Where the error of a factored cumulant chain lives at width 1024

Submission ID: 335513 (ARC White-Box Estimation Challenge 2026, Phase 2)

ARC White-Box Estimation Challenge 2026, Phase 2 (1024 × 16 ReLU MLPs) · Note for the algorithmic-contribution prize · Graded submission: **AIcrowd #335513** (internal name e54957), score 3.202e-9 · 18 October 2026

> *At width 1024 the factored third-cumulant chain gets the third cumulant right in level; the remaining error sits at fourth order and beyond, where no affordable physical closure beats the fitted one — only the exact, complete closure does, by 4–16 % on earlier chains and 4 networks.*

Reading conventions. `[R]` marks a figure re-measured for this note on the chain of #335513, over 24 public networks at n = 1024 (20 for the third-cumulant comparison), given as mean ± standard deviation across networks. The bench replays the 12 October bench pass of #335513 (numpy, without Strassen products), not the graded package; it reproduces that pass's MSE exactly on the 24 networks (mean 1.9619e-8). The sampling floor of the ground truth is 0.39 % of the final MSE (0.30–0.46 %). Source: `note-prix/verification/resultats/verification-chiffres*.txt`. (§163) points to a section of our research notebook (`WRITEUP.md`, 25 Aug–26 Sep 2026), cited only if nothing later contradicts it; those figures come from earlier versions of the chain, which is said each time. "MSE" is the final-layer mean-squared error against Monte-Carlo truth; bench MSE excludes the crest (worth 0.9 %, applied offline); the score is MSE × max(0.1, C/B), B = 2^{41} billed flops.

Summary

Our graded estimator is the organisers' factorised third-order cumulant chain, as made affordable by team 504aldo, with a fitted ("regenerated") fourth cumulant and our own additions. Measured on that estimator:

1. **κ_3 is right in level.** The (2,1) slice of the chain's third cumulant matches the truth to better than 1 % in level; its remaining incoherent error is about 7 % rms (5–9 %, growing with depth) `[R]`. At n = 256 the same construction misses a coherent 10–12 % (§136).
2. **At fourth order, the fitted closure beats every affordable physical one.** Without any κ_4 the error is ×7.5; replacing the fitted closure by the measured physical structure of κ_4, a scale mode, costs +42 % even with the true level, and still +29 % when the learned correctors that could favour the fitted closure are removed on both sides `[R]`. Only the exact, complete fourth-order core, which we could not afford, did better: −4 % and −16 % on earlier chains (§152, §124). Figure 1 shows the whole picture.
3. **What each piece is worth**, in a full ablation of #335513 (Table 4).

The third cumulant is not the main cause of the remaining error, but it is not free of it either: keeping all old sources dense or using rank-16 thin legs, both on the κ_3 side, would each cut the error by about 5 %. An appendix adds that verdicts on modifications do not transport across widths. We also report a central figure of our own notebook that the re-measurement overturned, and how the work was done: largely by AI agents (Claude from Anthropic, Codex from OpenAI) directed by one human participant.

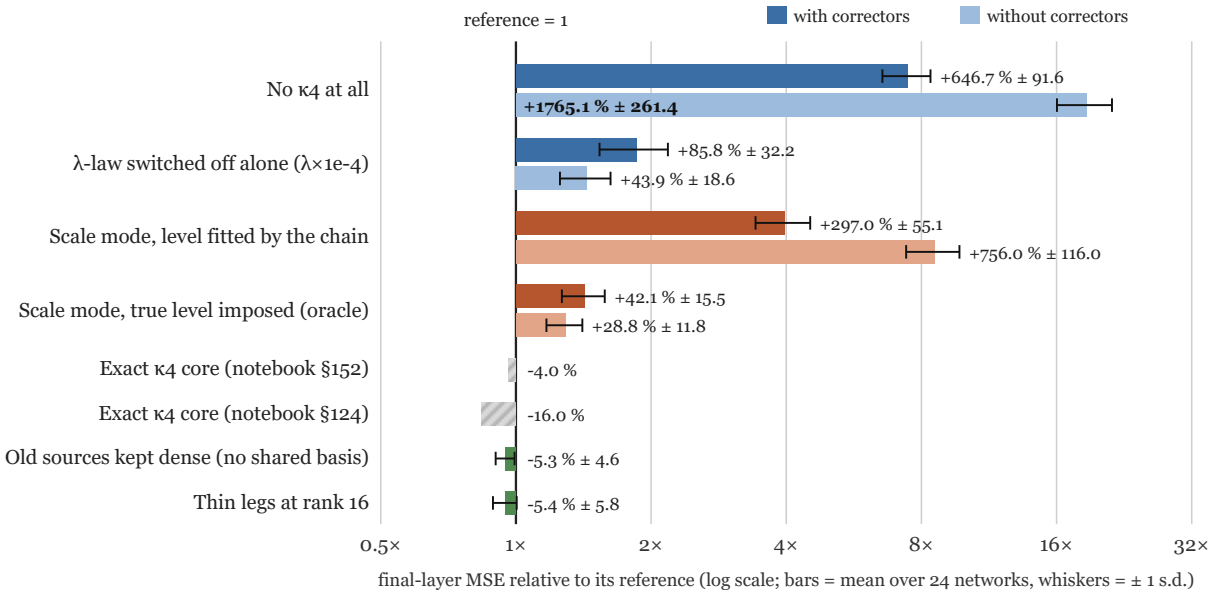


Figure 1. What the fourth-order closure and two κ_3-side changes do to the error (24 networks, `[R]`). Blue: no or reduced κ_4; orange: the physical scale mode in place of the fitted closure; dark bars: relative to #335513 with its learned correctors; light bars: the same change with the learned mean correctors removed on both sides, relative to the chain without correctors (section 4.3). Green: κ_3-side changes. Hatched grey: exact κ_4 core, from the notebook on another chain and 4 networks, no spread available.

