx 0.03$–0.26 against an anchor of 24). The sample-efficiency gap, quantified: FRL reaches the 0.5 band at $K{=}64$ per-node queries; a weight-space scalar broadcast there needs $N \sim 10^6$ evaluations.

The two literatures are one design space from opposite ends of the bandwidth axis (Fig. 2): EGGROLL pays face value, financed by coherence; node-space and vector-injection channels refinance on cheaper rows (§5).

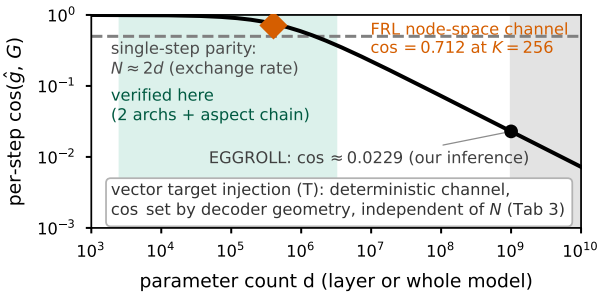

Figure 2: The operating map: per-step alignment at fixed $N = 2^{20}$ against parameter count (curve = Thm 1's law), showing the band verified here (two architectures + the aspect chain; ratios 0.93–1.12, 0.976–1.006), the EGGROLL operating regime (our inference), and the off-curve node-space point, where weight-space parity would need $N \approx 8.3 \times 10^5$ (3247× the node-space queries). Sealed figure (App. H).

5 The Design Space of Credit Channels

The measurements of §2–§4 were made channel by channel; read together they are a map: two primitives generate four laws, the laws induce a taxonomy, and the taxonomy supports three impossibility statements.

5.1 Two primitives, four laws

Information conservation — an update cannot carry more task information than its source injects, yielding the *source law*: every functional channel must contain a term through which fresh task information (y) enters from outside the model's own state. Its corollary is the death of self-referential modulation: a target m generated from the model's own state s injects $I(m; y \mid s) = 0$, and no gating, bounding, or attraction-shaping of such a term manufactures credit. Three mechanism-level autopsies close this slot experimentally: the unanchored repulsive rule collapses (proto-separation $0.004/0.002$, L_1/L_2; accuracy 27.11%); geometric self-modulation matches its shuffled control (40.21 ± 1.25 vs. $40.08 \pm 0.81\%$); a gated, bounded, attraction-only variant fares no better (40.22 ± 0.90 vs. $40.23 \pm 0.39\%$; App. D). A repulsive anchor partially rescues the first variant's *dynamics* (63.19%; App. D) — the anchor law at work, not a counterexample: no self-generated target carries y.

Channel capacity — a channel transporting c scalars per query carries at most $O(c)$ directional information per query; this yields the *bandwidth law*, promoted in §3 from a scaling statement to the scalar-broadcast family's exact constant, $N_0 = 2(m+2)(n+2)$.

Two dynamical requirements complete the set: any iterative channel needs a self-limiting external potential against drift and collapse (the *anchor law*; face: the measured collapse boundary $\gamma_c^{\text{emp}} \in (0.05, 0.1)$, single-seed sweep, registered as such; App. E); and any low-SNR channel can still train if its estimates are integrated coherently across steps (the *coherence law*; faces: Adam's $+5.49$pp on CIFAR-10 and §4's $180\times$ invoice). The four are not homogeneous in status: the bandwidth law is Theorem 1, the source law an information argument, the anchor and coherence laws measured regularities. Necessity is the claim: no exhibited channel works while violating any of the four laws.

5.2 The taxonomy the laws induce

The taxonomy of measured credit channels is presented in Table 4, contrasting theoretical limits with experimental outcomes.

In Table 4, every populated row is measured here (row lineages: (Wierstra et al., 2014; Salimans et al., 2017); (Fiete & Seung, 2006; Lansdell et al., 2020; Pozzi et al., 2020); (Lillicrap et al., 2016; Nøkland, 2016); (Hinton, 2022)).

Table 4: The design space of measured credit channels; N_0 entries are the noise floors of §3–§4; verdicts summarize §5.1–§6.

CHANNEL FAMILY	SOURCE TERM	SCALARS PER QUERY	NOISE FLOOR N_0	MEASURED EXEMPLAR	VERDICT
scalar global broadcast (ES)	loss difference $s = (F^+ - F^-)/2\sigma$	1	$2(m+2)(n+2) \approx 2d$ (exact, §3)	§3 arms; EGGROLL at scale (§4)	works only via coherence; $180\times$ GPU-hours
node-space perturbation + batch multiplexing	per-node loss differences	$O(BKn)$ per pass	$\ll 2(n+2)$; output layer 0.03–0.26 (§4.3)	FRL	0.5 band at $K{=}64$
vector target injection (T)	error-anchored targets	$O(n)$ per layer per sample	n/a (deterministic)	94.73% MNIST; 62.81% C10; T-adam 68.30%	functional; bandwidth-bound at depth (L1 cos ≈ 0.33)
compressed error broadcast (FA/DGL)	top error via fixed random feedback	$O(n_{\text{out}})$	n/a	FA 61.40%, DGL 66.86% (C10)	trains at dec ≈ 0: alignment unnecessary
per-layer local anchor (FF)	label-layer goodness	local	n/a	house FF at 10.02% (chance, C10)	falsified as implemented: alignment does not compose
self-referential modulation	none: $I(m; y \mid s) = 0$	—	—	v5 / v7 / v8 (App. D)	dead (source law)

5.3 THREE IMPOSSIBILITY STATEMENTS

Stated as sharply as our evidence allows — *impossible* meaning excluded by a stated law or by every measured instance:

1. **No self-referential source can work.** By the source law's premise (credit must carry fresh task information), self-generated targets inject none; three autopsies concur.
2. **No scalar-broadcast channel escapes the exchange rate.** $N_0 = 2(m+2)(n+2)$ ($\approx 2d$ at moderate aspect) is not an engineering constant awaiting a trick; it is the family's white-noise floor, now exact. Improvements are confined to shrinking the *effective* perturbation dimension (node-space, low-rank whitening) or multiplexing scalars per pass (batch structure): moving between Table 4 rows, not off a row's floor.
3. **"Approximate backpropagation" is the wrong objective.** Alignment with the BP direction is neither necessary (FA/DGL train at dec ≈ 0) nor predictive (FF and DGL share the dec ≈ 0 profile, Fig. 3: 10.02% vs. 66.86%; FF's layers track local objectives exactly, yet nothing composes). Channel design should optimize information per query under the four laws, not cosine similarity to BP.

5.4 WHAT THE LAWS LEAVE OPEN

Random-feature readouts (Rahma et al., 2026) sit outside the taxonomy by construction: only a convex last layer trains; the price is a kernel-capacity ceiling. The taxonomy leaves exactly one direction open, and the laws fix its shape: take the *direction backbone* from the strongest anchored design (error-anchored vector target injection); spend queries only where §2's probe flags under-delivery: node-space antithetic estimation at $N \propto n$ per flagged layer, not $2d$; integrate with Adam. All source terms are external; the design does not touch the closed slot. We name the synthesis precisely to *decline* claiming it: its predictions pre-registrable (Theorem 1 with FRL-family N_0; measured γ_c; Fig. 3's rails), its validation a new campaign. We claim the map, the laws, the impossibility statements; the coherence law's compounding kernel (§4) and the floor's rank/kurtosis generalization (App. A) stay open.

6 GUARD RAILS: ALIGNMENT IS NOT THE OBJECTIVE

The ruler of §2 measures cosine to BP; here pre-registered predictions show the coordinate is not the objective (Bartunov et al., 2018; Lillicrap et al., 2020). Six channels were trained on a single CNN-4 architecture (CIFAR-10/100). These include BP, anchored vector target injection (T), its Adam-integrated variant (T-adam), feedback alignment (FA), a decoupled-global-loss channel under

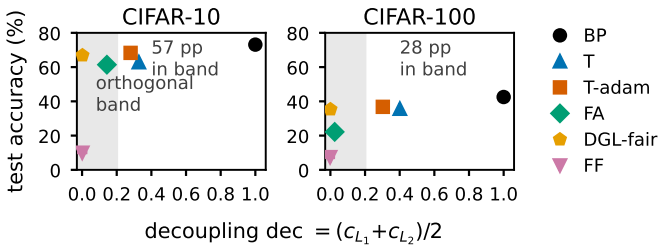

Figure 3: Alignment is not the objective: dec alignment to BP vs. test accuracy, six channels, CNN-4 (CIFAR-10/100, left/right); the shaded band denotes an orthogonal band (dec < 0.2) and spans ≈ 57 accuracy points (CIFAR-10), 28 (CIFAR-100). Sealed figure (App. H).

the same local rule (DGL), and an approach employing per-layer local-anchor learning (FF (Hinton, 2022)), all evaluated under the §2 framework.

Three pre-registered predictions, three confirmations. **P-7.6a** (resolution): FA reproduces the random-feedback depth signature of Lillicrap et al. (2016) (per-layer cosines $0.082 \to 0.205 \to 0.743 \to 1.000$, CIFAR-10). **P-7.6b** (alignment not necessary): DGL trains to 66.86% on CIFAR-10 at dec $= 0.002$, deep in the orthogonal band, 6.3pp under BP's 73.12%. **P-7.6c** (alignment not predictive): FF sits at chance (10.02%) with dec $= 0.001$, a profile indistinguishable from DGL's. Per-layer alignment alone does not compose.

The scatter of Fig. 3 (P-7.6d) spans dec $\in [0, 1]$ against acc $\in [7.44\%, 73.12\%]$ with no monotone relation, and CIFAR-100 replicates it (FA 22.16%, DGL 35.35%, dec ≈ 0).

