# The Compendium

*Everything we measured on the White-Box ARC Estimation Challenge, and what each measurement killed*

trim_qewas · companion to “Measuring the Ceiling of a Method Class” · submission 326725

17 August 2026

# Contents

**Everything we measured on the White-Box ARC Estimation Challenge, and what each measurement killed**

*Companion volume to Measuring the Ceiling of a Method Class. That document argues a thesis; this one is the evidence base behind it — the full register of what was measured, what the number was, what sealed it, and what it closed.*

## How to read this

The problem: given a bias-free He-initialised MLP (width 256, depth 32, ReLU) and `z ~ N(0, I)`, predict `E[ReLU(W₃₁ᵀh₃₀)]` analytically, under a FLOP budget.

> `score = final_layer_mse × max(0.1, C/B)` `C = F + 1e11·R` `B = 2.72e11`

Our submitted estimator is a two-moment Gaussian closure blended with a small Monte-Carlo branch. Best adjusted score `1.855e-07`, rank #66 at the Phase-1 deadline.

**The corpus behind this document:** 396 sealed findings, 87 memory shards, 600 measurement stands, ~5.5 MB of accumulated notes across 109 working sessions. The great majority of it is negative results, and that is the point: we spent the project mapping the boundary of a method class rather than tuning inside it.

**Every entry follows the same form:** *claim → number → seal → verdict.* A claim without a seal is marked as derived. Where a later measurement contradicted an earlier one, both appear, with the retraction.

**One warning that applies throughout.** Nearly every number below was, at some point, wrong for a reason that had nothing to do with the mathematics: a constant measured on a different subset, two different objects sharing a name, an adjusted score compared against a raw MSE. Part VII is the list of those. We consider it the most useful part.

The central instrument of the project. Rather than build a better method and see, we substituted **ground truth** into one component at a time and measured what a *perfect* version of that component would be worth. This prices any idea before it is written.

## I.1 The master table

Everything below is measured in `×score` on the official grader or in analytic precision converted through the blend model (§I.3). Our base is `1.0`.

| what is made PERFECT | ×precision | ×score | verdict |
|---|---|---|---|
| the step formula (exact per-step error) | 1.19 | 1.032 | ceiling of a whole class |
| linear correction in the closure's own basis | 1.027 | 1.001 | ceiling of a whole class |
| 2-D quadrature instead of truncated Mehler | 1.001 | 1.000 | the series is not the problem |
| genetic-programming search over expressions | 1.002 | 1.000 | converged to a hand-found constant |
| rank-r non-Gaussian step kernel (r = 4) | 1.04 | 1.001 | – |
| per-neuron Edgeworth on OUR moments ($\gamma_3$,$\gamma_4$) | 1.02 | 1.000 | needs exact moments to switch on |
| true OFF-DIAGONAL of C | 1.149 | 1.012 | the covariance channel is nearly empty |
| true DIAGONAL of C | 0.118 | 0.930 | ← makes it 8.5× WORSE |
| true C entirely | 1.098 | 1.008 | ← worse than off-diagonal alone |
| true m AND true C (`allm`) | 30.86 | 3.363 | ⇒ 5.5e-08, about 5th place |
| + true per-neuron $\gamma_3$,$\gamma_4$ on top (`edgm`) | ~160 | 22.78 | ⇒ 8.16e-09, between #2 and #3 |
| per-network blend weight W*_i | – | 1.029 | realisable version: ×0.967 |
| per-network scale a_i | – | 1.25 | realised on grader: ×1.003 |
| knowing the true (m,C) after layer K=1 | – | 1.00 | shallow layers are worthless |
| knowing the true (m,C) after layer K=24 | – | 3.7 | the prize is at depth |
| exactness through 3rd Hermite degree | 3.4 | – | ±0.6; our estimator sits on it |
| exactness through ANY Hermite degree | 7–9 | – | ceiling of SMOOTHNESS itself |
| any additive one-dimensional class | 1.89 | – | the bottleneck is additivity |
| weights read at fewer than 10 bits | – | – | closes coarsened-weight schemes |

**The last block needs one line of explanation, because two of its rows are often conflated.** Degrees 1–3 hold `0.74 ± 0.05` of the target's Hermite variance, so a method exact through third order is capped at `×3.4 ± 0.6`. But `10–16%` of the variance sits at degrees `≥ 7`, which no finite-degree method removes, so a method exact through *arbitrarily* high degree still saturates at `1/(0.10…0.16) ≈ ×7–9`. The first number is what third-order exactness buys; the second is the ceiling of smoothness itself, and it is the one that closes quadrature, cubature, QMC and polynomial control variates at any price. **Seal.** The mechanism is measured, not argued: replacing ReLU by `softplus_β` collapses the degree-`≥7` tail by `3–7×` while the function is still `55–66%` nonlinear — a kinked function has a power-law Hermite tail, and no polynomial takes it.

## I.2 The single most consequential row

**`allm` — a two-moment closure with *exact* moments at every layer — scores `5.5e-08`.**

```
#1 on the board   1.50e-09     ⇒ ×37 above a perfect two-moment closure
#2 on the board   7.20e-09     ⇒ ×7.7
#3 on the board   3.68e-08     ⇒ ×1.5
#4 on the board   4.62e-08
US WITH ORACLE    5.50e-08     ≈ 5th place
```

**No moment-based method reaches #1, even with a perfect oracle — and against #2 our two instruments straddle parity, so we make no claim there.** A full oracle including $\gamma_3$, $\gamma_4$ was measured twice and the two disagree by `×1.8`, so both are quoted:

| | | vs #1 (1.50e-09) | vs #2 (7.20e-09) |
|---|---|---|---|
| direct injection | 8.16e-09 | ×5.4 | ×1.13 |
| sieve | 1.48e-08 | ×9.9 | ×2.1 |

Direct injection (12 networks × 524288, convergence ratio `0.0005`) is the stronger instrument by rule 1 — inject the exact quantity, do not extrapolate a fit — and it is also the number least favourable to our own thesis, so it is the one we argue against. The structural statement that survives on both: **no moment-based state reaches #1**, by `×5.4` even on the most favourable measurement.

Seals: the `base` row of that stand reproduces the grader's own figure to four digits (`0.5072%` vs `0.5070%`); `allm` landed at `0.0913%` against a pre-registered prediction of `0.10–0.20%`; at layer 0, increasing `N` by ×64 dropped both figures by exactly √64; two independent replicas (seeds 777 / 31337).

## I.3 The blend model — how any analytic gain converts into score

A single knob in our estimator isolates the branches, and the **official grader** was run at both settings:

```
_W_BLEND = 1.0  →  pure analytic closure   final-layer MSE  2.3385e-05
_W_BLEND = 0.0  →  pure Monte-Carlo branch                  1.9227e-06
submitted blend                                             1.8115e-06
```

**Seal.** For two independent unbiased estimators, `MSE = V·E²/(V+E²)`. Predicted `1.7767e-06` against measured `1.8115e-06` — **2% agreement, on two different MLP subsets**. Every "×score" figure in this document passes through this model.

**The leverage is not constant** — the single most expensive thing we did not know for the first ~90 sessions:

```
precision ×k     score      ×score    leverage d log(score)/d log(k)
    1.5       1.694e-07     1.095        0.22
    2.0       1.510e-07     1.228        0.30
    4.0       8.67e-08      2.139        0.55
    8.0       3.21e-08      5.787        0.85
   15.0       1.03e-08     18.02         1.07
```

⟹ **while Monte-Carlo dominates the score, a ×1.05 analytic gain shows up as ×1.01.** We discarded ideas for sessions on end through this suppressed leverage. **Evaluate an analytic branch on its own error, never on the blended score.**

## I.4 The sieve — how much moment accuracy is worth

Measured on the grader's own instrument, 50 networks:

```
improve (m,C) by ×2   →  score ×1.190
                 ×4   →         ×1.652
                ×10   →         ×2.301
           perfectly  →         ×2.519   ← ceiling of the class from this direction
```

**The curve is concave.** This retro-predicts every one of our channel-correction results: "no channel correction ever exceeded ×1.13" is not a property of the corrections, it is a property of the curve. Fifteen branches returned `×1.0×` *by construction*.