1. The problem and submission #335513

Each test case is a ReLU MLP of width n = 1024 and depth L = 16 with frozen random weights and Gaussian input; the estimator predicts the expected post-activation of every neuron at every layer, under a hard budget of B = 2^{41} metered flops (flopscope), one core, 120 s of wall time and 0.4 s of non-metered "residual" time per network, 5 s of setup and 8 GB of address space (rules as recorded in 504aldo's community post and in our notebook §97).

Table 1. The graded submission. Official values from the AIcrowd page of #335513 (`boucle/notes-officielles/e54957-335513.json`) unless marked.

quantity	value	origin
score	3.202e-9	official
billed compute C/B	0.16836	official
implied final-layer MSE (score / C/B)	1.902e-8	derived
networks graded / failures	100 / 0	official
residual time per network: min / mean / max	0.279 / 0.306 / 0.394 s	official
wall time per network: mean / max	81.2 / 93.8 s	official
billed flops per network	≈ 3.70e11 (≈ 172 units of $2n^3$, ≈ 10.8 per layer)	derived: 0.16836 × 2^{41}
bench MSE, 100 public networks / the 24 re-measured	1.9362e-8 / 1.9619e-8	package record; `[R]`
score predicted before deposit / measured	3.2059e-9 / 3.202e-9	`CONTRADICTION-CHIFFRES-17-10.html`, row 12
score noise between two deposits of the same package	0.02 %	same file, row 13

For scale: the best graded entry we recorded, J2W #333893, scored 1.549e-9. We are roughly a factor 2 behind the leaders. This note does not claim a winning estimator; it locates, by measurement, where the error of this family of estimators sits.

2. The estimator, exactly

2.1 Lineage and what is ours

The backbone is the factorised K = 3 cumulant propagation of Wu, Lecomte, Winer, Robinson, Hilton and Christiano (arXiv:2605.05179, supplement S.4.3) with its exact power-Hermite Wick coefficients, ported from the organisers' reference code to flopscope (parity 1e-15, §188). The third cumulant is kept as a sum of sources, one born per ReLU layer and transported by the weights. Team 504aldo (public MIT code, write-up and findings log, September 2026) made it affordable and published the pieces we reuse: the exact $7 \to 4$ restructure with two dense legs per source (F63), the memoryless κ_4 regeneration $G_{off} = \lambda_l C_{off}$ with the diagonal kept exact (F68), the D21 feedback legs (F69), the trimmed last layer (F71), the age-gated shared basis for old sources (rank 384 from age 4, nested rank 224 from age 7, F72–F73), and Strassen–Winograd block products with pooled buffers (F79–F81). Our additions are listed in Table 3. The package is one assembled file `estimator.py` (14,720 lines, most of them inert research switches) plus `table-N.npz`; we describe only what is active.

2.2 State carried from layer to layer

- the mean μ and the full covariance C of the post-activations (dense n×n);
- the third cumulant as live sources: young sources (age < 4) with two dense legs A, P (n×n) and thin legs (rank 2, rank 4 for sources born at layers 11–14); older sources confined to a shared tier-1 basis (rank 384) and, from age 7, a nested tier-2 basis (rank 224, read at an interior rank of 160);
- for the fourth cumulant, only a diagonal vector g and the previous variance (no tensor memory).

2.3 Order of computation at layer l

1. **Linear step.** $\mu_{pre} = W\mu$; $X = WC$ over the living columns of the previous layer; var = rowsum(X ∘ W); $t = \mu/\sigma$. Neurons with $\Phi(t) < 10^{-3}$ are declared dead and moved last; the living count is rounded up to a multiple of 64 (keeps Strassen leaves even, §160); among the living, "linear" neurons ($t > 3$) are moved last. $C_{pre} = XW^T$ is formed as a symmetric triangle of blocks (at layer 0 as the Gram matrix WW^T). At the last layer only the variance is formed.
2. **Fitted κ_4 diagonal** (504aldo's recipe, our shape mix):

   $$dG = (W \circ W)\, g_{prev} + \lambda_l\,(\mathrm{var} - (W \circ W)\,\mathrm{var}_{prev}), \quad dG \leftarrow 0.6\, dG + 0.4\, \Pi_{\mathrm{var}^2}(dG)$$