Credit is thus information transport under the four laws of §5; the BP direction is the one coordinate in which Theorem 1 is exact.

7 Discussion and Limitations

Scale. All floors are measured per layer up to $d = 3.15 \times 10^6$ on one consumer GPU. The LLM-scale operating point (§4) is inferred from the law and the cited system's reports; what transfers is the form and the constant, verified where measurable.

Failures, verbatim. Four pre-registered verdicts failed, unpatched. **R-7.55a** demanded $\cos \geq 0.5$ at $N = 4096$ where the exact law needs $N \approx 7\times10^5$–2×10^6: instrument miscalibration, not estimator bias. **T1** anchored a node-perturbation channel to the weight-space constant $2(n+2)$, a mis-transfer; T1′ is a §4.3 finding. **T2** registered a non-decreasing ratio through the repulsive transition; it turns back at $\gamma = 0.15$ on an inflated denominator while absolute divergence grows (T2′; §5). **T-Phase8a** transferred a null gate across regimes; the meta-rule (*gates only on quantities calibrated in the same regime*) is itself an export.

Artifacts. Frozen repository; one autograd entry point; 74 harness checks; every number recomputed from raw archives (ddof $= 1$); Appendices G–H.

AI use statement

The authors used a large language model assistant for drafting and editing of this paper's text, for data analysis and experimental-workflow assistance under author direction, and for bookkeeping of the experimental record. **AI did not originate mathematical claims, proofs, preregistered experimental designs or synthetic data. All core scientific ideas were conceived solely by the authors.** References were manually curated.

All AI-generated code executes within our frozen preregistered test harness (74 checks). Every reported number is recomputed from raw archived measurements by an independent digest pipeline, and figures include automated consistency assertions. The authors reviewed all AI-assisted material and accept full responsibility for all content, claims and artifacts in this submission.

Ethics statement

This work is a measurement study of training algorithms on public benchmark datasets (MNIST, CIFAR-10/100). It involves no human subjects, no data release, and no deployment; we see no ethics concerns requiring elaboration.

Reproducibility statement

All experiments ran on a single consumer GPU under a frozen repository with a single autograd entry point (Appendix H). Every phase carries a pre-registered specification archived with the code, and every reported number is recomputed from per-phase raw archives by an independent digest pipeline (sample SD, ddof = 1). The measurement instrument of §2 is released as a standalone callable. Three-tier reproduction scripts (smoke / fast / full) and the spec-to-commit-to-test mapping are given in Appendices G and H; code is submitted as anonymized supplementary material.

References

Armen Aghajanyan, Sonal Gupta, Luke Zettlemoyer, and Anchit Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Sergey Bartunov, Adam Santoro, Blake Richards, Luke Marris, Geoffrey E. Hinton, and Timothy Lillicrap. Assessing the scalability of biologically motivated deep learning algorithms and architectures. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

Atılım Güneş Baydin, Barak A. Pearlmutter, Donnacha Syme, Frank Wood, and Philip Torr. Gradients without backpropagation. *arXiv preprint arXiv:2202.08587*, 2022.

Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard E. Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Ila R. Fiete and H. Sebastian Seung. Gradient learning in spiking neural networks by dynamic perturbation of conductances. *Physical Review Letters*, 97(4):048104, 2006.

Nikolaus Hansen. The CMA evolution strategy: A comparing review. In *Towards a New Evolutionary Computation: Advances in the Estimation of Distribution Algorithms*, pp. 75–102. Springer, 2006.

Geoffrey Hinton. The Forward-Forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Benjamin J. Lansdell, Pradyumna R. Prakash, and Konrad P. Kording. Learning to solve the credit assignment problem. In *International Conference on Learning Representations (ICLR)*, 2020.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations (ICLR)*, 2018.

Qiyao Liang, Jinyeop Song, Yizhou Liu, Jeff Gore, Ila Fiete, Risto Miikkulainen, and Xin Qiu. The blessing of dimensionality in LLM fine-tuning: a variance-curvature perspective. *arXiv preprint arXiv:2602.00170*, 2026.

Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7: 13276, 2016.

Timothy P. Lillicrap, Adam Santoro, Luke Marris, Colin J. Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, 2020.

Niru Maheswaranathan, Luke Metz, George Tucker, Dami Choi, and Jascha Sohl-Dickstein. Guided evolutionary strategies: Augmenting random search with surrogate gradients. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Timoleon Moraitis, Dmitry Toichkin, Adrien Journé, Yansong Chua, and Qinghai Guo. SoftHebb: Bayesian inference in unsupervised Hebbian soft winner-take-all networks. *Neuromorphic Computing and Engineering*, 2(4):044017, 2022.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

Isabella Pozzi, Sander Bohté, and Pieter Roelfsema. Attention-gated brain propagation: How the brain can implement reward-based error backpropagation. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Xin Qiu, Yulu Gan, Conor F. Hayes, Qiyao Liang, Yinggan Xu, Roberto Dailey, Elliot Meyerson, Babak Hodjat, and Risto Miikkulainen. Evolution strategies at scale: LLM fine-tuning beyond reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Atamert Rahma, Chinmay Datar, Ana Cukarska, and Felix Dietrich. Rapid training of Hamiltonian graph networks using random features. In *International Conference on Learning Representations (ICLR)*, 2026.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Bidipta Sarkar, Mattie Fellows, Juan Agustin Duque, Alistair Letcher, Antonio León Villares, Anya Sims, Clarisse Wibault, Dmitry Samsonov, Dylan Cope, Jarek Liesen, Kang Li, Lukas Seier, Theo Wolf, Uljad Berdica, Valentin Mohl, Alexander David Goldie, Aaron Courville, Karin Sevegnani, Shimon Whiteson, and Jakob Nicolaus Foerster. Evolution strategies at the hyperscale. *arXiv preprint arXiv:2511.16652*, 2026. v2.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *Journal of Machine Learning Research*, 15(1):949–980, 2014.

A Full derivation of Theorem 1

A.1 Setup and conventions

One layer $W \in \mathbb{R}^{m\times n}$; a fixed mini-batch on which the reference $G = \nabla_W \mathcal{L}_{\mathrm{BP}}$ is evaluated by the sole autograd entry point (§2). Antithetic rank-1 pairs $\pm E$, $E = ab^\top$ with $a \in \mathbb{R}^m$, $b \in \mathbb{R}^n$ independent, zero-mean, unit-variance, whitened ($\mathbb{E}[aa^\top] = I_m$, $\mathbb{E}[bb^\top] = I_n$); Gaussian in all experiments. The antithetic score and the estimator are

$$s_\sigma \;=\; \frac{f(W+\sigma E) - f(W-\sigma E)}{2\sigma}, \qquad \hat{g}_N \;=\; \frac{2}{N}\sum_{i=1}^{N/2} s_i E_i, \tag{3}$$

where N counts *forward evaluations* throughout this paper (two per pair; $\hat{N} = N/2$ pairs). In the linear regime $s_\sigma \to \langle E, G\rangle_F = a^\top G b$.

A.2 THE FOURTH-MOMENT IDENTITY, WITH KURTOSIS

Lemma 2 (contracted fourth moment)**.** *Let b have independent, zero-mean, unit-variance components with vanishing odd moments and kurtosis $\kappa = \mathbb{E}[b_i^4]$ (hence $\mathbb{E}[b_i b_j] = \delta_{ij}$). Then for every fixed c,*

$$\mathbb{E}_b\big[(c^\top b)^2 \|b\|^2\big] \;=\; (n - 1 + \kappa)\,\|c\|^2. \tag{4}$$

For Gaussian b ($\kappa = 3$): $(n+2)\,\|c\|^2$.

Proof. $\mathbb{E}[(c^\top b)^2\|b\|^2] = \sum_{i,j,k} c_i c_j\,\mathbb{E}[b_i b_j b_k^2]$. For $i \neq j$ every term contains an odd moment of some component and vanishes. For $i = j$: $\sum_k \mathbb{E}[b_i^2 b_k^2] = \mathbb{E}[b_i^4] + \sum_{k\neq i}\mathbb{E}[b_i^2]\mathbb{E}[b_k^2] = \kappa + n - 1$. □

The companion second-moment identity is $\mathbb{E}[(c^\top b)^2] = \|c\|^2$ (unit directional variance: one antithetic pair carries one unit-variance measurement of G along E). With general whitened laws the results below read $N_0^{\text{exact}} = 2(m-1+\kappa_a)(n-1+\kappa_b) - 2$; the product structure and the $\approx 2d$ scaling do not depend on Gaussianity, only the $+2$ per side does. All experiments use $\kappa_a = \kappa_b = 3$.

A.3 PROOF OF THEOREM 1, UNABRIDGED

Unbiasedness. $\mathbb{E}[sE] = \mathbb{E}[ab^\top a^\top G b] = \mathbb{E}[aa^\top]\,G\,\mathbb{E}[bb^\top] = G$, using independence of a and b and whitening.

Second moment of one pair. Write $s = a^\top G b = c^\top b$ with $c = G^\top a$. Then

$$\mathbb{E}\,\|sE\|_F^2 = \mathbb{E}\big[(c^\top b)^2\|a\|^2\|b\|^2\big] = (n+2)\,\mathbb{E}_a\big[\|G^\top a\|^2\|a\|^2\big] = (m+2)(n+2)\,\|G\|_F^2, \tag{5}$$

applying Lemma 2 on the b side (at fixed a) and then on the a side with c replaced along $\|G^\top a\|^2$.

Centering. Since $\mathbb{E}[sE] = G$ and $\langle G, \mathbb{E}[sE]\rangle = \|G\|_F^2$,

$$\mathbb{E}\,\|sE - G\|_F^2 = \big[(m+2)(n+2) - 1\big]\,\|G\|_F^2. \tag{6}$$

Averaging. Over $\hat{N} = N/2$ independent pairs, $\mathbb{E}\,\|\hat{g}_N\|_F^2 = \|G\|_F^2\left(1 + \frac{(m+2)(n+2)-1}{\hat{N}}\right)$.