**Channels multiply, they do not add.** `(m,C)` ×10 without γ gives ×2.30; oracle γ on *our* `(m,C)` gives ×1.002; together ×7.75. ⟹ **Gate: does the idea touch both channels? If not, it is dead before implementation.**

Fifty-four classes were closed with a number. Grouped by what they tried to fix.

## II.A Moment-based refinements — closed by the sieve above

| class | measured | mechanism |
|---|---|---|
| higher-order Mehler (K=4, K=24) | ×1.000 | the series was never the error |
| exact 2-D quadrature | ×1.001 | same |
| Edgeworth in the recursion, tabulated γ | ×0.80 … ×1.027 | half the gain survived the *inverted-sign* control ⟹ it was calibration, not information |
| Edgeworth with true per-neuron γ, our moments | ×1.02 | needs exact `(m,C)` to switch on |
| $\kappa_3$ (third cumulant) carried in the recursion | ×0.38 … ×1.088 | closed at **all four** price points; >half of what it fixes is *scale*, which the global constant takes for free |
| co-skewness / co-kurtosis tensors | 4.00e11 FLOPs = 1.47×B | later reopened by the rank-3 discovery (§IV.4) and closed again by the ceiling |
| saddle-point / CGF | ×3150 past the pre-registered threshold | a CGF truncated at the 4th cumulant **is not a CGF** |
| maximum-entropy density from 8 moments | ×31 in a stand that never ran the accumulation | superseded by the step ceiling ×1.19 |
| Johnson S_U | ×1.14 | this *is* the width of the white spot — not an order of magnitude |

**The umbrella verdict (§I.1):** injecting the **exact** step error gives ×1.19. Every row in this table is bounded by that one number, including rows nobody has written yet.

## II.B Sampling-side improvements — closed by one test

We have exactly **four** exactly-known truths available in this problem. The complete inventory:

```
second moments of z     E[zzᵀ] = I              → whitening        ×2.34  ✓
norm of z               E‖z‖ = 15.984382667     → spherisation     ×0.74
fourth moments of z     E[(u·z)⁴] = 3‖u‖⁴       → measured         ×1.00
E[h₁]                   ‖w‖/√(2π)               → the λ-block      ×1.14  ✓
```

⟹ **there is no fifth.** Everything else requires substituting our own estimate, which converts variance reduction into bias.

**And one test closed the remaining family at once.** Ten networks were run against **60 shared sample sets** — the same `z` batches through different networks. If a sample batch is "bad", it must be bad for everyone:

```
signed error, correlation between networks    +0.0120 ± 0.1459
magnitude of error                            −0.0004 ± 0.1044
†  control (independent batches)               +0.0103 ± 0.1367   ← identical
```

⟹ **universal sample quality does not exist.** Monte-Carlo error is network-specific. This kills RQMC, spherisation and fourth-moment corrections with a single argument instead of three.

**Law: when several features all return zero, do not test the next feature — test whether a common signal exists at all.** One measurement replaces an infinite search.

**Also closed here:** - **RQMC (Roberts lattice + Cranley-Patterson)**: first run ×1.14, verdict run (24 networks × 8 shifts, paired) **×1.0156, sign test 12/24, median ×0.9897**. The original number was noise. Mechanism: whitening already fixes all 32 896 second moments, leaving RQMC almost nothing. - **Non-uniform sample allocation across networks**: "+7% free" was an in-sample artefact, killed four independent ways — the seal was empty by construction, the aggregate is separable so the optimum is uniform *structurally*, 67% of the spread is realisation noise, and cross-seed the branch is worse than uniform. - **Robust / winsorised estimators**: output kurtosis +0.373, top 1% carries 10.28% of the variance against **8.41% for a Gaussian** — the tail is heavier by only ×1.22. Winsorising costs ×0.874. - **Rank truncation to cheapen the forward pass**: softest setting gives 0.377% bias against a pre-registered threshold of 0.06% — a miss of 6.3×, and the bias exceeds the entire Monte-Carlo error it was meant to buy.

## II.C Per-network personalisation — closed by one mechanism

Four separate quantities showed the same shape: **an oracle demonstrably helps, and every deterministic proxy for it is orthogonal.**

| quantity | oracle | best deterministic proxy | sample route |
|---|---|---|---|
| per-network blend weight W*_i | ×1.029 | ×0.967 (cross-validated) | – |
| coefficients a_k of a basis | ×1.718 | ×1.002 (55 features, 1000 networks) | ×0.956 |
| mixing-direction correction | ×1.34 | – | ×0.956 / ×0.588 |
| per-step error b_l | ×1.189 | ×1.000 (tabulated Edgw.) | ×1.011…1.024 |

**The mechanism for `W*`, measured rather than argued.** The natural proxy (`avg_variance`) correlates with `V` at 0.589 and with `E²` at 0.616 — and is useless *for exactly that reason*. `W*` depends on their **ratio**, in which the proxy cancels: slope of `log V` on `log avg_var` = 0.5514, slope of `log E²` = 0.5015, so the slope of `log(V/E²)` = **0.0499**.

**Law: a proxy must correlate with the TARGET, and the target may be a ratio.**

**The one survivor tells you why the others failed: sample size.** A learned *per-neuron* correction did cross a held-out split (`×1.011…1.024`, foreign control returning exactly `0.00`). Predicting a per-neuron quantity gives 196 608 examples; predicting a per-network constant gives 50–1000. **Closed.** A "micro-network per constant" (2209 parameters on 50 observations) is dead arithmetically — even a *single* scalar fitted on 50 networks already overfits (A ×1.036 → B ×0.9969). The right family size is **2–4 parameters**.

**Capacity law measured directly:** overfitting penalty = `11.76% · k/N` (k parameters, N networks with ground truth), coefficient stable at 9.6–15.4% across two independent runs.

## II.D Structural and inverse approaches

**The backward operator has effective rank ~3.** This closes an entire family.

`δ_out = B_l · δ`, where `B_l = Φ_{l+1}∘W_{l+1} … Φ_31∘W_31` — and `Φ` is already in the closure, so the cost is zero. We had measured forward transport many times and the spectrum of the *backward* operator never once.

```
l→31   eff. rank   S₁/S₂₅₆    energy of B in top-16   ERROR in top-16
   0      2.74     1.4e+31           99.89%              5.52%
   8      5.78     3.0e+29           99.35%              6.95%
  16     11.20     1.0e+31           96.53%              6.06%
  24     23.85     6.6e+28           83.05%              6.14%
  30    100.49     3.6e+14           29.84%              5.11%
†  isotropic control: 16/256 = 6.25%
```

**Condition number $10^{31}$.** Of 256 numbers at a shallow layer, **2.74** survive to the output ⟹ information is destroyed irreversibly. The class "inverse problem / reconstruct the state from the output" is closed **by degeneracy of the operator**, not by cost. This is a property of the *network*, not of our method.

**Mechanism.** It also explains, without any further run, why injecting exact moments after layer K=1 or 2 is worth `×1.00…1.01` while K=24 is worth ×3.7: the prize is at depth, the only exactly-known anchor is at layer 1, **and they sit at opposite ends**.

**And a clean second proof that "aim at part of the state" is dead:** the error in the top-16 subspace is 5.5–7.0% against an isotropic 6.25% ⟹ the error is *exactly isotropic* in the operator's own basis.