   where Π_{var^2} is the least-squares projection on the vector var² (§140, §143).
3. **Transport of the third-cumulant sources** by $WD = W\,\mathrm{diag}(w_{1,prev})$: $A \leftarrow WD\,A$, $P \leftarrow WD\,P$ for the young sources (batched, Strassen–Winograd, 5 levels); a source reaching age 4 joins tier 1 (at layers 7 and ≥ 9 without a new range finder: the source is projected on the existing basis); the oldest tier-1 source moves to tier 2 at age 7; old legs are re-formed from the basis each layer, because every read is a Hadamard product of legs (504aldo's "wall"). From layer 12 on, joins and moves are skipped.
4. **Newborn insertion.** The source born at layer l−1 enters with $A = W\,a_b$, $P = W$ (dead hubs zeroed).
5. **Wick coefficients and reads.** w(k,p) at (μ, var) from one unified table; D3 (diagonal of κ_3) and D21 (its (2,1) slice) read from all sources: two dense contractions $LA\,A^T + LP\,P^T$ per young source, the tier bases for the old ones. Rows of D21 of linear neurons are not formed; their contribution to the (2,2) sums is taken in the $t \to \infty$ limit.
6. **κ_4 slices**, with m = 2:

   $$g_4 = m\, dG, \quad g_{22,ij} = (m/6)(dG_i + dG_j), \quad g_{31,ij} = (m/2)\,\lambda_l\, C_{off,ij} \cdot \mathrm{var}_i / \langle \mathrm{var} \rangle$$

7. **Last layer only.** E[ReLU(z)] by exact integration against a three-moment law fitted to (μ, var, κ_3) instead of the Edgeworth series (§187), plus the κ_4 term $w(4,1)g_4/24$; the source reading is restricted to neurons with $|t| \le 3.1$; then the mean correctors (step 10) and a learned quadratic "crest" corrector clipped to $\pm 10^{-3}$ per neuron. Output: the 16 layer means.

8. **Other layers: post-activation moments** from the dense Wick term table applied to the slices {C_{off}, D3, D21, g_4, g_{22}, g_{31}}, converted to cumulants (mean, variance, covariance and the two-index slices the births need). The non-Gaussian part of the second moment enters the variance at 0.9 only (idea of §4).
9. **Birth of the new source**: birth leg a_b and P; feedback and residual thin legs by a randomised range finder (two passes); the $\kappa_4 \to \kappa_3$ feed carried as column weights ($c_1 = \lambda w_1$, $c_2 = w_1^2 dG$, $y = (m/4)w_2$), dosed at 0.85.
10. **Mean correctors, injected.** The post-activation mean receives three linear corrections on features of the local state (A2: 10 features, layers 0–15; M1: 30 features including row sums of C_{off}, D21, g_{31}, g_{22}, layers 1–14; Q2: 165 features with regime interactions, layers 1–15) and the covariance a two-term correction (Q2, layers 5–14). Their coefficients were fitted offline by DAgger, i.e. on the chain's own corrected trajectory (`chantier-dagger*/VERDICTS.txt`).
11. $\mu \leftarrow$ corrected mean; $C \leftarrow$ corrected covariance with the corrected variance on its diagonal; next layer.

A Gaussian mean/covariance fallback replaces the chain if it raises or returns a non-finite value, or if a shape would exceed 60 % of the budget. Setup allocates about 1.5 GB of persistent buffers and replays, without computing, a recorded trace of the first network (shapes, views, Strassen plans) so that the first graded network does not pay for it in residual time; Python's cyclic garbage collector is disabled.

2.4 What is fitted

Table 2. Constants of #335513 that are not exact consequences of the model (read from the package).

constant	value	how it was set
λ table (16 values)	5.24e-3 … 1.14e-2 … 5.87e-3	504aldo's published table × 1.05 on layers 1–7 and × 0.809 on layers 8–15 (index shifted by one; computed by comparing the two files)
κ_4 diagonal shape mix; (3,1) var weighting	0.6; on	100-network sweeps, §143–144
$\kappa_4 \to \kappa_3$ feed dose; variance part	0.85; 0.9	research loop, from 26 Sep
dead / linear / last-layer thresholds	$\Phi(t) < 10^{-3}$; $t > 3.0$; $\lvert t \rvert \le 3.1$	sweeps at 1024 (§156 for the dead threshold)
ranks	384 / 224 / 160 (tiers); 2, 4 (thin legs)	504aldo's age law for the tiers; ours for the rest
mean correctors A2, M1, Q2	156 + 420 + 2,384 non-zero coefficients	DAgger on public networks 16–99 (84), judged on 0–15
crest (quadratic ridge on 29 final-state features)	464 coefficients + intercept, clip 10^{-3}	fitted on 100 public networks
read-out table in (t, σ)	present, factor 0 (inactive)	—

About 3,400 fitted coefficients in all, about 2,960 of them in the injected mean correctors. The Wick coefficients, the transport and the readings are exact consequences of the model.

2.5 Additions relative to the published lineage

Table 3. What #335513 adds to the organisers' recipe and to 504aldo's V24–V29. “First evidence” is historical; Table 4 gives what the pieces are worth in #335513.

addition	what it does	first evidence
diagonal shape mix 0.6, (3,1) weighted by var_i	moves the fitted κ_4 toward the scale-mode shape at constant level	−3.3 % on 100 networks (§144)
λ reduced on late layers	the late core was over-fed	§167, §176
thin legs at rank 2–4, two finder passes	cheaper feedback and residual legs	§165, §168, research loop
dead-neuron pruning, living count rounded to 64; linear rows in the limit	products restricted to living rows; D21 rows of linear neurons not formed	−14 % flops for +1.5 % MSE (§100, earlier chain)
trimmed end: no joins after layer 12, last-layer read restricted, in-place joins	removes work whose effect on the output is small	research loop (campaign D)
three-moment exact read of the last layer	replaces the truncated Edgeworth term	−0.63 % on 4/4 networks (§187)
DAgger mean correctors (injected)	law-learned corrections of mean and covariance at every layer	M1 −4.6 % (16/16), then Q2 −9.5 % (16/16) on networks 0–15 (`VERDICTS.txt`)
crest read-out corrector	last-layer quadratic correction, clipped	−0.97 % (§144, earlier chain)
engineering: symmetric half-products, paired Strassen engine, twin products, persistent buffers, setup trace replay	lower bill and residual time at identical predictions	research loop, 28 Sep–11 Oct