Self-averaging. At fixed G, $\langle \hat{g}_N, G\rangle$ concentrates on $\|G\|_F^2$ at rate $\hat{N}^{-1/2}$: for Gaussian perturbations $\mathbb{E}[\langle E, G\rangle^4] = 3\|G\|_F^4 + 6\|G^\top G\|_F^2 \leq 9\|G\|_F^4$ (equality at rank-1 G), so $\mathrm{Var}\langle sE, G\rangle \leq 8\|G\|_F^4$; the same order holds for $\|\hat{g}_N\|_F^2$. Therefore

$$\cos^2(\hat{g}_N, G) \;\xrightarrow{p}\; \frac{\hat{N}}{\hat{N} + (m+2)(n+2) - 1} \;=\; \frac{N}{N + 2(m+2)(n+2) - 2}. \tag{7}$$

Integer-exact versus leading form. The exact floor is $N_0^{\text{exact}} = 2(m+2)(n+2) - 2$; the registered form $N_0 = 2(m+2)(n+2)$ drops the $O(1)$ term, a relative gap of $1/((m+2)(n+2)) \leq 3.2 \times 10^{-4}$ at the smallest layer used anywhere in this paper ($(m, n) = (10, 256)$) — inside the 0.05% stated in §3.1 and far inside every measurement's sampling noise. All gates and prediction tables use the leading form.

A.4 EXCHANGE-RATE ARITHMETIC

Two readings of the same law, kept distinct throughout:

$$\cos^2 = \tfrac{1}{2} \iff N = N_0 \approx 2d; \qquad \cos = \tfrac{1}{2} \iff \cos^2 = \tfrac{1}{4} \iff N = \frac{N_0}{3} \approx \frac{2d}{3}. \tag{8}$$

The paper's exchange-rate sentence ("single-step analytic-gradient parity needs $N \approx 2d$ forward calls") is the $\cos^2 = 0.5$ reading, with N in forward calls and $d = mn$ the layer parameter count.

A.5 RANK, KURTOSIS, AND THE FULL-RANK SIBLING

Full-rank sibling. Classical ES perturbs full-rank isotropically (Hansen, 2006; Wierstra et al., 2014; Salimans et al., 2017): an isotropic E in $\mathbb{R}^{mn}$ obeys the same fourth-moment identity with mn in place of n, giving $N_0^{\text{full}} = 2(mn{+}2) - 2 \approx 2d$. The two constants agree to $2(m{+}n)/mn$ at square aspect (0.4% at $m{=}n{=}1024$) but separate by 25% at 1:256; thus, the aspect experiment's exclusion of H4 ($2mn$: the full-rank leading form) certifies the rank-1 geometry specifically, while the $\approx 2d$ rate itself is geometry-robust at moderate aspect.

Rank-r interpolation. Theorem 1 is derived at $r{=}1$. The measured rank flatness of §3.3 ($r \le 32$) sits on near-square layers, where the rank-1 and full-rank limits differ by $\sim 0.4\%$; flatness there is expected, not diagnostic. Rank-r floors at extreme aspect (interpolating between $2(m{+}2)(n{+}2)$ and $2(mn{+}2)$) are an open, pre-registrable prediction, not a measurement of this paper.

Kurtosis. By Lemma 2, $N_0^{\text{exact}} = 2(m{-}1{+}\kappa_a)(n{-}1{+}\kappa_b) - 2$: discrete or quantized noise ($\kappa < 3$; Rademacher $\kappa{=}1$) shrinks the constant toward $2mn$. The system priced in §4 trains in integer arithmetic; its layers are near-square, so the kurtosis shift is $O(2/m{+}2/n) \approx 0.4\%$ — far inside the 0.007–0.023 inference band, which the d-range dominates.

Ambient versus effective dimension. The floor counts the ambient d, while the success of fine-tuning at $\cos \approx 0.02$ (§4) is consistent with the task gradient occupying a low-dimensional subspace (Li et al., 2018; Aghajanyan et al., 2021). An N_0 indexed by the gradient's effective dimension rather than the layer's parameter count is the natural next constant; it is a statement about the landscape, not the estimator family, and we do not measure it here.

Information per forward evaluation. The law bounds the channel's per-query yield: $\cos^2/N = 1/(N{+}N_0) \le 1/N_0 \approx 1/(2d)$, against backpropagation's $\cos^2 = 1$ at ≈ 3 forward-equivalents per step. Per unit compute the scalar-broadcast family is therefore down by $\Theta(d)$ per step at every N; the $180\times$ GPU-hour invoice of §4 is this bound made concrete, and what rescues the family is wall-clock parallelism and arithmetic simplicity, not information efficiency.

A.6 THE FINITE-σ EXPANSION (SUPPORTING §3.5)

Taylor-expand the antithetic score about $\sigma = 0$. Even derivatives of f cancel in the difference $f(W{+}\sigma E) - f(W{-}\sigma E)$, leaving odd orders only:

$$s_\sigma \;=\; \langle E, G\rangle \;+\; \frac{\sigma^2}{6}\, T[E,E,E] \;+\; \frac{\sigma^4}{120}\, V[E,E,E,E,E] \;+\; \cdots, \tag{9}$$

where T and V are the loss's third- and fifth-derivative tensors in weight space at W, acting on E by contraction.

Bias direction. Taking expectations,

$$\mathbb{E}[s_\sigma E] \;=\; G \;+\; \frac{\sigma^2}{6}\, t^* \;+\; O(\sigma^4), \qquad t^* := \mathbb{E}[T[E,E,E]\, E]\,. \tag{10}$$

For the rank-1 law $E = ab^\top$, Isserlis factorizes the fourth moment across the independent sides, $\mathbb{E}[a_p a_r a_t a_v] = \delta_{pr}\delta_{tv} + \delta_{pt}\delta_{rv} + \delta_{pv}\delta_{rt}$ (and likewise for b), so $\mathbb{E}[E_{pq}E_{rs}E_{tu}E_{vw}]$ is a sum of nine index-pair contractions. Hence t^* is a fixed linear combination of *partial traces of T over weight-space index pairs*, determined by the trained landscape at W, not by any estimator constant.

Variance correction. With s_σ as above,

$$\mathbb{E}\,\|s_\sigma E\|_F^2 = (m{+}2)(n{+}2)\,\|G\|_F^2 \;+\; \frac{\sigma^2}{3}\, C_2 \;+\; O(\sigma^4), \qquad C_2 := \mathbb{E}\big[\langle E, G\rangle\, T[E,E,E]\, \|E\|_F^2\big]\,, \tag{11}$$

a degree-6 moment (linear $\times$ cubic $\times$ quadratic in E), generically nonzero, and again a G–T contraction. Consequently

$$N_0^{\text{eff}}(\sigma) \;=\; N_0 \;+\; \tfrac{2}{3}\,\sigma^2\, \frac{C_2}{\|G\|_F^2} \;+\; O(\sigma^4). \tag{12}$$

This is the content of the two observations in §3.5. *First*, the leading correction to the floor is $O(\sigma^2)$, not $O(\sigma^4)$, and its coefficient is a contraction of trained-state tensors (T with itself and with

G). A parameter-free theorem for the excess therefore does not exist: the constant $2(m+2)(n+2)$ is a property of the estimator family, while C_2 is a property of the landscape the estimator happens to stand on. *Second*, the perturbative form self-limits: pure σ^2 growth extrapolates the measured $\sigma = 0.05$ hidden-layer excess (0.6–2.1) to 60–210 at $\sigma = 0.5$, against the measured 6.66 — sub-quadratic saturation, so $\sigma = 0.5$ lies outside the perturbative domain and the excess is reported as measured, mechanism-bounded, and unfitted. The random-initialization control of §3.4 sits on the same picture: on the untrained chain the full $\sigma = 0.05$ grid stays within 0.84–1.28 of the floor (Appendix C), i.e. $C_2 \approx 0$ before training and $O(1)$ after; the excess accrues with the trained state, as N_0's drift with training state does independently in §4.3.

B Full measurement tables for the white-noise floor

B.1 Phase 7.55: all measured cells

The ES-rank-r dual-pair estimator content probe $\cos(\hat{g}_l, \nabla_l \mathcal{L}_{\mathrm{BP}})$ (frozen weights; unique autograd entry `bp_reference_grad`; probe batch 256; step points $\{0, T/2, T\}$; seeds $\{42, 43, 44\}$; 9 samples per cell; SD with ddof=1). Arm A = MLP-4/CIFAR-10, 11 cells; Arm B = CNN-4/CIFAR-10, 4 cells. Registered exact law $N_0 = 2(m+2)(n+2)$ ($E = ab^\top$ white-noise base); layer shapes $(m, n) = (\mathrm{out}, \mathrm{in})$. Figure 4 plots the collapse of these cells on $\cos^2(N) = N/(N+N_0)$.

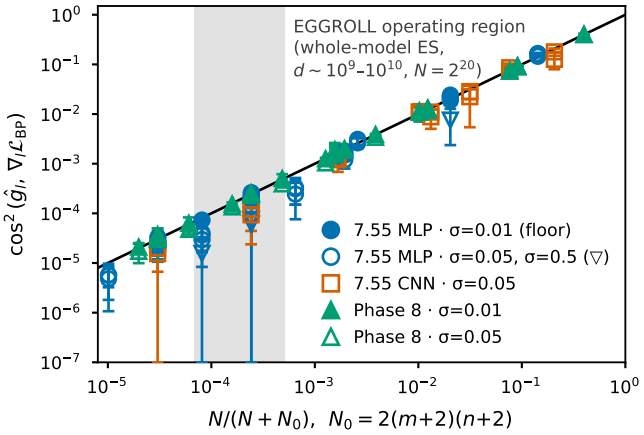

Figure 4: Data collapse onto the exact law: $\cos^2$ against $N/(N+N_0)$ for all cells of both architectures; the gray band marks the EGGROLL operating regime (our inference, §4); error bars: sample SD (ddof=1). Sealed figure (App. H).