**Also closed here: - Euler + Stein identity.** Homogeneity plus Gaussian integration by parts give the *exact* relation `E[out] = E[Δout]`, where `Δout` is a measure on the ReLU kinks. **Seal.** Verified analytically at L=1 (`‖w‖²δ(w·z) ⟹ σ/√(2π)`, exactly the truth) and statistically at L=32 (`mean(out−Δout)/se` = −0.19, −1.90, −0.27, +0.24). **Closed.** Dead because it needs the density at zero, and at layer 31 the error in `E[ReLU]` is `0.107%` while the error in `ρ(0)` is `7.27%` — a factor of **68**. As a control variate: `corr = 0.117`, variance reduction ×1.014, at ×3 the FLOPs. - **Hybrid (closure to layer K, then Monte-Carlo from `N(m_K, C_K)`)**: at K=24 with 4× the samples, ×0.068. The hybrid does not average away the closure error — it **preserves** it. - **Spectral (Wiener) blend per eigendirection**: cross-validated ×1.013, the gain lives strictly in the eigenbasis (`eigh` = 0.59% of budget) ⟹ eaten by its own cost. - **"Almost-linear network"** (freezing the mask at its most likely value): relative error **53.95%**, ~100× worse than the closure. - **Width extrapolation / Richardson in `n`.** The `1/n` law is real and now confirmed *inside a single network* (slopes −0.897 … −1.135 across n = 32…256, after correcting the instrument, which initially changed the *gate regime* rather than the width). But the entire ×1.22 gain is **pure scale**, which the global constant already takes: raw error ×0.986, **cosine error ×0.9999**.

- **Virtual weights — `k` representative linear surrogates.** For a fixed input the network *is* linear: `out(z) = z @ M_{D(z)}` with `M_D = W₀D₀ ⋯ W₃₁D₃₁`, hence `E[out] = E[(M_D − M̄)z]` — the answer is **purely the covariance between which linear map you got and which input you got** (seal: gates taken from a different sample give exactly zero, `0.96` of that quantity's own noise floor). Conditional on the gate pattern everything is exactly Gaussian, so a `k`-component mixture would have zero step error by construction. It fails on the size of the ensemble: its exact effective rank is **1034** (397…1558), and a linear rank-`k` truncation evaluated **out of sample** returns `17% / 26% / 37% / 49% / 74%` of the answer at `k = 8 / 16 / 32 / 64 / 512`, against a random-subspace control of `0.2 / 0.4 / 0.8 / 1.6 / 12.5%`. The concentration is real (`×30`) and still far too weak: the per-doubling gain peaks at `k ≈ 64` and decays, while beating the closure needs ~99.5% of the answer. **Mechanism.** The reason ~10 earlier low-rank schemes failed is now explicit: they were justified by the rank of the **state** (`165.6 → 2.8` with depth, `×59`), whereas what must be represented is the **ensemble of maps** (`782 → 201`, only `×3.9`). Same numbers, different objects — all roads lead to nearly the same place, but they remain very different roads. And the rank has a source: flipping a quarter of the negative weights to positive collapses it `200 → 9.6`, and flipping all of them gives exactly `1.01` — the linear network, which is solvable in closed form to `1.7e-06`. Criticality creates the rank; the rank predicts the closure's relative error (`corr = +0.80`, 50 networks).

## II.E Learned and amortised methods — closed at seven students plus a diagnosis

The family a modern reader reaches for first, and the one we spent the longest arc on. Target: a surrogate from an intermediate activation `a¹⁶` to the output — legitimate, because the second half of the network is nearly linear (`R²` of a linear head: `0.44` at layer 0, `0.88` at 16, `0.99` at 30).

```
student                              bias²        verdict
per-neuron affine  γ·z+β            ×1.00        provably a SUBSET of antithetic sampling
per-neuron nonlinear 1-6-6-1        600–2400×    pointwise fit error 1.9e-04 amplifies
wide linear (OLS) 256→256           1e-07 ✓      but needs E[a¹⁶] — the cost wall
wide nonlinear 256-128-64-256       2e-05        200× worse than the LINEAR student
residual/skip 256-32-256            6–40×        tail blow-up: pred |max| = 320, signal 0.5
bounded activation (tanh)           2e-05        stable, 24× worse when cost-matched
folding pairs of early layers       255× / 10⁶×  linear / MLP
```

**The task penalises capacity.** OLS guarantees a zero-mean residual for free; SGD does not, so flexibility bends exactly the estimated quantity. **Linearity is protection here, not weakness. Unboundedness, not architecture, causes the blow-ups**: every unbounded activation reaches `pred |max| = 2.7e5 … 7e11` on the heavy tails of `a¹⁶`; `tanh` holds `5.6` and loses on bias instead. **Knowing `E[control]` costs as much as the answer** — cost-matched from 4k to 200k samples the surrogate is `1.07–2.41×` worse than plain sampling and **never crosses**.

**Amortisation removes the cost and not the wall.** An offline-trained hypernetwork (60 networks, 48 train / 12 test) learning the *residual* of a cheap analytic base:

```
base only                                test MSE 5.64e-01
base + learned residual                          5.88e-01   ← worse, fits noise
†  corr(predicted residual, true residual)        0.0012
```

**Mechanism.** The diagnosis is about the object: the truth varies across networks (`rms 0.50 → 1.36`) while the cheap base is essentially **the same number for every network** (`rms ≈ 0.55`), because mean-field variance collapses to a fixed point under criticality. ⟹ **there is no cheap descriptor of an instance from which the answer can be recovered**; computing the descriptor *is* the work.

**And approximating ReLU is actively harmful**, which the same arc established from four directions: a learned `1-6-6-1` replacement with pointwise error `1.9e-04` gives `600–2400×`; a Gaussian anchor mid-stream `+2e9%`; truncating the recursion at layer 24 `7×`; averaging mid-stream `400×`. And there is no compressibility in the weights either: rank-`r` truncation gives `0.52 / 0.40 / 0.05 / 0.0071 / 6.8e-07` MSE at `r = 16/128/192/224/256` — discarding 12% of the spectrum is still `1700×` worse. iid Gaussian weights are full-rank **by structure, not redundancy**.

## II.F Approximation theory — six frameworks, one number

Rational approximation (Newman, Padé, AAA, lightning solvers) · ridgelets and the Radon transform · nonlinear approximation in Besov spaces · Gegenbauer reconstruction · Gaussian measure of polytopes · Euler–Maclaurin with jump corrections.

**Mechanism.** All six reduce to one question, because we need an *integral*, not an approximation, and in 256 dimensions the only object that integrates exactly against a Gaussian without sampling is a ridge function `g(w·z)`. So the shared ceiling is the variance share of `c + Σ_{i≤m} g_i(u_i·z)` with arbitrary `g_i`:

```
m              1      4     16     32     64    128
share taken  0.314  0.359  0.425  0.455  0.471  0.466   ⟹ ×1.46 … ×1.89
†  linear m=1 → 0.999 · Σ₈relu m=8 → 0.998 ·    random directions m=32 → 0.029
```

×1.89 equals the ×1.93 of a purely polynomial control variate ⟹ **an arbitrary one-dimensional class buys nothing over polynomials. The bottleneck is additivity, not the basis. Seal.** Newman's theorem tested at *our* parameter: 1384–1945 kinks along one Gaussian line, and the rational advantage is gone by K = 16.

**And the kinks are a dense medium, which closes the opposite family too.**

```
†  layer-0 control (∇p = w ⟹ d must be exactly |N(0,1)|)
quantile     0.05     0.25     0.50     0.75     0.95
measured    0.0621   0.3177   0.6737   1.1505   1.9577
theory      0.0627   0.3186   0.6745   1.1503   1.9600

median distance to the nearest kink   0.00011
mass within ε of a kink               ε = 0.001 → 99.1% · ε ≥ 0.003 → 100.0%
†  forced by counting: ~8000 hyperplanes × φ(0) ⟹ 1.6e-04 predicted, 1.1e-04 measured
```

⟹ cells, poles, Gegenbauer and targeted sampling near the hyperplanes are **the wrong class**, because there is nothing to localise. Exact per-cell integration is closed without a stand: a cell has diameter ~$10^{-4}$ in 256 dimensions.

