A Provably Near-Optimal Spherical-Cubature Estimator for Mean Activations of Deep Random ReLU Networks under a FLOP Budget

Solution write-up — ARC White-Box Estimation Challenge 2026 (Phase 1)

`skibidi_toilet` (AIcrowd) — submission #325170*

15 August 2026

Abstract

We describe our final Phase-1 submission to the ARC White-Box Estimation Challenge 2026: predict the per-neuron mean activation of a random width-256, depth-32 ReLU MLP under standard-normal inputs, scored by final-layer mean squared error (MSE) multiplied by the fraction of a 2.72×10^{11}-FLOP budget actually spent. The estimator exploits the positive homogeneity of bias-free ReLU networks to factor the Gaussian expectation exactly into a radial constant times a spherical integral, and evaluates the spherical integral with the antipodal closure of the union of all 129 real mutually unbiased bases of $\mathbb{R}^{256}$ (a Kerdock-code line set): 66,048 points forming a *tight* spherical 4-design (a 5-design by antipodal symmetry). Exact analytic corrections pin the layer-1 mean and the layer-2 pre-activation variance; a 1,344-point pilot propagated at full width drives per-layer neuron pruning with an exact compensation term; and the surviving computation is executed at close to the analytical Strassen–Winograd FLOP count under the contest's operation-level billing model. On 100 freshly generated MLPs never used for tuning, the submission attains final-layer MSE 2.599×10^{-7} and an expected adjusted score of 1.068×10^{-7} at 41% effective-compute utilisation, with zero failures (public-leaderboard reading 8.55×10^{-8}); the best published, method-disclosed competitors score 3.7× and 16× worse. We further show this operating point is nearly the floor of its class: the design meets the Welch–Sidelnikov bounds with equality at degrees 2 and 4, no 66,048-point antipodal configuration can improve the degree-6 residual by more than 0.195%, only ∼36% of the integrand's variance is annihilable by any affordable cubature, and a blending bound closes the analytic/hybrid route ($M \leq 9V$ required; measured free-feature correctors reach $R^2 = 0.61$ against the 0.978 needed). Finally, we report cost-model findings we believe are of independent value to the organizers: a dtype-billing interaction worth a deterministic 2.00× on the score, a residual-time law (cost $\propto$ instrumented-call count), and a tail-risk mechanism by which timer attribution, not computation, can zero an otherwise healthy evaluation.

Contents

*AIcrowd profile: https://www.aicrowd.com/participants/skibidi_toilet. Final Phase-1 artifact `02_opus5_wb_v15__expected_adj_1.068e-07.tar.gz`, SHA-256 `dc3caef52889c2365bb4f17eba4ed789216b3dc0e5d4f95fddceaecd5575908e`. Contact: marbinner@gmail.com.

1 Introduction

The ARC White-Box Estimation Challenge asks a sharp question: given the weights of a randomly initialised ReLU MLP, can you predict its expected behaviour *more cheaply than running it*? Concretely, the grader hands the estimator one width-n=256, depth-d=32 network and a FLOP budget $B = 2.72 \times 10^{11}$, and asks for the expected post-ReLU activation of every neuron at every depth under $x \sim \mathcal{N}(0, I_n)$. The ranked metric multiplies final-layer MSE by the fraction of the

budget actually used, so accuracy and frugality are priced against each other on the same axis (Section 2).

Our submission is a *cubature* estimator: a sampling method whose sample locations are chosen so that all low-degree components of the integrand are integrated *exactly*, leaving only high-degree variance. Three exact structural facts about the problem make this work far better than generic Monte Carlo:

1. Bias-free ReLU networks are positively homogeneous, so the Gaussian expectation factors *exactly* into a known radial constant times an integral over the unit sphere (Section 3.1). The radial dimension of the problem costs nothing.
2. The sphere in $\mathbb{R}^{256}$ admits an extraordinarily good, explicitly constructible point set: the union of all 129 real mutually unbiased bases (MUBs) obtained from Kerdock codes. Its antipodal closure is a *tight* spherical 4-design (5-design by symmetry) of 66,048 points that meets the Welch–Sidelnikov moment bounds with equality at degrees 2 and 4, and—as we prove below—is within 0.195% of the best possible degree-6 moment for *any* antipodal configuration of its size (Section 3.2).
3. The first two layers admit cheap exact corrections: the exact layer-1 mean (a norm), and the exact layer-2 pre-activation variance via the ReLU arc-cosine kernel (Section 3.5). Pinning these low-layer moments removes the dominant coherent error before depth can amplify it.