2.6 Cost

One unit is $2n^3 = 2^{31}$ flops; B is 1024 units. #335513 bills about 172 units per network, about 10.8 per layer. Structurally, a young source costs two transports and two reads, $4n^3$ (2 units) per layer at dense price; an old source costs $8n^2r$ per layer (1.5 units at $r = 3n/8$, 0.88 at $r = 7n/32$) for forming and reading its legs (§147); five Strassen–Winograd levels bill $(7/8)^5 \approx 0.513$ of a dense product (§138). The last cost map measured on this lineage (24 Sep) put the shared basis at 24.5 % of the bill and the thin legs at 11.9 % (§174). We have **not** measured the per-component bill of #335513: the re-measurement bench runs without Strassen products and does not reproduce the graded bill, so Table 4 reports error only.

3. The third cumulant is right in level (result 1)

The truth for the slices comes from 400,000 × 2 samples per network (two independent seeds; layers 6–15, pre-activation means taken exact from the official 10^9-sample layer means), whose relative noise on D21 is 7–13 % depending on the layer. Against it, on the living block of 20 networks `[R]`:

Table 5. Third-cumulant (2,1) slice of #335513 against the truth, by layer (20 networks; mean over networks). The last column is the chain's own incoherent error implied by the gap to the ceiling, $\sqrt{((\text{ceiling}/\text{corr})^2-1)}$, computed by us from the same file.

layer	truth noise (rel.)	slope chain/truth	slope truth half A/half B	correlation	ceiling of an exact chain	implied chain error (rms)
6	13.1 %	0.981	0.975	0.9900	0.9914	5.3 %
8	10.7 %	0.985	0.983	0.9923	0.9943	6.4 %
10	9.2 %	0.986	0.989	0.9930	0.9957	7.4 %
12	8.1 %	0.988	0.993	0.9935	0.9967	8.1 %
14	7.3 %	0.988	0.992	0.9936	0.9974	8.7 %
6–14 (all)	—	0.986 ± 0.007	—	0.9926 ± 0.0012	0.9952	—

The slopes are attenuated by the truth's noise to about the same degree as the slope between the two independent halves of the truth, so the level is right to better than 1 %. The correlation stays below the noise ceiling at every layer: the chain's own error is incoherent, about 7 % rms, rising from 5 % at layer 6 to 9 % at layer 14. The diagonal D3 behaves alike (slope 0.987, correlation 0.9954). An earlier one-network measurement agrees (§144). This is width-specific: at n = 256, even the exact K = 3 chain misses a coherent 10–12 % of D21 at depth (§136). Slices must be compared on the living block; an apparent deficit at 1024 came from the rows of dead neurons, which the chain zeroes (§136).

4. At fourth order, the fitted closure beats every affordable physical one (result 2)

4.1 The true fourth cumulant is close to a scale mode

On sampled truth at n = 128 and 256 the matrix $\kappa_4[aabb]$ is 88–97 % rank one in Frobenius energy at every layer, and $\kappa_4[iijj]/(\mathrm{var}_i \mathrm{var}_j)$ is the same on every pair and grows with depth (§131): the signature of a scale mixture $z = s\,\tilde{z}$, which ReLU's positive homogeneity transports unchanged:

$$\kappa_{4,abcd} \approx k_s\,(C_{ab}C_{cd} + C_{ac}C_{bd} + C_{ad}C_{bc}), \quad \text{so } \kappa_{4,iiii} = 3k_s \mathrm{var}_i^2, \quad \kappa_{4,iiij} = 3k_s \mathrm{var}_i C_{ij}.$$

One scalar explains the three repeated slices to about ten percent at n = 128 and 256 (§131.6). At n = 1024, on network 0, the true κ_4 diagonal follows the σ^4 law with R^2 0.93–0.96 from layer 4 on, with k_s rising from 0.0032 (layer 1) to 0.0227 (layer 15) (§134). For comparison, the weight-averaged kurtosis of Noci et al. (2021, arXiv:2106.06615), $\gamma_2 = 0.227$ at n = 1024, L = 16 (our survey, `RAPPORT-ETAT-DE-L-ART.md` §3.2), corresponds, by our own conversion, to $k_s \approx 0.076$: for a scale mixture $\kappa_{4,iiii} = 3k_s \mathrm{var}_i^2$, so the excess kurtosis is $\gamma_2 = 3k_s$ and $k_s = \gamma_2/3 = 0.227/3$. That is about three times the value we measure at fixed weights (one network); the comparison is ours, not the authors'. Noci's kurtosis is over weights at a fixed input; ours is over inputs at fixed weights, on one network: an order-of-magnitude comparison, not a test. A non-scale remainder, about 30 % of the (2,2) slice's rms, is real (an independent draw reproduces it at correlation 0.98–1.00) and is mostly the accumulated memory of the $\kappa_3 \to \kappa_4$ feed (§148).