**MLMC over depth**, measured with the full Giles telescope: a pre-registered ladder gives 0.961× ± 0.117, the best searched ladder 1.480× (chosen on the same data, so an overestimate), **Seal.** with both controls exact (one level 1.0000×, foreign network 0.26×). **Below our plain estimator.** What survives is the $\rho^2(\ell)$ profile, which caps *any* "anchor at layer $\ell$ plus sampling" scheme.

**Finite-size AMP / state-evolution corrections**: priced before opening the branch by interpolating the closure toward truth on layers 1–8. Perfect closure there is worth ×1.2535 and saturates; the next Plefka order is a $256^3$ tensor. Entry cost known, return bounded.

## II.G Budget and cost tricks — all empty

```
economising the multiplier      we sit 0.4% above the floor; buying it back
                                costs exactly what it returns ⟹ net zero
many submissions, take the best the grader is deterministic: two submissions of the
                                same archive gave identical MSE and 0.05% score spread
client-side compute via residual DISALLOWED by the organisers mid-competition
```

**The leader sits on the same budget floor we do** (BUDGET USED 9.59% vs our 10.04%) ⟹ **there is no budget trick**; the entire difference is MSE. And our whole analytic closure costs 1.66e9 = **6.1% of the floor** ⟹ a deterministic scheme could be **×16 richer**. **Cost is not the gate, and has not been for 40 sessions.**

Six things, in order of contribution. Everything else in this document is a negative result.

```
whitening the sample                  ×2.34 in the Monte-Carlo branch
λ-block (exact E[h₁] = ‖w‖/√(2π))     ×1.1369 value, 0.1324% of B cost ⟹ payback ×10.9
dead-neuron pruning                   ×0.776 MSE (−22%), 0 failures
κ = 0.9985, a single scalar on C      ×1.0039 on the grader, at ZERO cost
per-network scale a_i                 ×1.003 on the grader (oracle said ×1.25)
all_layer_means (free, panel only)    ×41345 on a secondary score, cost zero
```

## III.1 κ — the third kind of change

Our own dichotomy had been "channel correction" (never exceeded ×1.13) versus "global replacement" (never below ×3.1). κ is neither: **a global knob inside the recursion.**

```
sub116 = sub99 + κ = 0.9985 on layers i > ksig, with α re-fitted jointly
LOCAL:   mini ×1.0092 (33/50, p=0.016) · full ×1.0209 (36/50, p=0.0013)
         pure closure (W_BLEND=1.0)  2.338512e-05 → 2.215759e-05  ×1.0554
GRADER:  1.86762e-07 → 1.859e-07     ×1.0039
cost:    65536 FLOPs × 23 layers = 1.5e6 = 0.0006% of B
```

**Seal.** LOO across 50 networks selected `κ=0.9985` in **all 50 folds**. A 2-D scan with separate multipliers on the diagonal and off-diagonal put the optimum exactly on `kd = ko` — a direct pass of the consistency gate. **Seal.** Prediction before the run: `2.2155e-05`. Measured: **2.215759e-05** (four digits).

**The motive was wrong and the result still holds.** The closure *understates* σ, monotonically with depth: −0.08% (L1), −4.24% (L16), **−10.56%** (L31). A shared multiplier removes −16.2% of that. **But the optimum for κ lies BELOW 1** — i.e. what is being cured is not σ, it is the *balance*. This is the fifth time damping inverted the expected sign.

## III.2 Two traps κ cost us, both repeatable

```
(1) κ on ALL layers breaks the λ-block. `_closure_parts` returns not only `m_deep` but
    also the supports s_cl, m_cl, q_cl, P_cl, a_cl, b_cl — the control variate of the
    Monte-Carlo branch, which is 91% of the score. Shifting them gave ×0.57.
       LAW: a stand measuring ONE branch cannot see that your change feeds the OTHER.
(2) α must be re-fitted jointly with κ, and it is specific to each layer configuration.
    Carrying α from "κ on all 32" into "κ on i>8" gave 3.11e-05 instead of 2.21e-05.
    †  SEAL: the penalty was predicted as (Δα·RMS m)² = 9.0e-06; 2.2145e-05 + 9.0e-06
       = 3.11e-05 against measured 3.113026e-05.
       LAW: calibration constants do not transfer between configurations.
```

## III.3 The cosine insight

```
min_a ‖a·m − T‖² = ‖T‖²·(1 − cos²(m, T))
```

⟹ **the MSE after the optimal scale is literally a cosine loss**: magnitude does not matter, only angle. Holding the scale with one constant across all networks is therefore a loss — and in 100 sessions nobody had asked. The oracle per-network `a_i` is worth ×1.2525 in the closure, and **Monte-Carlo noise does not interfere** (×1.2519), because `a_i` is one scalar fitted over 256 numbers.

**Consequence adopted as a law: measure any correction to `m` on the COSINE error, not the raw error.** The gap between them is ×1.9, and inside that gap sits exactly what the global scale constant takes for free. Three separate branches passed LOO, A/B cross-validation *and* the foreign control while delivering nothing, because none of those three catches a scale-only effect.

These are the findings we would keep if the competition disappeared.

## IV.1 A tuned closure is a balanced system of cancelling errors

```
true diagonal of C substituted            ×0.118   (8.5× WORSE)
sharpening C toward its true spectrum     ×0.008   (125× worse), rank hit its true value
the σ channel: predicted at R² = 0.836    optimal correction coefficient 0.00
true C entirely  <  true off-diagonal alone
κ optimum 0.9985 < 1, while the true C at L31 is ~25% LARGER
```

**The measured direction of the error and the direction that helps are opposite. ⟹ Practical rule: substituting truth into one channel of a tuned closure is expected to hurt.** Any evaluation of a "more accurate submodel" must re-fit the remaining constants jointly, or it measures their staleness. We violated this rule once and reported a ×1.031 that shrank to ×1.011…1.024 under joint recalibration.

## IV.2 The error is born near the output, and it is not additive

```
per-step error at exact inputs       0.139–0.162%
accumulated error                    0.48–0.67%      ⟹ accumulation is a factor of 3–4
‖b_l‖/‖T‖ by layer (0,1,8,16,24,31): 0.061 0.079 0.095 0.079 0.064 0.046 %
sum of norms ≈2.2% against accumulated 0.48%         ⟹ step errors partly CANCEL
injecting the exact step error                       ×1.19 only
†  the accumulated figure is stand-dependent (0.48% on the 16-network step stand,
   0.67% on the earlier one); the factor is 3–4 either way and nothing turns on which
```

**The local correction claims 16% of the error; the remaining 84% is not "transport as a separate object" — it is that after the very first layer the state has DIVERGED, so a correction measured at exact inputs is no longer the right correction. ⟹ "fix the step" ≠ "fix the method".** The shortest explanation for why a hundred sessions of local corrections returned ×1.0x.

**Closed.** The hypothesis that the correction merely misses its point of application was tested directly (transporting `b` into our own state by homogeneity) and **refuted**: +0.7%.

## IV.3 The network freezes with depth, and that saturates the error

| layer | 0 | 1 | 4 | 8 | 16 | 24 | 31 |
|---|---|---|---|---|---|---|---|
| median \|t\| | 0.003 | 0.466 | 0.965 | 1.588 | 2.191 | 2.593 | 3.274 |
| switching neurons | 100% | 99.9% | 90.0% | 66.5% | 49.7% | 40.5% | 30.9% |
| frozen fraction | – | 0.318 | – | 0.834 | 0.882 | 0.923 | 0.943 |
| eff. rank of Cov(h) | – | 165.6 | – | 22.7 | 10.3 | 4.5 | 2.8 |

`frozen = ‖m‖²/(‖m‖² + tr C)` — the share of the state that no longer depends on the input. Effective rank decays ×0.8663 per layer ⟹ **half-life 4.83 layers**.

**A seal we did not plan:** `frozen = 0.943` implies a typical `t = m/s = 4.07`, while an independently recorded `t* = √SNR` reads **4.20** ⟹ the mixing metric and the gate regime are **the same quantity**.