The remaining work is to *spend the FLOPs well*. The contest meters compute with an analytical, operation-level cost model (`flopscope`), which turns implementation into an optimisation problem with unusual and exactly-known prices. The submission propagates the design through the network with pilot-driven neuron pruning plus an exact sparse path for rarely-firing neurons, executes the dense work as a non-stationary five-level Strassen–Winograd recursion, and is engineered so that essentially every billed FLOP is load-bearing arithmetic (Section 4).

Results and claims. All headline numbers are measured on freshly generated evaluation suites that were never used for tuning, with paired statistics (Section 5). Our contributions:

C1. A state-of-the-art estimator. Final-layer MSE 2.599×10^{-7} and expected adjusted score 1.068×10^{-7} on 100 fresh MLPs, and zero failures on every graded and dedicated-host panel (1,599 of 1,600 documented evaluations clean; the single exception is a timer-attribution event under heavy host contention, analysed in Section 7). The best published, method-disclosed competitors are 3.7× (randomised-QMC + Rao–Blackwell) and 16× (trajectory-calibrated cumulant chain) behind (Section 5).

C2. Optimality of the design, with a short new argument. The design is tight at the degrees where configurations can differ, sits within 0.39% of the Delsarte–Goethals–Seidel size floor for its strength, and its degree-6 moment—the first non-exact degree—is provably within 0.195% of the minimum attainable by *any* 66,048-point antipodal set (Proposition 1). Improving the point set is not an available lever.

C3. How much of the problem any affordable cubature can reach. Measured on controlled panels: only ∼36% of the integrand's variance lives at spherical degrees ≤5; the other 64% sits at degree ≥6, where a design would need ≥ 2,861,952 points (43× our budget) to be exact (Section 6).

C4. Closure of the analytic and hybrid routes. An optimal unbiased/biased blend needs partner MSE $M \le 9V \approx 1.85 \times 10^{-6}$ at negligible cost; the organizer-referenced cumulant method at the contest shape delivers 6.8×10^{-5} cheaply or 7.2×10^{-6} at 3.04× the *entire* budget, and a learned free-feature corrector caps at held-out $R^2 = 0.61$ versus the 0.978 required (Section 6.4).

C5. Cost-model findings for the organizers. A dtype-promotion interaction that silently doubles billed FLOPs (fixing it is worth a deterministic 2.00× on the adjusted score); a measured law that residual wall-time is a per-call floor (∼11 µs × 10,402 calls), making residual proportional to instrumented-call count; and a demonstrated failure mode in which timer *attribution* (not computation) can consume 85% of a per-MLP budget with no affordable in-process defence (Sections 4.4, 7, 8).

Throughout we keep the evaluation methodology explicit: what was swept where, which panel each number comes from, and why we rank on deterministic FLOP counts with a calibrated residual rather than on public-leaderboard readings (Section 5.1). The submitted artifact, its hashes, environment pins, and reproduction commands are catalogued in Section 10 and Appendix A.

2 Task and scoring model

Networks. Each evaluation instance is an MLP of width n and depth d (contest shape $n = 256$, $d = 32$) with weight matrices $W_\ell \in \mathbb{R}^{n\times n}$, $\ell = 0, \dots, d-1$, drawn i.i.d. from the He initialisation $\mathcal{N}(0, 2/n)$, no biases. For input $x \sim \mathcal{N}(0, I_n)$ the layers are $h_0 = \mathrm{ReLU}(W_0^\top x)$ and $h_\ell = \mathrm{ReLU}(W_\ell^\top h_{\ell-1})$. The estimator must return the $d \times n$ matrix of means $\mu_{\ell,i} = \mathbb{E}[h_{\ell,i}]$.