Verdict (registered, verbatim). R-7.55a = False (pre-registered gates; calibration error of the gates, not an estimator defect). **R-7.55a′ (upgraded):** the ES-rank-r low-rank dual estimator is, at our scale, a theoretically exact unbiased noisy-gradient estimator – $\cos^2(N) = N/(N+N_0)$, $N_0 = 2(m+2)(n+2)$ ($E = ab^\top$ white-noise base), holding exactly in the $\sigma \to$linear regime on both architectures' output layers and on all MLP layers; hidden layers develop higher-order contamination excess variance for $\sigma \geq 0.05$. Exchange rate: single-step analytic-gradient parity needs $N \approx 2d$ forward calls (d = layer parameter count; $\cos^2 = 0.5$ at $N = N_0$). Measured cells: Tables 5–11.

Table 5: CNN-4 Arm B: same quantity, all 4 pre-registered cells. conv1/conv2/fc3 hidden, fc4 output positive control.

(r, N, σ)	conv1	conv2	fc3	fc4
$(1, 64, 0.05)$	0.141±0.053	0.034±0.008	0.004±0.002	0.104±0.011
$(1, 512, 0.05)$	0.350±0.065	0.092±0.019	0.009±0.003	0.271±0.010
$(8, 64, 0.05)$	0.167±0.023	0.035±0.004	0.004±0.001	0.104±0.014
$(8, 512, 0.05)$	0.421±0.035	0.104±0.008	0.010±0.003	0.292±0.014

Table 6: MLP-4 Arm A: $\cos(\hat{g}_l, \nabla_l \mathcal{L}_{\mathrm{BP}})$ mean $\pm$ SD (ddof=1, 9 samples), all 11 pre-registered cells. L1–L3 hidden, L4 output positive control.

(r, N, σ)	L1	L2	L3	L4
$(1, 64, 0.05)$	0.002 ± 0.001	0.004 ± 0.001	0.005 ± 0.001	0.053 ± 0.005
$(1, 512, 0.05)$	0.005 ± 0.002	0.012 ± 0.002	0.014 ± 0.002	0.145 ± 0.008
$(1, 4096, 0.05)$	0.015 ± 0.005	0.034 ± 0.006	0.039 ± 0.004	0.382 ± 0.010
$(8, 64, 0.05)$	0.002 ± 0.001	0.005 ± 0.001	0.005 ± 0.001	0.051 ± 0.006
$(8, 512, 0.05)$	0.006 ± 0.002	0.014 ± 0.001	0.014 ± 0.001	0.150 ± 0.007
$(8, 4096, 0.05)$	0.017 ± 0.004	0.039 ± 0.004	0.042 ± 0.002	0.403 ± 0.008
$(32, 64, 0.05)$	0.002 ± 0.001	0.005 ± 0.001	0.006 ± 0.001	0.055 ± 0.006
$(32, 512, 0.05)$	0.006 ± 0.002	0.014 ± 0.001	0.015 ± 0.001	0.155 ± 0.004
$(32, 4096, 0.05)$	0.018 ± 0.005	0.039 ± 0.004	0.042 ± 0.002	0.407 ± 0.004
$(1, 512, 0.01)$	0.009 ± 0.001	0.016 ± 0.001	0.016 ± 0.001	0.137 ± 0.011
$(1, 512, 0.5)$	0.003 ± 0.002	0.006 ± 0.005	0.008 ± 0.005	0.082 ± 0.030

Table 7: $N_{0,\mathrm{eff}} = N(1/\cos^2 - 1)$ divided by the exact-law $N_0 = 2(m+2)(n+2)$, per measured cell (Arm A MLP). Linear floor at σ=0.01 (all-layer range 0.93–1.12, digest-locked); output layer exact at σ=0.05 (L4, r=1 subset, 0.918–0.976); hidden excess grows with σ/depth; σ=0.5 L1 $\approx 15d$ (ratio 7.66).

(r, N, σ)	L1	L2	L3	L4
$(1, 64, 0.05)$	2.41	1.63	1.37	0.92
$(1, 512, 0.05)$	3.07	1.79	1.27	0.97
$(1, 4096, 0.05)$	2.96	1.67	1.30	0.98
$(8, 64, 0.05)$	1.89	1.40	1.13	1.00
$(8, 512, 0.05)$	2.36	1.26	1.17	0.91
$(8, 4096, 0.05)$	2.20	1.30	1.10	0.86
$(32, 64, 0.05)$	2.05	1.18	0.95	0.85
$(32, 512, 0.05)$	2.17	1.25	1.12	0.84
$(32, 4096, 0.05)$	2.08	1.26	1.08	0.84
$(1, 512, 0.01)$	1.12	0.93	1.01	1.09
$(1, 512, 0.5)$	7.66	5.79	4.25	3.05

Table 8: Hidden-layer excess (range of $N_{0,\mathrm{eff}}/N_0$ over the σ=0.05 cells, all values recomputed from the raw archives) and the σ^2-extrapolation check: scaling the σ=0.05 L1 excess (0.6–2.1) by $(0.5/0.05)^2$=100 predicts 60–210, while the measured σ=0.5 L1 excess is 6.66 (7.66−1).

layer	ratio range (σ=0.05)	note
MLP L1	1.89–3.07	hidden/conv excess
MLP L2	1.18–1.79	hidden/conv excess
MLP L3	0.95–1.37	hidden/conv excess
CNN conv1	1.13–1.86	hidden/conv excess
CNN fc3	1.97–2.90	hidden/conv excess
MLP L1 σ^2-extrap	60–210	$(0.6\text{–}2.1) \times 100$
MLP L1 measured σ=0.5	6.66	= 7.66−1

Table 9: Pre-registered P2 (rank-flatness, N=512, $|\cos(r{=}1) - \cos(r{=}32)|$) and the Arm B conv1 exception ($|\cos(r{=}8) - \cos(r{=}1)|$ per step point, N=512). The pre-registered P2 scope is the hidden layers at N=512, where ≤ 0.002 (gate 0.05, PASS); L4 (output, out of scope) is 0.010, shown for completeness. conv1 d=864: at high rank the low-dim probe sits closer to $N(0, I)$, r=8 raises cos (0.094/0.086 > 0.08 soft gate).

check	L1	L2	L3	L4
$\lvert\cos(32) - \cos(1)\rvert$ @N=512	0.001	0.002	0.001	0.010
Arm B conv1, N=512, per step point: 0.094/0.086 (soft gate 0.08: P2(soft) False)				

Table 10: Null instrument (label-shuffled reference gradient, $N{=}4096$, $r{=}1$): per-layer $\cos_{\text{null}}$ mean±SD. Hidden layers sit near zero; the L4 residual ($+0.086 \pm 0.099$) factorizes as the product dilution of Table 11. CNN null not measured (null cell (1,4096) not on Arm B grid, all-NaN).

	L1	L2	L3	L4
$\cos_{\text{null}}$ mean±SD	$-0.002{\pm}0.005$	$-0.002{\pm}0.009$	$+0.002{\pm}0.009$	$+0.086{\pm}0.099$

Table 11: Product-dilution model of the null residual, $\cos_{\text{null}} \approx \cos_{\text{ref}} \times \cos_{\text{true}}$ (MLP L4, seed-42 basis): ratio $1.05/0.91/0.84$ at step points $0/5865/11730$ (digest-locked) – the high-dim cross term $\sim 1/\sqrt{d}$ vanishes.

step point	$\cos_{\text{null}}$	$\cos_{\text{ref}}$	$\cos_{\text{true}}$	$\cos_{\text{ref}}\cos_{\text{true}}$	ratio
0	0.090	0.218	0.395	0.086	1.05
5865	0.100	0.296	0.371	0.110	0.91
11730	0.117	0.355	0.391	0.139	0.84

B.2 Node-space floors (FRL): the companion table of §4.3

The detailed noise floor measurements for the node-perturbation channel are summarized in Table 12.

Table 12: Noise floors of the node-perturbation channel (FRL) against the weight-space anchor; per-cell fit R^2 0.970–0.999; held-out $K{=}256$ predicted 0.7168 vs. measured 0.7121.

LAYER	$2(n{+}2)$	N_0^{fit} **(EP1/EP4/EP5)**	**RATIO**
L1 (hidden)	1028	68.6 / 242.2 / 261.4	0.067–0.254
L2 (hidden)	1028	18.4 / 51.8 / 34.6	0.018–0.050
L3 (output)	24	0.06 / 0.26 / 0.03	0.001–0.011

C Aspect-ratio experiment: full grid

C.1 Phase 8: extreme-aspect-ratio decisive probe

Aspect chain 784→2048→8→2048→256→4096→256→10 (pure MLP, tanh, frozen weights, zero training). Layer shapes (out, in): W1 (2048,784) *record-only* (7.55-type repeat), W2 (8,2048) bottleneck decisive, W3 (2048,8) mirror, W4 (256,2048) large-N_0, W5 (4096,256) large-N_0 decisive, W6 (256,4096) mirror, W7 (10,256) output positive control. Registered exact law $N_0 = 2(m{+}2)(n{+}2)$ ($E = ab^\top$ white-noise base). Grid $N \in \{64, 512, 4096\} \times \sigma \in \{0.01, 0.05\} = 6$ cells $\times$ 7 layers; 9 samples per cell (seeds $\{42, 43, 44\} \times 3$ noise re-draws; SD ddof=1); unique autograd entry `bp_reference_grad`. The pre-registered chain spans an N_0 range of $341.5\times$ across the decision layers W2–W7 (W1 record-only excluded; rounded to $340\times$ in the main text). The full grid, per-layer ratios, gate outcomes, and per-cell deviations are reported in Tables 13–16.