**The chain that explains a hundred sessions in one sentence** — stated with its domain, because we first stated it without one:

```
mixing (half-life 4.83) → frozen 0.94 → t ≈ 4 → ReLU nearly linear over the bulk
→ E[ReLU] ≈ m → THE SHAPE OF THE DISTRIBUTION DOES NOT ENTER THE ANSWER
```

**Caveat.** Every link is depth-dependent. `median |t| = 3.50` is a *deep-layer* value; at layer 1 it is `0.466`, a factor of 7 off. The argument survives because shallow layers barely reach the output (backward rank 2.74 at `l=0`).

**Lesson.** It also explains the saturation law `error ~ L^0.28`: the error does not fail to accumulate because it is "forgotten", but because **the source of nonlinearity is exhausted** — each deeper layer adds less new switching.

## IV.4 96% of the non-Gaussianity comes from dependence between neurons

A shuffle control: each coordinate of `h` is permuted independently, preserving marginals and destroying dependence.

```
layer   eff. rank Cov(h)   γ₄ real   γ₄ SHUFFLED   corr real   corr SHUFFLED
   1        165.6           0.0380     0.0281        0.2328        0.2494
  16          9.5           0.2799     0.0090       −0.0172        0.1379
  31          2.8           0.3947     0.0142       −0.0098        0.4034
```

**Warning.** `γ₄` falls from `0.395` to `0.014` under shuffling ⟹ **the non-Gaussianity is a dependence effect, not a marginal one**. **Seal.** The mechanism is *demonstrated*, not argued: on shuffled data the correlation **revives** to 0.4034 exactly where it is zero on real data ⟹ it is the dependence that kills it.

**Mechanism.** `γ₄ × eff.rank` ≈ 1.1 at layer 31 ⟹ **a preactivation is effectively a sum of ~3 terms, not 256.** This is why the naive `1/256 = 0.39%` never matched anything. **Seal.** Replicated across 27 sessions: PR-rank 161.6 → 2.5 versus 165.6 → 2.8.

## IV.5 The bulk/tail gate

```
deep layers (t ≈ 3.3)   the zero sits 3.3σ out ⟹ E[ReLU] ≈ m lives in the BULK
                        density at zero, orthant probabilities live in the TAIL
†  confirmed from both sides, two independent stands:
   with depth the error in E[ReLU] FALLS   0.228% → 0.107%
   with depth the error in ρ(0)     RISES    0.64% → 7.3%
```

**This explains an orthogonality we caught three separate times and could not name.** ReLU activation masks carry a real, long-range, non-moment structure: inter-layer co-activation deviates from the Gaussian orthant prediction by **4.883%** against a noise floor of **0.276%** (×17.7), and it does *not* decay with layer separation. Five seals, including the cleanest in the project: `N` ×32 dropped the noise ×5.7 while the signal moved 4.810% → 4.809%.

**And it is worth nothing to us** — because it is a *tail* object and our functional is a *bulk* object. They are orthogonal not through noise but because they are **different parts of the same distribution**.

**Gate adopted for every subsequent idea: is it about the bulk or the tail?**

## IV.6 The error is a finite-width `1/n` effect

```
n            32     64     96    128    160    192    224    256
error      8.45%  4.54%  2.52%  2.22%  1.67%  1.38%  0.93%  0.958%
slopes       −0.897 −1.100 −0.964 −1.007 −1.011 −1.135 −1.047   ⟹ exactly 1/n
†  slopes are PER-NETWORK, then aggregated; the error row is the aggregate ⟹ the two
   rows do not recompute from each other (pairwise slopes of the error row: −2.6 … +0.2)
```

**⟹ the closure error is a fluctuation of the specific realisation of `W`, not a missing term in a moment series.** This explains in one sentence why Mehler K=4 changed nothing, why no shared residual basis exists across networks, why the `κ₃` ceiling is only ×1.42, and why ~10 low-rank schemes all died at the same 10%.

**Warning.** The instrument lied at first: truncating `W[:n,:n]` leaves `Var(W) = 2/256`, so it changes not the width but the **gate regime**. Without the `√(256/n)` renormalisation the slopes read −17.5 … −7.2 (garbage).

## IV.7 The ground truth is itself Monte-Carlo

`metadata.json` of the public dataset: `"n_samples": 1000000000`, with a column `avg_variance` = `mean_j Var(out_j)`.

```
noise floor of final_layer_mse = avg_variance/1e9 = 4.949e-11 (mini) · 5.219e-11 (full)
†  seal: mean(true²) = 0.909290 against a previously recorded 0.9097
```

⟹ the leading entry at `3.628e-09` sits **×73 above the floor** ⟹ the competition has not hit the problem's ceiling.

**And it decodes the leader.** Their per-layer[0] figure equals `6.769e-10`, which is *the sampling noise of the ground truth itself* (theory: `Var(ReLU(W₀z))/1e9 = 6.8316e-10`; direct measurement on the dataset: `6.8714e-10`) ⟹ **they compute layer 0 exactly and do not compute layers 1..30 at all** — not even as a by-product. Their computation does not traverse the network.

**Caveat.** Two floors, and conflating them is easy: `4.949e-11` is the floor of the **final** layer, `6.83e-10` the floor of **layer 0**. They differ by `×13.8`, because the network freezes with depth. The leader's layer-0 figure matches the layer-0 floor to `×0.99`; the `×73` headroom above is measured against the final-layer floor. Both statements are true, of different quantities.

### The rest of the leader profile, for the record

```
NOT moment-based       full (m,C,γ₃,γ₄) oracle = 8.16e-09 (direct) / 1.48e-08 (sieve)
                       ⟹ ×5.4…×9.9 above #1; against #2 the instruments straddle parity
NOT layer-wise         layer-0 error = the truth's own noise (above)
NOT pair-coordinate    information ceiling in (t_i, t_j, ρ, depth) = 0.795, need 0.992
DETERMINISTIC          vs-sampling ratio 1782× · per-network spread ×9.6 (ours: 88.5%)
CAN AFFORD ALGEBRA     2.61e10 effective FLOPs = 15.7× our whole closure, with 10×
                       headroom unused ⟹ cost does not bind them either
```

**Caveat.** These come from the leader's public submission panel plus our own stands. The panel fields are the same ones any entrant who leaves `all_layer_means` unfilled would show, so we read them as evidence about *their* method only where they differ from ours — rows 2, 4, 5.

Reusable. Each one is the reason some verdict is trustworthy.

```
_W_BLEND = 1.0 / 0.0        isolates the analytic and sampling branches ON THE GRADER
oracle substitution stand   ceiling of ANY channel in ~15 minutes on one GPU
the foreign control         same correction shape, content from a DIFFERENT network
the scaling seal            N ×16 ⟹ noise falls ×4, real error stands still
the tautology seal          exact quantity over exact inputs must return exactly 0
paired grader runs, JSON    `plain` rounds to 3 digits and hides a 0.45% effect
AICrowd API                 returns MSE separately from budget (`score_secondary`),
                            deterministic to 17 digits ⟹ resolution 0.1% → 0.001%
```

## V.1 Seals that caught the instrument rather than confirming it

These are the entries we are most confident in, because the seal failed first.

```
`injm` must be exactly 0 — in the smoke run it read 1.89%, catching an indexing error
   (the input of the NEXT step is (m_{l+1}, C_{l+1}), not C_l). Without it the verdict
   on the entire "better step" class would have been wrong.
`μ=0` must reproduce the base 0.507% — gave 0.4178%; `μ=1` must reproduce allm 0.0913%
   — gave 0.0237%. This revealed that re-injecting an oracle at EVERY step suppresses
   accumulation, which is 84% of the error, and that the stand's score column is a
   shape and not a level.
the width instrument gave slopes of −17.5 before renormalisation (§IV.6)
`ground.py` was found to have a FAILED seal and to have issued a verdict anyway
```

**Law: an oracle stand must carry a seal at BOTH ends of its axis** — reproducing a known number at zero substitution and at full substitution.