4.2 What the chain needs

Table 6. The fourth-order closure of #335513 replaced or removed (24 networks; error change, mean ± s.d.; networks worse out of 24) [R], and the exact core from the notebook.

variant	Δ MSE	worse
no κ_4 at all (SANS_REGEN: fitted diagonal g, off-diagonal law and $\kappa_4 \to \kappa_3$ feed all removed)	+646.7 ± 91.6 % (×7.5)	24
λ-law switched off alone ($\lambda \times 10^{-4}$; the diagonal transport stays)	+85.8 ± 32.2 % (×1.86)	24
scale mode in place of the fitted closure, level fitted by the chain	+297.0 ± 55.1 %	24
scale mode, true per-layer level imposed (oracle, 65,536 samples per network)	+42.1 ± 15.5 %	24
λ × 0.80 / 0.90 / 1.05 / 1.10 / 1.15 / 1.25	+5.07 / +2.70 / −0.77 / −0.75 / +0.13 / +0.69 %	22, 21, 8, 9, 11, 13
exact κ_4 core with five-leg births, thin legs 64, dense, no basis, no pruning (notebook, 19 Sep chain, networks 0–3)	−16.0 %	—
exact κ_4 core + five legs (rank 128) on the research chain (ChaineP) with the 21 Sep settings (job 246), networks 0–3, for +18 % of bill	−4 %	—

- **Opposite depth profiles.** At n = 448 the fitted (4,) slice decays with depth (rms 0.041 → 0.0038) while the physical mode grows ($3k_s \mathrm{var}^2 \approx 0.006 \to 0.03$) (§134).

- **The physical structure costs; the exact one gains.** The scale mode with its true level is 42 % worse than the fitted closure on every one of the 24 networks. On earlier chains, bringing the read-out κ_4 diagonal to its true level cost +24 % (§144), and the fitted (3,1) slice, at 29–34 % of the truth from layer 8 on, could not be raised alone (+8 % at ×1.5, +119 % at ×3, §145–146). But the exact, complete core gains: −16 % (§124) and −4 % (§152); the exact κ_4 store gave 2.108e-8 against 2.298e-8 on network 0 (§115); and at n = 256 the three true κ_4 slices fed to the exact research chain gave −54 % (§136, §148). Its cost (dense five-leg births; +18 % of bill in the cheapest form we built, §152) was beyond our budget.
- **λ is on a flat optimum.** Scalings from 1.05 to 1.25 stay within about 1 %, with large network-to-network spread; × 0.90 costs +2.70 % (21/24) and × 0.80 +5.07 % (22/24). The −2.0 % that × 0.80 gave on an earlier chain (§167) is already inside #335513, which carries a factor of about 0.8 on layers 8–15 (Table 2).

4.3 A bias in these comparisons, measured

In every row of Table 6, the mean correctors (about 2,960 coefficients, whose inputs include the fitted g_{22} and g_{31} slices) and the λ table stay active, and both were fitted on the chain with the fitted closure. To measure that advantage we removed the three mean correctors on both sides (the bench has no crest; λ stays at its table, since the fitted closure *is* λC_{off}). The chain without correctors has a mean MSE of 2.26e-8, +15.68 ± 6.79 % above #335513. Against that chain, 24 networks `[R]` (`verification-chiffres-V6.txt`):

Table 6b. Fourth-order variants without the learned mean correctors, relative to the fitted closure without correctors.

variant	Δ MSE without correctors	range	worse	with correctors (Table 6)
scale mode, true level imposed	+28.8 ± 11.8 %	+9.4 to +55.2 %	24/24	+42.1 %
scale mode, level fitted by the chain	+756 ± 116 %	+544 to +1,038 %	24/24	+297.0 %
no κ_4 at all	+1,765 ± 261 %	+1,359 to +2,605 %	24/24	+646.7 %
λ-law switched off alone ($\lambda \times 10^{-4}$)	+43.9 ± 18.6 %	+21.2 to +117.0 %	24/24	+85.8 %

The correctors did favour the fitted closure against the true-level scale mode (+42 % becomes +29 %) and against switching the λ-law off (+86 % becomes +44 %), but modestly; without them the fitted closure matters even more against the two cruder alternatives. The conclusion holds: no physical closure we could afford beat the fitted one, and only the exact, complete one did. The true level was obtained from 65,536 samples per network (two halves of 32,768), the per-layer k_s being the least-squares fit of the sampled (4,), (2,2) and (3,1) slices on the scale form, averaged over the two halves.

Interpretation. The covariance update consumes the (2,1) slice of κ_3, the (2,2) and (3,1) slices of κ_4, and the truncated old sources. With κ_3 right in level, the fitted fourth-order closure is what keeps the variance on track: it is worth a factor 7.5 in MSE. We read it as a fitted closure that carries part of the scale mode through its (3,1) slice and its feed leg (§133) and absorbs what the chain lacks elsewhere (§137), not as a model of the network's fourth cumulant.