Table 13: Phase 8 full grid: $\cos(\hat{g}_l, \nabla_l \mathcal{L}_{\text{BP}})$ mean±SD (ddof=1, 9 samples), all 6 cells $\times$ 7 layers. W1 record-only (shaded); W2/W5 decisive, W3/W6 mirrors, W7 positive control. Rows omitted from the main text are included here.

	W1[†]	W2	W3	W4	W5	W6	W7
$\sigma = 0.01$							
$N{=}64$	0.0045±0.0005	0.0407±0.0071	0.0390±0.0060	0.0078±0.0013	0.0054±0.0008	0.0060±0.0010	0.1060±0.0190
$N{=}512$	0.0124±0.0005	0.1144±0.0074	0.1113±0.0043	0.0222±0.0025	0.0152±0.0005	0.0158±0.0016	0.2704±0.0163
$N{=}4096$	0.0354±0.0009	0.3047±0.0056	0.3031±0.0044	0.0621±0.0012	0.0439±0.0011	0.0440±0.0006	0.6299±0.0111
$\sigma = 0.05$							
$N{=}64$	0.0041±0.0008	0.0419±0.0081	0.0430±0.0082	0.0069±0.0007	0.0056±0.0008	0.0053±0.0017	0.1028±0.0181
$N{=}512$	0.0116±0.0004	0.1084±0.0040	0.1129±0.0077	0.0198±0.0015	0.0158±0.0013	0.0159±0.0012	0.2711±0.0158
$N{=}4096$	0.0322±0.0009	0.2977±0.0064	0.3021±0.0062	0.0583±0.0011	0.0445±0.0006	0.0434±0.0011	0.6340±0.0066

[†] W1 = record-only (7.55-type layer, not a decision layer; main text reports W2–W7 only).

Table 14: $N_{0,\text{eff}} = N(1/\cos^2 - 1)$ over the exact-law $N_0 = 2(m+2)(n+2)$, σ=0.01, N=4096 (G1 cells). All six decision layers land in 0.976–1.006 (digest-locked). W1 record-only ratios at $N \in \{64, 512, 4096\}$ (recomputed; not gated): 0.960, 1.035, 1.014.

layer	shape (m, n)	N_0=$2(m+2)(n+2)$	cos	$N_{0,\text{eff}}$	ratio
W2	$(8, 2048)$	$41,000$	0.3047	$40,025$	0.976
W3	$(2048, 8)$	$41,000$	0.3031	$40,500$	0.988
W4	$(256, 2048)$	$1,057,800$	0.0621	$1,059,446$	1.002
W5	$(4096, 256)$	$2,114,568$	0.0439	$2,118,408$	1.002
W6	$(256, 4096)$	$2,114,568$	0.0440	$2,114,051$	1.000
W7	$(10, 256)$	$6,192$	0.6299	$6,229$	1.006

Table 15: Pre-registered gates (`_8spec.txt`; any FAIL → verdict FALSE, registered as-is). G1 all six decision-layer ratios in [0.85,1.20] and every separable alternative excluded (W2/W3 vs H1=0.100×H0, H4=0.799×H0; W4/W5/W6 vs H1/H2/H3). G2 symmetric pairs; G3 $\sqrt{N}$ constancy W2/W3/W7. Null (hidden W2–W6, label-shuffled, N=4096) FAILs at max $|\cos_{\text{null}}|$ = 0.1287 $\gg$ 0.02.

gate	check	value	verdict
G1	all 6 ratios ∈ [0.85,1.20]	0.976–1.006	PASS
G2	$\lvert\cos(\text{W2}) - \cos(\text{W3})\rvert$	0.0016	PASS
G2	$\lvert\cos(\text{W5}) - \cos(\text{W6})\rvert$	0.0000	PASS
G3	$N_{0,\text{eff}}(512)/N_{0,\text{eff}}(4096)$	0.965/1.008/1.042 (W2/W3/W7)	PASS
Null	max $\lvert\cos_{\text{null}}\rvert$ hidden (W2–W6)	0.1287 $\gg$ 0.02	FAIL

Table 16: Measured vs pre-registered prediction (`_8spec.txt` §four), σ=0.01 per cell. dev = measured − prediction. max |dev| = 0.006 (W7 @ $N = 512$; digest-locked).

layer	pred (64/512/4096)	measured	dev (64/512/4096)
W2	0.039 / 0.111 / 0.301	0.0407 / 0.1144 / 0.3047	+0.002 / +0.003 / +0.004
W3	0.039 / 0.111 / 0.301	0.0390 / 0.1113 / 0.3031	+0.000 / +0.000 / +0.002
W4	0.0078 / 0.022 / 0.062	0.0078 / 0.0222 / 0.0621	+0.000 / +0.000 / +0.000
W5	0.0055 / 0.016 / 0.044	0.0054 / 0.0152 / 0.0439	−0.000 / −0.001 / −0.000
W6	0.0055 / 0.016 / 0.044	0.0060 / 0.0158 / 0.0440	+0.001 / −0.000 / −0.000
W7	0.101 / 0.276 / 0.631	0.1060 / 0.2704 / 0.6299	+0.005 / −0.006 / −0.001

Why the failed null does not touch G1. The null cell reuses the estimator machinery with a shuffled-label *reference*; its 0.1287 measures a correlation between two references (true- and shuffled-label gradients on the same batch and weights) dilution-factored by the estimator's true alignment (Table 17). In the G1 cells the estimator never sees shuffled labels, and at σ=0.01 it is unbiased by construction (the finite-σ bias is $O(\sigma^2)$, Appendix A). Two measured quantities bound any residual estimator-side contamination. (i) An N-independent additive component in $\hat{g}$ would break the $\sqrt{N}$ constancy: G3 holds N_0^{eff} constant between N=512 and 4096 to within 0.965–1.042, bounding such a component to $\lesssim 4\%$ of the signal; $\lesssim 0.012$ in cos at W2/W3, under half the H4 exclusion margin ($0.333 - 0.3047 = 0.028$). (ii) A layer-asymmetric systematic would break the mirror symmetry G2 ($|\Delta \cos| = 0.0016$ / 0.0000). Finally, any residual inflation moves the measurement *toward* H4's higher prediction; the measured $0.3047/0.3031$ sit on H0's 0.301, so the H4 exclusion is conservative against precisely this failure mode.

Verdict (registered; rendered, not patched). T-Phase8a = FALSE (any FAIL → FALSE, registered as-is); **T-Phase8a′ (revised registration) = decisive confirmation**. G1/G2/G3 all PASS → H0=2(m+2)(n+2) decisively confirmed on every separable layer (ratios 0.976–1.006; H1/H2/H3/H4 excluded at ≥1.25×/2560×; decisive cell W2/W3 @ N=4096 measured 0.3047/0.3031 vs H0 prediction 0.301); **Null FAIL** (hidden max $|\cos_{\text{null}}|$=0.1287 $\gg$ gate 0.02) → FALSE. Attribution (gate not fixed): shuffled-label CE gradients non-orthogonal on the same input batch/weights (output $\langle p-y, p-y' \rangle \approx -0.1$ structural correlation propagated through random weights), $|\cos_{\text{null}}| \approx 0.1$ systematic signal not estimator noise; the 0.02 gate is a calibration error relative to the extreme-aspect-ratio chain (7.55 hidden convention ≤0.017 at MLP-4/CIFAR max width 1024; chain has width 4096, bottleneck 8).

Table 17: Null instrument (label-shuffled reference gradient, N=4096; per-layer max $|\cos_{\text{null}}|$ over the 9 samples). W1/W7 not null-tested (record-only / output). Mechanism (registered attribution, gate *not* patched): shuffled-label CE gradients are *not* orthogonal on the same input batch / frozen weights – the output layer $\langle p-y,\, p-y' \rangle \approx -0.1$ structural correlation propagates through the random weights, so $|\cos_{\text{null}}| \approx 0.1$ is a systematic signal, not estimator noise. The 0.02 gate was calibrated per the 7.55 hidden convention ($\leq$0.017, MLP-4/CIFAR max width 1024); the extreme-aspect chain (7 layers, width 4096, bottleneck 8) lies outside that calibration regime. Product-dilution model $\cos_{\text{null}} \approx \cos_{\text{ref}} \times \cos_{\text{true}}$ (MLP L4 ratios 1.05/0.91/0.84, seed-42): see Table 10.

	W2	W3	W4	W5	W6
max $\lvert\cos_{\text{null}}\rvert$	0.1287	0.0410	0.0385	0.0327	0.0317

D AUTOPSIES OF THE SELF-REFERENTIAL SLOT (V5 / V7 / V8)

The source law's corollary (§5.1) says a target generated from the model's own state injects $I(m; y \mid s) = 0$, and no gating, bounding, or attraction-shaping of such a term can manufacture credit. The corollary is a statement about *all* mechanizations; the three autopsies below close the slot experimentally at three independent mechanism levels: unanchored self-generated targets (v5), self-model *geometric* modulation of a working anchored channel (v7), and the same modulation fully mechanized; gated, norm-bounded, attraction-only (v8). All verdicts below are registered, unpatched.

D.1 V5: SELF-MODEL TARGETS COLLAPSE (PHASE 6, MNIST)

Mechanism. The target generator reads the model's own state: prototype-EMA targets t_l are built from the network's own activation prototypes and injected in place of the error-anchored targets of the T channel. No fresh y enters the update beyond what the state already contains.