The distilled version. Each one was paid for with at least one wrong verdict.

## On measurement

1. **Measure the ceiling of your class before optimising inside it.** One afternoon of oracle substitution would have saved us many weeks.
2. **Measure a ceiling by injecting the EXACT quantity, not its best approximation.** We spent three sessions grading approximations of a step correction and never asked what a *perfect* correction was worth. One run answered it and closed six branches.
3. **Measure the TRANSFER FUNCTION of a channel, not its own error.** The largest raw error (off-diagonal, 4.08%) is the cheapest channel (×1.16); `m` carries ×22.59.
4. **Multiply an improvement by its already-measured transfer function BEFORE running.**
5. **Measure the VARIATION of a phenomenon across networks, not its magnitude.** A phenomenon can be large, real, and sealed five ways, and still deliver zero.
6. **Small runs do not adjudicate.** Verdicts need $\geq 24$ networks, submissions $\geq 60$. Our sign flipped six times on small slices.
7. **Test at the working depth (L=32) immediately.** The error saturates rather than accumulating (`~L^0.28`), so at L=2..8 methods are indistinguishable and a fitted `(L/n)^K` law misses by 58×.

## On controls

1. **A foreign-correction control is worth more than an effect size.** Half of our measured gains survived a plausible derivation and died here.
2. **Know its blind spot: it does not catch a scale-only effect.** Neither does LOO, nor A/B cross-validation. All three passed on a branch whose true gain was zero.
3. **A foreign control can be empty by construction** (when the fitted coefficient can flip sign). The working substitute is a **strength scan**: your own correction must peak at `λ > 0`, a foreign one at `λ → 0`.
4. **Fit the coefficient; never nail a correction at weight 1.** We discarded four corrections that actually help. Related: the optimal correction is **shrunk**, not exact — an oracle at `R²=1.0` can be worse than a proxy at `R²=0.6`.
5. **When every member of a family shows its optimum at exactly zero, suspect the grid step, not the gradient.** Ours was 50× too coarse; a parabola through three points recovered the optima.

## On numbers and their provenance

1. **A calibration constant must be MEASURED, not derived by dividing two remembered numbers.** This produced a 4.61× error that three consecutive conclusions stood on.
2. **A number produced inside a model is not a measurement.** Trust the ratio (the bias cancels), not the absolute.
3. **Before quoting a number, open its log and recompute the aggregate.** Three load- bearing figures of this project turned out to be artefacts, one of which appears in no log at all.
4. **Before substituting a quantity into a formula, check WHAT IT DESCRIBES.** Our $\gamma$ table described *projections*, the formula needed *per-neuron* values. Same name, same order of magnitude, different objects — and the substitution returned exactly zero instead of −0.99.
5. **A null correlation is a reason to audit the instrument, not to declare orthogonality.**
6. **When comparing yourself to a leaderboard, verify both numbers are the SAME quantity.** The board shows adjusted score, the submission panel shows raw MSE, and the multiplier between them differs by 100× across entrants. This cost us a load-bearing "×53" that was really ×7.7.
7. **A number taken from outside has an expiry date — date it.** Ours died not from an algebra error but because the organisers re-scored the competition.
8. **Having found a "compared different quantities" error, grep the corpus for the PARTNER NUMBER.** Fixing the one place you found creates an illusion of closure; the same pair survived in three other sections.

## On agreement between numbers

1. **A good agreement between two numbers is a reason to build a control, not a conclusion.** Put the control on the *mechanism of the agreement*.
2. **And ask whether the number could have come out otherwise.** A "0.2% match with our prediction about a rival's method" turned out to be empty: those panel fields are emitted by anyone who does not fill in the optional output.

## On closing branches

1. **Record the CAUSE of death, not the circumstance.** "Closed because expensive" is an invitation — cost is a moving target, a ceiling is not. One correction was resurrected three times purely because its death was recorded as a circumstance.
2. **Grep your notes by the name of the MECHANISM, not the name of the branch.** We returned to Edgeworth three times not because we closed it with a weak number, but because we closed it under the wrong index.
3. **A ceiling has a domain of applicability — state it alongside the ceiling.** "`k < 1.2` never pays" was derived for ideas *with a cost*; at zero cost it does not apply, and a `k = 1.027` idea went on to move the score.

We consider this the most useful section, and we publish it in full because each entry was load-bearing at the time.

| retracted claim | what killed it |
|---|---|
| "$\gamma_4$ only switches on at exact moments" | Intermediate points show it turns on **gradually** (×1.10 → ×2.39 → ×4.50). Our claim came from two endpoints with nothing between them. Consequence: moment accuracy is twice as valuable as we had computed. |
| "Non-uniform sample allocation gives +7% free" | In-sample artefact, killed four independent ways (§II.B). |
| "Richardson extrapolation in width works (×1.22)" | The gain was **pure scale**. Raw ×0.986, cosine ×0.9999. LOO, A/B and the foreign control all passed while the true gain was zero. |
| "RQMC beats Monte-Carlo by 14%" | 24 networks × 8 shifts, paired: ×1.0156, sign test 12/24, median ×0.9897. Noise. |
| "A learned correction gives ×1.031" | Measured against a base holding **stale** constants. Joint re-fitting leaves ×1.011…1.024. Our own rule, violated by us. |
| "`cos = 0.93` after transport implies a large gain" | The backward operator has effective rank 2.74, so any two transported vectors look nearly collinear. `cos` after a low-rank projection measures nothing. |
| "The step error is orthogonal to Edgeworth" | The Edgeworth term had been built from **projection** γ, not per-neuron γ. With per-neuron γ: `cos = −0.9915`, `R² = 0.983`. |
| "Even a perfect two-moment oracle is ×53 from the podium" | Two errors compounding: our *adjusted score* compared against another entry's *final-layer MSE*, using board figures from before the Phase-1 re-scoring (in which one entry's adjusted score moved ×92 at unchanged MSE). Correct: ×7.7 from #2. |
| Relative errors quoted against `E[m²] = 0.858` | Direct measurement: `0.909` (mini), `0.946` (full). ~3% systematic overstatement. |
| "The variance-map gain is ×6.63 / ×44.80" | Really ×1.137 / ×2.771. The overstatement *grew with K* — the signature of shared noise between oracle and target. |
| "Norm spread is 18% and worth exploiting" | `Var(R)/E[R]² = 0.00196`, i.e. `sd(R)/E[R] = 4.4%`, not 18%. **Seal.** Theory for $\chi_{256}$: `0.5/256 = 0.00195`. The norm concentrates far more sharply in 256 dimensions. |
| "The leader's panel matches our prediction to 0.2%" | Those fields are emitted by anyone who leaves `all_layer_means` unfilled. They describe the *task*, not the method. The match could not have failed to occur. |
| "Bias is 46% of MSE" | Direct measurement with a calibrated zero: **7.6%**; ceiling of "bias = 0" is ×1.08, not ×1.85. |
| "Mehler truncation is ×4.14 more accurate than exact moments" | Entirely an artefact of the global scale constant α. Exact quadrature versus the truncated series differs by **0.1%**. |

396 sealed findings. **The filenames are historical titles**, written when each finding was made, so a few carry numbers that were later superseded — "we are 55th" predates the Phase-1 re-scoring that put us at #66, and the retraction table above lists every such case. The filenames are themselves sentences; this is the map of what exists, grouped by what each one does. Anything not listed in a summary is still here.