5. What each piece of #335513 is worth (result 3)

Table 4. Ablations of #335513, one change at a time (n = 1024, 24 networks; error change, mean ± s.d. across networks) `[R]`. Cost changes were not measured (section 2.6); improvements are not known to add up.

component removed or changed	Δ MSE	networks worse / 24
thin feedback and residual legs removed	+850.5 ± 163.8 %	24
source born at layer 0 suppressed (NAISSANCE_DES = 2: no source enters before layer 2)	+825.2 ± 259.6 %	24
no κ_4 at all	+646.7 ± 91.6 %	24
λ-law switched off alone	+85.8 ± 32.2 %	24

shared basis at rank 16	+2,078.9 ± 400.3 %	24
shared basis at rank 128	+296.4 ± 44.6 %	24
old sources kept dense (no shared basis)	−5.29 ± 4.61 %	1
thin legs at rank 16 instead of 2–4	−5.37 ± 5.77 %	0
all three mean correctors removed	+15.68 ± 6.79 %	23
Q2 alone removed / M1 alone / A2 alone	+9.53 / +4.64 / +0.81 %	23 / 24 / 18
crest read-out corrector removed (in sample: fitted on these networks)	+0.92 ± 0.76 %	21
five-leg births (their own legs at rank 4) in place of the two-leg (“A”) births	+1.46 ± 7.56 %	16
five-leg births (legs rank 16) with thin legs 16, against thin legs 16 alone	−3.56 against −5.39 %	7 against 0
no-op controls of the two five-leg runs	−0.00 / −0.01 %	10 / 9

The five-leg births are refused by assertions on the deposit path (living-hub axis, grouped leg writing), so they were measured off that path, against paired controls that reproduce #335513 to 0.01 %: their effect is null within the spread. One variant could not be measured: dropping the old sources entirely (the SANS_BASE switch is broken in the current chain). Reading the table: the thin legs and the fitted κ_4 are what make the chain accurate; the earliest source is indispensable; the shared basis is not a free compression — keeping all old sources dense would cut the error by about 5 %, and the basis trades that 5 % for compute; rank-16 thin legs would also cut it by about 5 %, at a cost we did not measure. The injected mean correctors are worth about 16 % of MSE, the read-out corrector less than 1 %.

6. Limits, and an internal figure the re-measurement overturned

- **An overturned figure.** Our notebook (§163) claimed that the chain creates 6.5–7.4 times more error than survives at the output, with an inherited error uncorrelated with the total (+0.00), and built a “compensation” reading of the chain on it. Re-measured on #335513 over 24 networks, the ratio is 1.75 ± 0.15 (1.49 ± 0.13 once the truth's noise is removed) and the correlation is 0.800 ± 0.006; at the last layer the inherited part carries 78 % of the MSE and the generated part 21 % `[R]`. A control that deliberately misaligns the neurons reproduces the old signature (correlation 0.005, ratio 10.4), which points to a misalignment in the original script. We dropped that result.
- **Other internal figures that the site contradicted.** An internal “cross-judge” described as measuring 100 networks measured 8, confirmed on 24; a setting predicted at −0.97 % of score measured +0.25 % when deposited (#335716); our residual-time predictions erred by 3–5 times the safety margin (`CONTRADICTION-CHIFFRES-17-10.html`). The target “ε(D21) 19 % → 5 %” of §65 was measured on an older chain and is superseded by §188.
- **Which chain, how many networks.** Tables 4, 5 and 6 (upper part) are measured on the bench replay of #335513 over 24 public networks (20 for Table 5). The exact-core figures (−4 %, −16 %) and the mechanistic κ_4 figures (§131–148) rest on 1 to 4 networks and earlier chains.
- **Truth.** The official truth uses about 10^9 samples; our slice truths use 400,000 × 2 samples (D21 noise 7–13 %), the scale-mode oracle 65,536 (2 × 32,768), and our earlier κ_4 truths at 1024 used 300k × 2 draws with relative noise 12–16 % on K22 and 27–30 % on K31 at depth (§145).
- **Fitting and selection.** The DAgger correctors were fitted on public networks 16–99 and the crest on all 100 public networks, so their rows in Table 4 are partly or wholly in sample, and they bias Table 6 (section 4.3). According to §126 the graded networks are not in the public dataset, so the official score is out of sample for these fits, but our choice of what to deposit used many graded scores. A private re-evaluation on sealed networks was announced (§107); we do not know its outcome.
- **Costs not measured.** Table 4 gives error only.
- **Not first.** We did not reach the leaders (1.549e-9 against 3.202e-9).

7. How this work was done

Most of this work was done by AI agents, directed by one human participant. Claude models from Anthropic (mainly Claude Opus, at times Sonnet; as coders, explorers and research strategist) wrote nearly all the code, including the 14,720-line estimator, designed and ran the measurements, kept the research notebook and drafted this note. From late September, agents running on OpenAI's Codex took part in a research loop that proposed, built and judged variants (a tree of research “ways”, a builder, a bench and a cross-judge on a rented 12-core, 32 GB server; §176–190). The human participant set the objectives and constraints, decided budgets and when to stop, approved every deposit, and repeatedly asked for re-checks.

For this note, the re-measurements, a contradictor's review of the draft and the novelty audit were also done by AI agents, separate from those who produced the original results; there was no independent human check of the figures. Three consequences matter to the reader. (i) Speed came with mistakes: wrong arithmetic (§30), a rank closed on the wrong clock (§49), settings silently ignored by the deposit chain (§174), instruments that misjudged time, and a central figure of our own notebook that did not survive re-measurement (section 6). (ii) Many variants were tried (more than 350 numbered jobs by 24 September, §173, then the loop's own trials), so any single gain under 1 % should be treated as unconfirmed. (iii) Where the re-measurement or the contradictor disagreed with the notebook, we kept the re-measurement and the narrower claim.