Registered numbers (seeds 42–46; mean $\pm$ SD, ddof $=$ 1): accuracy $32.69 \pm 7.80\%$ (per seed: $27.11/34.32/22.60/41.83/37.60$) against the shuffled-prototype control $20.57 \pm 3.86\%$ ($20.65/24.35/14.33/23.09/20.42$); unstable across seeds and close to its own shuffled floor, against anchors T 94.73% and BP 97.35%. The proto-separation diagnostic collapses to its floor by epoch 20 ($6.24 \rightarrow 0.03$; floor ratio ≈ 0.004). Verdict: **falsified** (double-red). Commit `a70430e`, `results/phase6_self_model.npz`.

Anchor reproduction. The $\gamma = 0$ point of the Phase-7.7 sweep (Appendix E) reproduces the collapse under the registered rule unchanged (proto-separation L1 $6.46 \rightarrow 0.03$ vs. Phase 6's $6.24 \rightarrow 0.03$; accuracy 27.11% at seed 42, matching the value quoted in §5.1).

Partial rescue, and its limit (Phase 6.5). Adding a repulsive anchor to the same self-generated targets causally rescues function: $63.19 \pm 4.57\%$ ($67.67/68.20/60.22/61.93/57.91$) against its shuffled control $29.70 \pm 5.77\%$, which is a $\approx$ 33.5pp separation, with the probe showing re-anchoring at the top layer (L3 ≈ 0.998) while L1/L2 remain orthogonal ($0.19/0.06$). But 63.2% stays far under the anchored channel's 94.73%: the repulsive structure stabilizes the dynamics (that is the anchor law's job, §5.1) without externalizing the source. Commit `6a2b575`, `results/phase65_repulsive.npz`. Within the v5 lineage, function grows monotonically with repulsion strength (accuracy 68.77% at $\gamma = 0.5$ vs. 63.19% at $\gamma = 0.25$; registered as an auxiliary record) — a statement about the external anchor, not about the closed slot, and unrelated to the v7/v8 death verdicts below.

D.2 V7: GEOMETRIC SELF-MODULATION IS INERT (PHASE 7.2, CIFAR-10 MLP-4)

Mechanism. On top of the working anchored inverse channel (§5.1; T channel at scale), a self-model readout *geometrically modulates* the delivered update by reshaping, not replacing, the credit direction with a state-derived geometric term.

Registered numbers (seeds 42–44): modulated $40.21 \pm 1.25\%$ ($39.11/41.57/39.95$) against the shuffled-modulation control $40.08 \pm 0.81\%$ ($39.21/40.82/40.21$): $\Delta \approx 0$, indistinguishable. The unmodulated anchored channel sits at $45.04 \pm 0.25\%$. Verdict: **INERT** (double-red). Commit `332e530`, `results/phase72_modulation.npz`.

D.3 v8: GATED, BOUNDED, ATTRACTION-ONLY — STILL INERT (PHASE 7.4, CIFAR-10 MLP-4)

Mechanism. The same self-model slot, now fully mechanized in the direction the corollary names: the modulation term is *gated*, *norm-bounded*, and restricted to *attraction-only* shaping.

Registered numbers (seeds 42–44): gated v8 $40.22 \pm 0.90\%$ ($39.37/41.17/40.13$) against the shuffled-gating control $40.23 \pm 0.39\%$ ($39.89/40.65/40.15$); again indistinguishable, and again under the anchored channel ($45.04 \pm 0.25\%$; v7 reused at $40.21 \pm 1.25\%$). Verdict: **INERT** (double-red). Commit `02a7f02`, `results/phase74_v8.npz`.

D.4 WHAT THE THREE AUTOPSIES CLOSE

Table 18 is the experimental content of the source-law corollary. The three manipulations are mechanistically independent, namely replacing the source term, reshaping a working channel geometrically, and mechanizing that reshaping under gating, bounding, and attraction-only constraints, and each is matched by a shuffled control that keeps the machinery while destroying the self-reference. The one manipulation that moves function (repulsive anchoring) acts on the dynamics, not the source: it rescues stability without delivering task information, and its ceiling remains far below the anchored channel it is bolted onto. We know of no remaining mechanization level at which the slot could be re-opened without injecting fresh y — which is precisely the source law.

Table 18: The self-referential slot, autopsied at three mechanism levels. Accuracy mean ± SD (ddof $= 1$); controls are the matched shuffled variants. All values recomputed from the raw archives.

VARIANT	MECHANISM LEVEL	ACC. vs. CONTROL	ANCHOR	VERDICT	COMMIT
v5 (Phase 6)	self-generated targets, unanchored	32.69 ± 7.80 vs. 20.57 ± 3.86	T 94.73 / BP 97.35	falsified	`a70430e`
v5 + repulsive anchor (6.5)	same targets, dynamics stabilized	63.19 ± 4.57 vs. 29.70 ± 5.77	T 94.73	partial rescue, not externalized	`6a2b575`
v7 (Phase 7.2)	geometric self-modulation of T channel	40.21 ± 1.25 vs. 40.08 ± 0.81	T 45.04 ± 0.25	inert	`332e530`
v8 (Phase 7.4)	gated, bounded, attraction-only	40.22 ± 0.90 vs. 40.23 ± 0.39	T 45.04 ± 0.25	inert	`02a7f02`

E THE NOISE–REPULSION PHASE MAP (ABSOLUTE-DIVERGENCE FORM)

E.1 NOISE–REPULSION PHASE DIAGRAM

Two probe dimensions locate the working regimes cited by the main text. **(i) Noise axis** (7.55 ES-rank-r, r=1, N=512, MLP-4/CIFAR-10): pre-registered magnitudes $\sigma \in \{0.01, 0.05, 0.5\}$. σ=0.01 is the linear floor where the registered exact law $\cos^2(N) = N/(N + N_0)$, $N_0 = 2(m + 2)(n + 2)$, holds (all-layer $N_{0,\text{eff}}/N_0$ ratio range 0.93–1.12); σ=0.05 develops hidden-layer excess; σ=0.5 saturates (L1 ratio 7.66, excess 6.66). **(ii) Repulsion axis** (7.7 v5 repulsive-anchor sweep, $\gamma \in \{0, 0.05, 0.1, 0.15, 0.25, 0.5\}$, MNIST, seed=42, 20 ep): collapse for small γ ($R_\gamma < 0.5$), growth for large γ ($R_\gamma > 2$); crossover boundary $\gamma_c \in (0.05, 0.1)$ (main text §5). Tables 19–23 report the two axes in full.

Table 19: Noise axis: $\cos(\hat{g}_l, \nabla_l \mathcal{L}_{\text{BP}})$ mean±SD (ddof=1, 9 samples) at r=1, N=512, all four MLP layers, for the three pre-registered σ. Digest-locked cells.

σ	L1	L2	L3	L4
$\sigma = 0.01$	0.0085±0.0010	0.0161±0.0007	0.0155±0.0008	0.1371±0.0108
$\sigma = 0.05$	0.0051±0.0019	0.0117±0.0017	0.0138±0.0019	0.1448±0.0075
$\sigma = 0.5$	0.0033±0.0024	0.0065±0.0046	0.0076±0.0045	0.0823±0.0296

Table 20: Noise axis, same cells: $N_{0,\text{eff}}/N_0$ ratio (exact law $N_0 = 2(m{+}2)(n{+}2)$). σ=0.01 linear floor, all-layer ratio range 0.93–1.12 (digest 0.93–1.12); σ=0.05 hidden excess (L1 3.07, L2 1.79, L3 1.27) with the output layer exact (L4 0.970); σ=0.5 deep excess, L1 ratio 7.66 $\approx 15d$.

σ	L1	L2	L3	L4
$\sigma = 0.01$	1.123	0.933	1.010	1.085
$\sigma = 0.05$	3.073	1.789	1.269	0.970
$\sigma = 0.5$	7.664	5.788	4.245	3.050

Table 21: Repulsion axis: collapse ratio R_γ =sep[ep20]/sep[ep1] per layer and final test accuracy (single seed 42, per spec; no cross-seed mean, ddof n/a). γ_c crossing between γ=0.05 (both layers $R < 0.5$, collapse) and γ=0.1 (both $R > 2$, growth): $\gamma_c \in (0.05, 0.1)$. The registered monotonicity gate fails at γ=0.15 (see Table 22): R peaks there then falls as high-γ repulsion inflates the ep1 denominator.

γ	R_γ L1	R_γ L2	test acc	regime
$\gamma = 0$	0.004	0.002	0.2711	collapse
$\gamma = 0.05$	0.007	0.004	0.4642	collapse
$\gamma = 0.1$	4.486	62.356	0.7755	growth
$\gamma = 0.15$	10.489	56.473	0.6836	growth
$\gamma = 0.25$	6.611	18.345	0.6767	growth
$\gamma = 0.5$	2.738	2.562	0.6877	growth

Table 22: Absolute proto separation (the quantity T2′ moved the paper phase diagram onto): proto-sep at ep1 and ep20 per layer, digest-locked prose values. ep20 diverges monotonically with γ (L1: 0.03 $\rightarrow$ 314 $\rightarrow$ 350); the R peak at γ=0.15 is a denominator artifact – high-γ repulsion inflates the ep1 initial spread 20–100$\times$ (L1 ep1: 6.5 at γ=0 vs 127.8 at γ=0.5).