### A. Ceilings measured (26)

- `bbp refuted direction is easy coefficient is the wall`
- `bias lives on rarely switching neurons probe ceiling 1.6x`
- `capacity law and why q has no pattern`
- `cheapening the sample by rank truncation is dead with a ceiling`

- closure correction ceiling
- cv beta optimal rho is the wall
- ideal two moment closure is only fifth place
- lambda families were capped at 8 layers for a dead reason
- leader sits at the class ceiling layer 29
- linear correction in closure basis has ceiling 1.027
- linear filter class is closed signal not noise is the wall
- local closure formula ceiling is 1.111x
- marginal is solved pairwise is the wall
- maximum is only 9x the mean but 47 sigma away
- perfect step ceiling is 1.19 so 1e 8 is unreachable from the step side
- s2 anchor depth ceiling unreachable
- s2 closure ceiling is 2.4x no far field
- s2 kappa3 ceiling is 1.42x
- s2 perfect closure ceiling is 12x and that caps everything
- sigma branch ceiling is 2.83x target is 6.8x
- sigma injection is pair nongaussianity and cumulants cap at 4x
- sigma residual is structurally unreachable and signs cannot enter it
- two moment ceiling 2.72x and switching series diverges
- two moment ceiling was noise deterministic branch is alive
- variance is the wall branch closed
- weights only features ceiling 1pct

## B. Classes closed (54)

- a3 fit beta revives four dead branches
- virtual weights ensemble has rank 1034 so k linear surrogates are dead
- adjoint rank 8 does not reproduce
- analytic pruning is dead gaussian tail fails
- backward operator is rank 3 excavation is impossible
- budget is per mlp not a common pot allocation is dead
- closed form for relu pair with third cumulants
- closure error does not depend on layer order
- closure error is a clean 1 over n law but its direction is not
- control variates are dead the exactness correlation dilemma
- cost axis midstart dead lambda is a function of n
- cost model v10 does not touch us
- dead neuron compensation breaks lambda
- dead neuron pruning
- dead neuron pruning is already past the true dead count
- early layers are worthless prize is at depth
- edgeworth does not overshoot for relu only for abs t

- edgeworth works mixture and input strat dead
- error recursion is exactly closed but sources are not measurable
- gamma invariant holds but rank collapse explanation fails
- gamma4 works only on exact moments table fails foreign control
- kappa a single scalar on C moves the score at zero cost
- kappa3 is rank one at depth and relu does not inflate it
- kappa3 transport is closed at every price point
- locks real but empty
- low rank nongaussian step is dead error lives in accumulation
- mc variance is closed and b2 is a blend artifact
- mehler higher orders are exactly zero at large t
- micromodel surrogate head fails
- mixing directions are known exactly but spectrum is not a lever
- one over n root is supported locally but does not extrapolate
- partition is not fractal error is between cones
- per network blend weight is empty variance cancels in the ratio
- per neuron lever does not exist error is isotropic
- pilot replacement buys 1.095x and per neuron lambda is dead
- pr is not cp rank both instruments calibrated
- pruning of dead neurons biases
- r2 is not the binding quantity
- rank sketch forward closed
- rao blackwell does not commute with composition
- relu masks carry non moment info but it does not vary across networks
- rqmc is eaten by whitening and jury writeup channel found
- s2 combinatorics lives where error does not
- scalar knobs on transport are empty and closure stand lies
- search never beats ols
- second audit pilot reuse was closed by the wrong gate
- sigma plus edgeworth pair is x114 on stand and dead on grader
- skeleton fails spectrum is answer
- spherical variance reduction is eaten by whitening
- sums game exact only at layer zero
- truth is monte carlo and leader computes only layer zero
- universal kurtosis worthless
- user backprop input search closed
- value is in the tail but error is not

### C. Mechanisms explained (45)

- activation rank collapses but answer lives in the tail
- alpha was miscalibrated because a i was added without refitting

- answer is born early error is made late
- board ranks budget tuning not method and we are 66th
- closure error is accumulation not local form
- covariance error rank2
- cumulant recursion dies on independence
- edgeworth a2 correction dominates the leading term
- edgeworth correction is the live lever
- edgeworth cumulants are the live branch
- edgeworth gain lives in a fragile cancellation
- edgeworth has no zone where it is both useful and convergent
- edgeworth in the chain helps m and ruins the submission
- edgeworth overshoots twofold at depth
- effective rank collapse
- effective rank of C is measurable free but neither target nor proxy
- error is 1d edgeworth gamma3 gamma4 are vectors
- error is born from variance error
- error lives in six directions
- euler stein identity is exact but lives in the tail
- half the ranking set is a sealed private split
- hermite cv variance full rank
- joint kappa3 of the pair explains step error
- kappa3 is low rank at depth
- kappa3 is transported linearly so cp rank carries it cheaply
- kappa3 lives in the low rank subspace
- lambda absorbs any mechanism you fix
- low rank plus diagonal cannot represent C oracle bound
- maxent beats edgeworth
- mc miss lives in three directions not 256
- network freezes with depth only 31 percent of neurons switch
- network is a ray nongaussianity is one scalar
- non gaussianity lives in the spectrum tail
- nongaussianity concentrates but not where accuracy lives
- nongaussianity is 96 percent from dependence and state rank is 3
- phase1 final rank is private reevaluation on fresh seeds
- precision leverage grows and transport has no structural knobs
- relu gates kappa3 and preserves rank exactly
- s2 closure bias is the edgeworth kappa3 term
- s2 natural transport beats any fitted direction
- s2 variance lives between cells not inside
- second order transport is two orders too small
- signal and bias are the same mechanism

- step error is exactly edgeworth with per neuron gammas
- structure lives two orders above target

### D. Positive results (45)

- activity pattern is a live control variate
- aicrowd api gives mse separately from budget
- all layer means is free 40000x and we throw it away
- antithetic halves two matmuls for free
- attractor is a left eigenvector and deep spectrum is real
- bad networks are not predictable from weights
- birth is shallow delivery is uniform
- blend weight signal is real but denominator noise eats it
- blend weight varies 57 percent across networks
- copula not marginals and mixed cumulants are the only live lead
- cosine loss means scale is free per network
- deep anchor beats exact shallow one
- deep channels die against the real submission
- free local correction from closure features
- free second moment channel
- free structure saturates at one percent
- full split has 1000 networks with exact truth
- gam gate argument multiplier is the third free knob
- gamma3 per network is real the across network average hid it
- hard trim threshold kills live neurons
- heavy tail is between networks not within sample
- last layer sigma reopens the class
- leaderboard reality we are 55th and top is 500x ahead
- learned anchor works
- learned mehler makes pair correction free
- learned per neuron correction works where per network constants failed
- learned recursion beats engine
- local dataset and grader testset are different networks
- mu scalar correction works
- narrow deep relu networks die object exists only above width 12
- online calibration signal real lever tiny
- payback gate and kappa21 is real but unaffordable
- prize and anchor live at opposite ends
- remove beats predict across 30 verdicts
- s2 fixed overhead is real but residual scales
- s2 gaussian tilt gives 1.10x free
- sample moments into the closure recursion beats sub97

- spectral blend beats scalar blend on cross validation
- split whitening free gain
- stand said 4.5x real evaluator said 0.61x
- submission only works in a narrow width band
- symmetry tags buy free budget
- tfit free t feature gives 1.056x
- variational closure works but channels are too small
- weave residual beats everything

### E. Instruments & seals (16)

- archive audit of c96 numbers mc is 0.1508 not 0.187
- bank truth noise inflates every mse we measure
- circle probe kinks measured null
- cubature deterministic flat
- first line by line audit of the submitted estimator
- grader budget per submission
- grader charges wall time not flops
- grader two point bias is lower than the leader
- lambda fit is at the optimum on grader
- local closure error was measurement noise
- measuring instrument was the bottleneck
- n sweep on grader refit and depth both hurt
- noise must be measured by redraws not derived from sigma over sqrt n
- partial channel fixes always hurt three instruments
- residual is python time and measurable
- second sample cannot measure what v already knows

### F. Budget, cost, competition (30)

- anchor cost identity
- anchored cv prize and two blockers
- bias floor equals the leader score
- budget is per network and path sampling cancels
- deep collective statistic prize
- depth information equals cost
- fit residual is correlated across layers and that kills the class
- flopscope call tax
- leader method reconstructed m recursion with sigma and shape
- leader pays almost full budget in flops not walltime
- leader runs per neuron cumulants not matrices
- leader spends wall time not flops
- leader submission page decoded