8. Novelty audit

We compared each result with 504aldo's write-up, findings log (97 entries) and claims; our state-of-the-art files (`white-box/etat-de-l-art/`); the organisers' paper as summarised in §60, §65 and in the width debate files; and the public notes and pages read in §60, §61, §65, §80, §92, §126. Phase 1 forum posts are cited from our survey files, which verified them; we did not re-read them.

Table 7. Verdict per result. **new** = not found in the material above; **known** = present there, credited; **uncertain** = close precursor or incomplete search.

result	verdict	closest prior work and credit
The factorised chain's κ_3 is right in level at 1024 (incoherent error ≈ 7 %), with a coherent deficit at 256	**new**	504aldo: feeding the exact D3 and D21 reproduces the exact chain (F65), and the $\varepsilon(D21)^2$ law (F71); no comparison with sampled truth at 1024 found
$G_{off} = \lambda_l C_{off}$, diagonal kept exact	**known**	504aldo F68
The true κ_4 is close to a scale mode k_s Sym(C⊗C), growing with depth	**known** in kind	a shared log-normal scale factor with kurtosis growing with L/n: Hanin & Nica 2019, Noci et al. 2021, Li, Nica & Roy 2022 (annealed over weights; our survey §3.2); Puffi's Phase 1 closure from covariance-response modes (§60); 504aldo F78 (a scale-mixture closure of the covariance chain, dead). New: the measurement at fixed weights (k_s = 0.0227 at layer 15 against ≈ 0.076 implied by the annealed kurtosis)
No affordable physical κ_4 closure beats the fitted one (+42 % with the true level); the exact complete one does (−4 to −16 %)	**new**	504aldo reports the regeneration as a projection of the augmented chain's core (F68) and the dense augmented reference as slightly better; not compared with the true κ_4
Full ablation of a frontier-class chain at 24 networks	**new** (figures)	504aldo's version ladder on 8 networks (V16b–V29)
Isolated truer pieces degrade a fitted chain (Annex C)	**known** in kind	Phase 1: jamesrahenry (topic 18157), trim_qewas (18182), amalgonim (18151), baltsat (18149); 504aldo F68/F69
Verdicts on modifications invert across width (Annex B)	**known** in kind	Wu et al. 2026 §6.3 (“increasing K can worsen performance at small values of n”); 504aldo F28–F29 (the expansion diverges outside its domain; optimal truncation depends on n/L). Our quantification (rank agreement) is new
Laws in n^{-p} and the role of L/n	**known**	Wu et al. 2026 ($c_K(L/n)^K$), Hanin 2024, Roberts–Yaida–Hanin 2022

Collective deformation; the independent prediction has the wrong trend with depth (§2)	**uncertain**	annealed finite-width theory (odd cumulants vanish); ARC's presumption-of-independence agenda frames it
Rank collapse of the propagator; no n^2 closure of D21; read-out corrector saturation	**known**	504aldo F66, F72–F73; F65, F82; F46–F47
Dead/linear neurons and threshold pruning (§100)	**uncertain**	other teams prune (reds' telemetry, §156), method unpublished
Errors correlated across teams (§126, §147, §188)	**new**	read from public pages
DAgger-trained correctors injected into the chain	**uncertain**	DAgger (Ross, Gordon, Bagnell 2011) is standard; departs from the Phase 1 rule "read out, never inject"; training on the chain's own corrected trajectory ("a posteriori" training in the turbulence-closure sense) is what lets injection work here; literature not searched in depth

9. Credits

The method is built on the factorised K = 3 cumulant propagation of Wu, Lecomte, Winer, Robinson, Hilton and Christiano (arXiv:2605.05179) and the organisers' reference code (`alignment-research-center/mlp_cumulant_propagation`, MIT). Team 504aldo's public estimator, findings log and write-up (`github.com/504aldo/whest-p2-cumulant-k3`, MIT) gave us the regeneration law, the 7 → 4 restructure, the feedback legs, the trim, the age-gated tiers and the Strassen and residual engineering; 504aldo also disclosed that his code was LLM-generated under his direction. The scale-mixture view of deep ReLU networks is due to Hanin & Nica, Noci et al. and Li, Nica & Roy; Puffi's Phase 1 write-up used covariance-response modes for κ_4. The Phase 1 posts of jamesrahenry, trim_qewas, amalgonim and baltsat described how truer pieces can hurt simpler chains. The leaderboard pages of J2W, marius_binner, reds, thegamers and others constrained our hypotheses. Any error in reporting their work is ours.

10. Reproducibility

- **Package of #335513:** `white-box/soumission/soumission-candidate54957.tar.gz` (manifest, `estimator.py`, `table-N.npz`; whestbench 0.16.1, flopscope 0.12.1, Python 3.12). Assembled by `banc/faire_estimateur_k3f.py` from `banc/k3f_m.py` (class ChaineM) and `banc/k3f_tables.py`. Checks: `banc/fabriquer_candidat.py`, `banc/verifier_paquet.py`, `banc/essai_client_strict.py`, `banc/pre_correcteur.py`, `banc/demarrage_officiel.py`, `boucle/trace_1024.py`.
- **Re-measurement for this note:** `white-box/note-prix/verification/`: `construire.py` (variant definitions, bench pass `jf_e54957` run by `k3f_v_flops.py`), `v1_heritage.py`, `v2_kappa3.py`, job folders under `resultats/`, summaries `verification-chiffres*.txt`.
- **Earlier results:** `banc/verite_K4.py`, `banc/oracle_k4_echelle.py`, `banc/ajuster_lambda_mu.py`, `banc/juger_variante_mu.py` (§131–146); `campagne-k3-vieux/r_regen.py` (§155); `banc/cache_chaine.py`, `banc/corriger.py` (§27); `chantier-dagger*/` (correctors); width: `recherche/arbre/projets/boite-blanche/banc-rapide-fidele/`.
- The research notebook (`WRITEUP.md`) is available on request; it records negative results and corrections in the order they happened.