γ	ep1 L1	ep20 L1	ep1 L2	ep20 L2	R L1/L2
$\gamma = 0$	6.462	0.028	1.589	0.004	0.00/0.00
$\gamma = 0.05$	9.704	0.065	2.461	0.010	0.01/0.00
$\gamma = 0.1$	14.806	66.414	3.981	248.222	4.49/62.36
$\gamma = 0.15$	22.524	236.254	6.722	379.630	10.49/56.47
$\gamma = 0.25$	47.486	313.954	21.456	393.605	6.61/18.35
$\gamma = 0.5$	127.783	349.907	154.403	395.606	2.74/2.56

Table 23: T1 variance law (FRL K-sweep archive, seed=42, σ=0.05, probe batch 1024): N_0 fit per (epoch, layer) to $\cos^2(K) = K/(K + N_0)$, $K \in \{1, 4, 16, 64\}$, vs theoretical anchor $N_{0,\text{th}} = 2(n{+}2) = \{1028, 1028, 24\}$. All 9 cells fail the registered band $[0.5, 2] \times N_{0,\text{th}}$ (ratio 0.001–0.254) $\rightarrow$ T1 = FALSE. Fit R^2 0.970–0.999; e4-L1 K=256 extrapolation (excluded from fit) $\cos(256) = 0.7168$ vs measured 0.7121 (ratio 0.993). Report items: structure slope $s_{\text{inv}} \approx -1$.

epoch	layer	$N_{0,\text{fit}}$	$N_{0,\text{th}}$	ratio	R^2	PASS
ep1	L1	68.552	1028	0.067	0.991	FAIL
ep1	L2	18.367	1028	0.018	0.997	FAIL
ep1	L3	0.065	24	0.003	0.990	FAIL
ep4	L1	242.187	1028	0.236	0.997	FAIL
ep4	L2	51.759	1028	0.050	0.995	FAIL
ep4	L3	0.257	24	0.011	0.981	FAIL
ep5	L1	261.352	1028	0.254	0.999	FAIL
ep5	L2	34.618	1028	0.034	0.999	FAIL
ep5	L3	0.034	24	0.001	0.970	FAIL

Verdicts (registered). T1 = FALSE (0/9 cells): $N_{0,\text{fit}}$ far below $2(n + 2)$; the exact-form structure $\cos^2(K) = K/(K + N_0)$ holds (R^2 0.970–0.999), but N_0 is set by the estimator's noise geometry, not by $2(n{+}2)$. **T2 = FALSE** (the registered "non-decreasing" gate fails on both layers): the phase transition exists and is two-layer consistent at $\gamma_c \in (0.05, 0.1)$, but $R(\gamma)$ falls back after γ=0.15 – a ratio artifact (high-γ repulsion inflates the ep1 denominator), while absolute divergence keeps growing.

F THE T-CHANNEL LINEAGE

This appendix is the audit trail behind the calibrated bands and anchors used in the main text: how the instrument's floors and ceilings were set, and how the two channel families that populate Table 4 arrived at their registered verdicts. All accuracies are mean $\pm$ SD (ddof $= 1$); commits and archives are listed per phase (Table 24; full spec mapping in Appendix G).

Table 24: Phase lineage, Phase 0 through 7.5 (pre-registered eras). Headlines are the registered per-phase numbers; "—" marks phases whose full treatment is elsewhere (autopsies: Appendix D; law-grade phases 7.55/7.7/8: §3–§4 and Appendices B–E).

PHASE	VARIANT / PURPOSE	HEADLINE (registered)	COMMIT
0	BP / FF (Hinton, 2022) / SoftHebb (Moraitis et al., 2022) baselines, MNIST, seed 42	ruler calibration: BP ≈ 1.000, FF ≈ 0, SoftHebb head ≈ 1.000	`8802246`, `99f482b`
1–1.5	FRL node-perturbation learner; K-sweep verdicts	unbiased noisy-BP estimator; L1 $\cos = 0.7121$ at $K{=}256$ (e4)	`e1efb4f`, `06776c4`
2	GFN v1: natural-metric geometry core	metric condition number 35k–70k (nat) vs. 1.0 (euclid); GDI 0.06–0.16	`c5df051`
2.5	multi-seed replication (seeds 42–46)	nat $+0.52 \pm 0.06$pp over euclid; McNemar $p < 0.01$ all seeds	`861307c`
3	GFN v2: task-metric targets	$+0.23 \pm 0.05$pp, shuffled control ≈ 0 (-0.05 ± 0.12)	`fda4e37`
4–4.5	GFN v3: residual injection; readout forensics	inert ($+0.16 \pm 0.13$pp); forensics: readout reads injected features, not BP credit	`46d55df`, `1716ead`
5	GFN v4: target injection (T channel), MNIST	$91.97 \pm 0.13 < 94$ gate; L1 cos $0.532 > 0.5$; joint verdict **FALSE** (Phase-5 lineage era)	`7511041`
5.5	bounded hardening (noise-trained inverse, one-shot)	94.73 ± 0.17; Δ over bad-decoder control $+1.71 \pm 0.14$pp; functional channel established	`45aa2ce`
6 / 6.5	v5 self-model targets; repulsive-anchor rescue	falsified / partial rescue — Appendix D	`a70430e`, `6a2b575`
7.0	CIFAR-10 MLP-4 scale base	BP ceiling recalibrated to 47.99 ± 0.72 (amendment A-05)	`140fbc0`
7.1	bandwidth law at scale	T 45.04 ± 0.25 vs. BP 47.99; decoder-mediated L1/L2 0.333/0.409 vs. bad-decoder 0.175/0.065; green under A-06	`efa5d05`
7.2 / 7.4	v7 / v8 self-model modulation	both inert — Appendix D	`332e530`, `02a7f02`
7.3a	CNN-4 + CIFAR-100 scale	T 35.65 ± 0.20 vs. BP 42.53 ± 0.36; bandwidth over conversion	`4e31cba`
7.5	coherence rescue (Adam integration)	C10 $+5.49$pp with probe band intact ($0.0476 \leq 0.05$): coherence hypothesis holds; C100 $+1.14$pp in the pre-registration gap ("undecided", amendment A-08)	`7d84511`

The GFN family. The geometry-first line asked whether credit could be carried by redefining the geometry rather than transporting the error: v1 established that a natural metric redefines credit geometry (condition numbers 35k–70k against the Euclidean identity) and that the geometric term is a valid carrier of a small, significant gain ($+0.52$pp, replicated; $+0.23$pp for the task-metric variant, causally tied to task information by the shuffled control); v3 residual injection was inert; v4 target injection failed its first joint gate (channel-established = FALSE, the Phase-5 lineage-era verdict retained in the taxonomy); v5's self-model closure and its autopsy line are Appendix D.

The anchored inverse channel. The T line begins at Phase 5.5's hardening (functional channel at 94.73% on MNIST, separated from its bad-decoder control by $+1.71$pp) and is carried to scale in §4.3: the bandwidth law at CIFAR-10 (7.1) and CNN-4/CIFAR-100 (7.3a); first-layer under-delivery with decoder-mediated layers cleanly separated from the control, and the coherence rescue (7.5), whose C10 success and C100 "undecided" are both registered as such. The protocol amendments that this lineage forced — A-02 (probe protocol, convention C1), A-03/A-04 (FRL K-sweep verdicts), A-05 (ceiling recalibration), A-06 (anchor-layer exclusion), A-08 (the gap ruling) — are fixed in Appendix G; every amendment preserves the original pre-registered caliber alongside the amended one, and no verdict was ever retroactively rewritten.

G FALSIFICATION PIPELINE: SPEC / COMMIT / TEST MAPPING

This appendix fixes the measurement instrument of §2 in full, then maps every verdict used in this paper to its pre-registered specification, execution commit, and test coverage.

G.1 NOTATION AND CONVENTIONS

$\hat{g}_N$, G	per-layer estimate; BP reference (same batch, same weights; C1)
cos, dec	cosine alignment; decoder-mediated aggregation over L1/L2 (§2)
N	forward evaluations (two per antithetic pair)
N_0, N_0^{eff}	white-noise floor $2(m{+}2)(n{+}2)$ (leading form; exact: $2(m{+}2)(n{+}2) - 2$); its measured effective value
σ, r	perturbation scale (per-sample SD); perturbation rank
T, T-adam	anchored vector target injection; its Adam-integrated variant (§6)
FA, DGL, FF	feedback alignment; decoupled-global-loss channel; per-layer local anchor (§6)
FRL	node-space perturbation family (§4.3)
H0–H4	the five floor candidates of §3.4
dec bands	> 0.5 BP-equivalent; 0.2–0.5 partial; < 0.2 orthogonal (§2)
G1–G3, P-7.6a–d, R-7.55a, T1/T2, T-Phase8a	pre-registered gates and verdicts (this appendix and App. H)

G.2 INSTRUMENT SPECIFICATION

Measurement object (convention C1). The measured object is the rule's *signed* per-layer gradient estimate $\hat{g}_l$, compared by cosine similarity against `bp_reference_grad` evaluated on the same mini-batch and the same weights. Per family, $\hat{g}$ is: BP and SoftHebb head: ∇_W of cross-entropy; FF (Hinton, 2022): ∇_W of the local goodness objective; SoftHebb (Moraitis et al., 2022) hidden layers: $\Delta W_{\text{oja}}/\eta$ (Oja, 1982); FRL: $\delta_l^\top a_{l-1}/B$ (batch-averaged node-perturbation estimate); ES-family arms: the antithetic score average of §3.1. The displacement-based caliber was retired under C1; $\cos \approx +1$ means the rule is secretly identical to BP.

Probe protocol (A-02b). At step 0 of every epoch and every 100 steps thereafter: take the *current* training batch and *current* weights, compute the rule's $\hat{g}$ for this step and the reference gradient on the same batch and weights, record per-layer cosines, then apply the rule's normal update. Outputs land in `results/{algo}_probe_cos.npz`; the epoch-cumulative caliber is kept as a secondary record only. *Idle rule*: a layer whose full-epoch $\|\Delta W\|_2 < 10^{-8}$ is recorded NaN-idle with its training-rule source annotated — never as zero alignment (this is how FF's untrained output head appears throughout).