- mc score is budget invariant flat budget theorem
- mehler self residual predicts per neuron bias
- mehler truncation costs nothing
- multilayer cv prize is 4x leader
- no cheap sigma offdiagonal costs 3x
- our advantage over mc melts with budget
- per neuron shape prize is 1.17x but needs bias predictor
- prize map prices any channel before running
- residual is cost per call
- residual jumps at allocator thresholds
- residual phase map
- residual wall is the invisible axis
- rotation prize no eyes
- we are 3.5 percent from the cost floor
- we compared our adjusted score against a rivals raw mse
- whitening deeper costs exactly what it gains
- width 1024 collapse was a fixed budget artifact

### G. Process & methodology (1)

- width extrapolation 1 over n holds inside one network but signal is all scale

### H. Everything else (165)

- accuracy at depth is a cliff not a slope
- all four exactly known truths in the problem are now inventoried
- analytic beta path
- anchor consumes only variance diagonal
- anchor depth sensitivity
- antisymmetric generators orthogonal
- antithetic sampling buys 1.59x through whitening
- arccos cv and anchor drift
- archive session corrections geometric model and sigma bypass
- arcsine family mostly zca shadow
- arcsine gate anchor
- attractor is just a typical sample
- bank arena 800 nets
- bank precision is a label not a method
- bell is mean direction
- bench target is biased along the fitted directions
- bias is 8 percent not 46
- carrying kappa3 through the recursion is worse than not carrying
- client side compute prohibited top10 under review

- closure bias is a universal function of 8 neuron features
- closure bias outruns noise by layer 8
- closure channels cancel and tables only buy scale
- closure error is a pair property learnable to r2 0.666
- closure error is covariance not shape
- closure is a balance point not an approximation
- closure jump at layer two
- closure to score is linear in k and quadratic in error
- coefficients are curves not constants
- conjugate network error conservation
- control variate dies where closure error exceeds sample noise
- correcting C alone makes recursion worse
- corrections are coupled covariance is the blocker
- corrections must be orthogonalized against alpha
- covariance not needed marginal chain
- deep anchor dies on closure error not on lambda
- depth map not contracting
- dual surface form turns 10e700 cells into 8192 kink integrals
- eight traces narrowest point
- elementwise lambda profile is noise
- error birth uniform across depth
- error is 98 percent variance tail tracks scale
- error is additive across all 32 layers
- error is made by relu switching
- error is one scalar plus covariance
- exact formula is nngp plus one over n
- exact sums survive approximations die
- fdt stein covariance
- field shape is an attractor set by weights
- final error is sample curvature
- forum writeups confirm our map
- fourth order has no cheap handle
- fp64 bench cannot arbitrate fp32 lambdas
- full sample gate is a small win
- function is worse than uniformly high dimensional
- gamma4 is a gradual amplifier not an exact moment switch
- gate channel needs q four times more accurate
- gate criticality not an axis
- gates are pinned not coins
- geometry vs frame depth axis
- half gate delta expansion

- half splice anova
- half split knows level not deviation
- hermite spectrum closes cubature
- holographic reference beam dies on draw noise
- hot samples cannot be targeted in 256 dimensions
- identity and correlation at opposite ends
- importance sampling needs the answer itself
- jjk term is the missing 93 percent of kappa3
- kappa3 must enter m and C together solo is negative
- label noise not the binding floor
- ladder path whittling
- lambda basis has no nonlinearity left
- lambda channels buy only 1.28x not 3.7x
- layer alignment is frozen by high dimension
- layer scalars are exactly what alpha already does
- layer0 cv coefficient one
- lowest draw taken as the level twice in one evening
- mean self averages
- mehler series makes closure cheap
- microgroups exponential wall
- micronet learns closure error
- minimax weights lose to ols
- moment estimator is the same estimator
- mu sigma errors cancel
- my own verdicts were half baked three were wrong
- nact channel survives but selffit eats 84 percent
- negative weights create dimension and variance drives determinism
- no one dimensional shadow
- noise is one scalar
- numerical density integration buys shape not scale
- official dataset with exact truth was on disk unused
- only the mask family stays clean at depth
- optimal quantization collapses in 256 dims
- orbit neighbour transform
- order four is the first uncontrolled moment of z
- our control variate only uses 11 percent
- output second moment channel pays exactly its price
- pair non gaussianity is third order not fourth
- path to 1e 8 requires exact moments plus gamma4
- per network personalization needs an unobservable projection
- per network scale knob is worth under 1.09x

- per neuron moments suffice but cannot be obtained
- pilot cut is net beneficial not a floor
- pilot samples are computed and thrown away
- pointwise step functions are exhausted but gam varies per network
- poisson reduces to surface
- preactivation is heavy tailed but shape is per neuron
- qmc is fully absorbed by whitening
- qmc sobol loses to iid
- radial homogeneity vr
- radial stratification null
- recursion as search input
- running mean path carries no information
- s2 beta weights derived not fitted
- s2 better 1d basis buys nothing
- s2 caustic concentrates both errors
- s2 cheap surrogates decorrelate
- s2 closure error is a 1 over N finite size effect
- s2 closure errors cancel fixing one piece hurts
- s2 collapse is lyapunov
- s2 degree spectrum bounds the whole exact family
- s2 error is 93 percent variance sketch dominated by topk
- s2 error removal at output 142x
- s2 high degree tail is relu kinks
- s2 known mean and correlation are anticorrelated
- s2 per element beta is worth 1.8x blocker is predicting r
- s2 signal and noise share the same recursion
- s2 widths fast per object slow per class
- same sample tautology theorem
- sample correlations at 29 carry no sign
- sample derived closure cannot beat the mc it is built from
- sample derived correction cannot beat the sample
- sample game fully mapped
- sampling is exhausted 65k samples give nothing
- scale noise is predictable but only within a network
- self iteration has one fixed point no cycles
- shadow equals antithetic
- shape is blind to location
- shrinkage weight is computable per network
- sigma error is a geometric sum of offdiagonal step error
- sigma is the whole problem and its requirement map
- signal abundant wall is fit

- single neuron marginal closure
- smoothness homotopy bias falls 31x but delta carries everything
- stale artifact had a different sigma channel
- stein control variates are orthogonal
- sub91 is worse sub92 blend 0883 is better
- sub93 cleans q with same pass cumulants
- submitted estimator forbids numpy and os
- superposition dimension is high univariate part is linear
- synthetic deep start
- tail quadrature needs 64 dims
- the mean channel carries everything not the off diagonal
- the q channel is poisoned by closure bias
- trimming bias is computable but too expensive
- turbine shaft error propagates with r2 0.94
- two estimators on one draw carry no direction
- two state decomposition is shrinkage of the product
- v2 error is mc noise not bias
- v2 exact through third moment
- v2 is exact order two
- v2 synergy decomposition
- whitening carries depth antithetic dies
- width 1024 needs the opposite estimator
- within layer errors are structured

The submitted estimator is documented separately — flow, constants with provenance, FLOP budget, and known defects. Measurements above were produced either by the official grader (with the blend knob set to isolate a branch) or by stands run against the public `arc-whestbench-public-2026` dataset, whose `final_means` are exact at $N = 10^9$. Ground-truth stands use disjoint seeds for oracle and target where both are sampled.

Sandbox constraints worth stating, because they shaped what was possible: the estimator may import only `flopscope` and the standard library — no NumPy, no SciPy, no Torch — so a $256^3$ matmul is billed in full (3.355e7 FLOPs), `eigh` costs 1.509e8, `svd` 4.36e8, and our entire analytic closure costs 1.66e9, which is 6.1% of the budget floor.

*Numbers we could not seal are marked as derived. Where a measurement contradicted an earlier claim of ours, both appear.*