Annex A. Complements

- **The deformation is collective.** At n = 256, L = 16, the skewness predicted by treating neurons as independent decreases with depth (0.020 → 0.0066) while the measured one increases (0.032 → 0.289): 2 % captured at layer 15 (§2).
- **The propagator collapses in rank.** At age 4, rank 384 holds 99.9 % of the transported propagator's energy against 21 % for a random matrix of the same size (§80).
- **Half the neurons are dead or linear.** At the last layer $\varphi(t) < 10^{-3}$ for 30 % of the neurons and $< 10^{-2}$ for 42 %; as hubs they generate almost no third cumulant, but their P legs and the κ_4 feed do not vanish with |t| (§100). Pruning by thresholds, not by counts, is what works (§100, §155).
- **No n^2 closure of D21.** A hub factorisation of the history is 300–600× off, or 3–9× worse than no closure (§61); a one-step closure from the young sources has a median error of 69 % against a 4 % criterion (§174).
- **Read-out correctors saturate.** Fourteen forms land within 2 % of each other (§27).
- **Everyone makes the same errors.** Per-network log-MSE of different teams correlate 0.6–0.87 (§126; J2W against us 0.63, ratio of final MSE 0.765 ± 0.072, §147); the difference to the leaders was later found to be mostly cost, error ×0.78 and compute ×0.66 (§188).

Annex B. Width

Two agents debated this for three rounds on 17 October (synthesis: `banc-rapide-fidele/LARGEUR-ET-TESTS-MODESTES.html`; measured statements only). Deviations from Gaussianity are O(1/n) (n × skewness ≈ 44–47 from n = 128 to 512), finer ones $O(1/n^2)$; a method's read-out error falls as n^{-p} with p set by what it tracks, and depth acts through L/n.

Table B1. Exponent p of the last-layer read error with the true per-neuron cumulants, fitted from n = 128 to 512 on 5 networks per width.

what the read uses	p, limit	p, 8 layers	p, 16 layers
Gaussian (mean, variance)	2	2.11	1.85
+ per-neuron skewness	2	2.24	2.21
+ one common tail number	3	3.17	2.71
+ per-neuron tails	4	4.23	3.48

Our chain's MSE is 2.16e-5 at n = 128, 3.62e-6 at 256 and 2.0e-8 at 1024 (slope −3.75 between 256 and 1024). Laws transport; verdicts do not, because the remaining error mixes pieces with different n-laws whose weights shift (by 5–15 % from 512 to 1024 at 16 layers, by a factor 1.4–3 from 128 to 1024), while modifications move the error by 0.1–1 %. Section 3 is an instance: the third cumulant has a coherent deficit at 256 and none at 1024.

- Gains measured at n = 128 against labels at 1024 (each on 7–8 networks): Spearman −0.46 over 219 pairs (95 % interval by batch resampling [−0.64; −0.23]), −0.57 on the means of 72 batches; for comparison, two halves of the 1024 networks agree at 0.79 (23 variants) (`contradiction-tour-2.txt`).
- An exact term added alone (the square of the skewness): +40 % at 128, +3 % at 256, −24 % at 512.
- Without the read-out table and the crest, the error is 2.9 % lower at 128 and 5.2 % lower at 256, but 15–21 % higher at 1024.
- Rounding the living dimension downward: flops −4.2 % and MSE −2.9 % at n = 128, MSE ×11 at 1024 (§160); attenuating the old sources: −10 to −51 % at n ≤ 128, +6 to +294 % at 1024 (§155); the λ table fitted at 1024 is 4–8× off at 256 (§133).

Practical rule: judge modifications on the real 1024-wide network made cheaper (no flop counter, no Strassen, reused prefixes, a sequential stopping rule on the number of networks), and use small widths only for the sign of large, exact effects.

Annex C. Truer pieces in fitted chains, measured earlier

Table C1. Truer pieces placed into a fitted chain, on earlier chains (named). “x” is a gain factor (> 1 is better); percentages are changes of the error (+ is worse).

piece made truer	chain, width	result	source
true per-layer variance injected alone	early mean/cov chain, n = 512 / 1024	0.60x / 0.75x	§4
σ drift corrected at its measured physical value	early chain, 1024	+149 %; optimum at c ≈ 0	§33
true diagonal step injected at every layer	early chain	1.133e-6 against 5.013e-7 without it	§44
old κ_3 regenerated from n^2 features at the oracle	research chain, n = 64–128	+8 to +17 %, tolerance 2–4 %	§155
old κ_3 coherently attenuated (0.95 … 0.7)	deposit chain, 1024	+6 % to +294 %	§155
three-block fresh κ_4 terms added to the fitted closure	research chain (ChaineP), 1024, network 0	+6 % alone, +2.5 % with the net term	§115