Aggregation (A-06). Constructed-anchor layers are excluded from headline aggregation, because the anchored channel's top-layer target $t = a - \alpha(p - y)$ contains W of the layer above by construction, forcing $\cos \approx 1$ independently of the decoder; anchor layers are reported standalone ("constructed anchor"). Headline aggregation (*dec*) covers the decoder-mediated layers only (L1/L2 on MLP-4 and CNN-4). The exclusion applies symmetrically to both arms of every control pair, and the original pre-registered caliber is preserved alongside the A-06 caliber (dual-caliber policy: no historical red reading is erased, no verdict is retroactively rewritten; Phase 5/5.5's pass criteria were already decoder-mediated, and Phase 7.0's BP self-probe is a true BP recheck, not a constructed anchor).

Firewall. Every rule is implemented closed-form under `torch.no_grad()`; rules never see a computation graph or `.grad`. The single autograd entry point in the codebase is `bp_reference_grad(model, x, y)`. The harness's firewall scan (block 77) mechanically checks that new drivers contain zero autograd calls.

Validation and bands. The instrument is calibrated on known answers: a BP "rule" probed against the BP reference returns $\cos \approx 1.000$ on all layers (calibration deviation 0.0% against the 0.5% tolerance); FF hidden layers sit at the floor (≈ -0.01); SoftHebb Oja hidden layers at ≈ 0.004; the

true-CE identity head of the anchored channel returns exactly 1.000. The dec bands (> 0.5 BP-equivalent, 0.2–0.5 partial, < 0.2 orthogonal) are a coordinate system, not a success criterion (§6). Resolving power is demonstrated by reproducing the feedback-alignment depth signature (P-7.6a, §6).

G.3 Harness and the number pipeline

The frozen test harness grew monotonically with the program: 41 checks entering Phase 7.0; 44 after it; 47 after Phase 7.1; 72 entering Phase 8; and **74** after Phase 8 added the block-76 schema validation and the block-77 firewall scan ($74/74$ green at the frozen state). Separately from the harness, every number in this paper is recomputed from the per-phase raw archives (`results/*.npz`, per-seed arrays) by an independent digest pipeline, using sample standard deviation with ddof $= 1$ throughout (the one known divergence from historical README prints, ddof $= 0$, is registered in the digest: Phase 7.5 C100 Tb-adam $\pm 1.09 \rightarrow \pm 1.34$). The digest is the sole number source for the paper. All figures are bound to archived values by automated assertions and were first sealed at their plotting commits (`8ceb277`; `5443fd9`; `bdc2037`), then regenerated at final print size with data pipelines byte-identical and re-sealed at the submission freeze commit `22f9926` cited in the main-text captions.

G.4 Verdict → spec → commit → test mapping

Every phase was executed under a specification registered *before* execution, committed once, and closed by an adjudication record, as shown in Table 25. The registered verdicts stand verbatim in the repository, including every FALSE: R-7.55a (instrument miscalibration, §3), T1 and T2 (§4.3, §5), and T-Phase8a (gate transfer across regimes, §3.4).

Table 25: Spec / commit / test mapping. "Spec" is the pre-registered specification archived with the code; "exec. commit" is the single execution commit (verification and registration commits listed where separate). All npz archives live under `results/` and are never overwritten (one file per arm).

PHASE	SPEC	EXEC. COMMIT(S)	ARCHIVE / TESTS
0 / 1–1.5 / 2 / 2.5	`_75spec3.txt` (original blocks)	`8802246`, `99f482b`; `e1efb4f`, `06776c4`; `c5df051`; `861307c`	`{algo}_history.npz` (Phase 0); `frl_*.npz`; `gfn_*.npz`; `phase25_seed_sweep.npz`
3 / 4–4.5 / 5 / 5.5 / 6 / 6.5	per-phase registrations (README/digest sections)	`fda4e37`; `46d55df`, `1716ead`; `7511041`; `45aa2ce`; `a70430e`; `6a2b575`	`phase3_task_metric.npz`; `phase4_*`, `phase45_*`; `phase5_*`; `phase55_*`; `phase6_*`; `phase65_*`
7.0 / 7.1 / 7.2	`_75spec.txt` (blocks 1–4)	`140fbc0`; `efa5d05`, `e3ad54b`; `332e530`	`phase70_bp.npz`; `phase71_scale.npz`; `phase72_modulation.npz`; harness 41 → 47
7.3a / 7.4 / 7.5	`_75spec.txt` (appended blocks)	`4e31cba`; `02a7f02`; `7d84511`	`phase73a_scale.npz`; `phase74_v8.npz`; `phase75_adam.npz`
7.55	registered in adjudication record	`9569a42` (verify), `975e63b` (digest)	`phase755_esr_mlp.npz` + `phase755_esr_cnn.npz` (separate files, overwrite forbidden)
7.6	registered in adjudication record	`6b234a1` (ruler), `881d698` (verdicts)	`phase76_ruler.npz`; checkpoint `phase76_ruler_ckpt.json`
7.7	`_77spec.txt`	`e4b2ece`, `3f070a4` (registration)	`phase77_variance.npz` + `phase77_collapse.npz`
8	`_8spec.txt`	`1c82da2`	`phase8_aspect_ratio.npz`; harness 72 → **74**
errata	registered errata (digest §"errata")	`2eec70b`, `291013e` (F4), `ff78e6d` (digest errata A1–A5 + script mirrors)	append-only; no registered number ever edited in place
appendix tables	scripted regeneration (this paper, App. B/C/E)	`224fe7d`, mirrored `ff78e6d`	103 / 100 / 67 assertions bind every printed cell to the archives

H REPRODUCTION GUIDE

All experiments ran on a single consumer GPU; the model is recorded as RTX 2070S (8 GB) for the scale phases 7.0–7.5, 7.55, and 8, while the earlier MNIST phases predate hardware logging. Code is submitted as anonymized supplementary material; every phase's pre-registered specification is archived with the code (Appendix G); the reproduction tiers and measured wall-clock costs are given in Tables 26 and 27.

H.1 THREE TIERS

Table 26: Reproduction tiers. All runs are non-destructive: `NBP_RESULTS_DIR` / `NBP_CONFIG_PATH` redirect outputs to a temporary directory, never touching the audited `results/` archives.

TIER	COVERAGE	WALL-CLOCK (single GPU)
`reproduce.sh --smoke`	mini-epoch end-to-end smoke (57 unit-test calls); skips the reconciliation gate	minutes
`reproduce.sh --fast`	MNIST key arms (bp / ff / softhebb / frl, seed 42), reconciled against the archive (`results/summary.txt` lines 1–4) within ± 0.5pp	core $<$ 10 min; budgeted $<$ 1 h
`reproduce.sh --full`	fast core plus the nine heavy stages below; heavy stages start manually	$\approx$ 8–11 h total

Table 27: Heavy-stage wall-clock: registered reproduction budget (*est.*; README manifest) vs. archived measurements (*measured*; npz/log arrays, summed over runs). "Not recorded" means the archive carries no wall-clock field; nothing is back-filled.

STAGE	EST.	MEASURED
gfn v1 geometry	minutes	not recorded
v5 / v6 self-model	$\sim$15 min	not recorded
7.0 CIFAR-10 MLP-4 scale (BP)	$\sim$1–2 h	33 min
7.1 bandwidth at scale (T/Tb)	(with 7.0)	72 min
7.2 v7 modulation	$\sim$30 min	79 min
7.3a CNN-4 ruler (18 runs)	$\sim$1 h	6.9 h
7.4 v8 gating	$\sim$30 min	82 min
7.5 Adam integration (12 runs)	$\sim$30–90 min	2.75 h
7.6 ruler (FA/DGL-fair/FF, 18 runs)	(with 7.5)	5.04 h
7.55 ES-rank-r grids	$\sim$3–4 h	3.9 h
7.7 phase diagram	$\sim$1.5 h	not recorded
Phase 8 aspect-ratio grid (probe-only)	$\leq$1 h	140.1 s

H.2 REPRODUCER NOTES

1. *No checkpoint-resume on heavy stages*: 7.55 ($\sim$3–4 h) and 7.7 ($\sim$1.5 h) run without resume support; only the 7.6 ruler checkpoints (`results/logs/phase76_ruler_ckpt.json`). Budget uninterrupted windows for the heavy stages.

2. *Archives are append-only*. The Phase-0 baselines (`results/{algo}_history.npz`, `{algo}_probe_cos.npz`) are seed-42 single-run archives; any rerun must go through the `NBP_RESULTS_DIR` redirection (automatic under `reproduce.sh`), since a direct rerun would overwrite the archive.

3. *Canonical probe entry*: `nbp.probe`; `src/nbp/metrics.py` is a thin legacy proxy with no independent logic.

4. *Tracked artifacts*: Phase 7.7's zero-new-training fit reuses the tracked archive `results/frl_k_sweep.npz`; `results/frl_ckpt_epoch*.pt` are optional local caches, not part of the audit chain.

5. *Console encoding* (Windows, GBK code page): run via `reproduce.sh`, or set `chcp 65001` / `PYTHONIOENCODING=utf-8` before invoking drivers directly.

H.3 VERIFYING THE NUMBERS WITHOUT RETRAINING

The audit chain is checkable independently of the heavy stages: the 74-check harness (≈5 min) validates schemas and the autograd firewall; the digest pipeline recomputes every reported mean ± SD from the per-seed archive arrays (ddof $= 1$); the figure scripts assert that every plotted point equals the archived value; and the appendix-table scripts bind all printed cells to the archives (103 / 100 / 67 assertions for Appendices B, C, E respectively). Scale experiments use seeds $\{42, 43, 44\}$ (compute constraint); MNIST phases use seeds $\{42, \ldots, 46\}$; single-seed phases (0, 1, 2, and the T2 sweep) are registered as such.