Ground truth. The grader compares against Monte-Carlo means baked from $N = 10^9$ input samples per MLP; the induced MSE floor, $\overline{\mathrm{Var}}/N \approx 2 \times 10^{-10}$, is ∼0.1% of our MSE, i.e. the target is effectively exact.

Score. Each `predict(mlp, budget)` call runs inside a `flopscope` budget context that counts every array operation analytically (F_m FLOPs) and measures the *residual* wall time R_m: time inside the call that is neither instrumented-kernel execution nor framework dispatch (participant Python, GC, allocator traffic). Effective compute and the ranked metric are

$$C_m = F_m + \lambda R_m, \qquad \texttt{adjusted\_final\_layer\_score} \;=\; \frac{1}{M}\sum_{m=1}^{M} \mathrm{MSE}_m^{\mathrm{final}} \cdot \max\!\big(0.1,\; C_m/B\big), \quad (1)$$

with $B = 2.72 \times 10^{11}$, $\lambda = 10^{11}$ FLOP-equivalents per second of residual, and lower is better. If $C_m > B$, or a wall-clock cap (60 s per MLP) trips, the MLP's predictions are *zeroed* and the multiplier is forced to 1.0: since the zero prediction has MSE ≈0.9, a single failure costs roughly $10^7\times$ our per-MLP score, which shapes the entire engineering posture (Section 7). All-layer MSE is published as a tie-break diagnostic.

Two consequences of (1) organise everything that follows. First, with MSE $\propto V/k$ for a k-point unbiased sampler of per-point variance V and $F \propto ck$,

$$\text{score} \;\approx\; \frac{V}{k} \cdot \frac{ck + \lambda R}{B} = \frac{V(c + \lambda R/k)}{B}, \quad (2)$$

so the score is *flat in* k up to the residual term, and exactly three levers exist: per-point variance V, per-point FLOPs c, and residual R. Section 6 shows all three are at, or within measurement

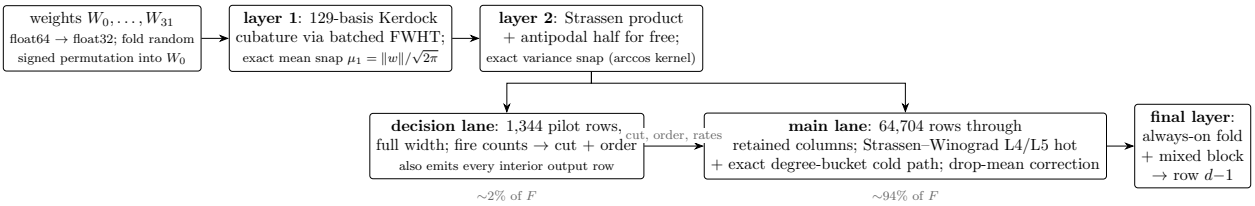

Figure 1: The estimator. One cubature state of 66,048 rows is created at layers 1–2 (positive halves materialised; antipodal halves recovered analytically), then split into a full-width pilot that makes all pruning decisions and a pruned main state that carries ∼94% of the arithmetic. Interior prediction rows come from the pilot; the ranked final row from the full state.

noise of, their floors for this class. Second, the multiplier floor at 0.1 is out of reach for any method whose MSE benefits from spending (we operate at $C/B \approx 0.41$); the incentive is a genuine joint optimisation, not a race to the floor.

3 Method

Figure 1 sketches the pipeline; Algorithm 1 gives the step order. We develop it top-down: the exact factorisation, the point set and its optimality, then the layer-by-layer machinery.

3.1 Exact radial–spherical factorisation

The network has no biases, so every layer map, and hence the whole network $M : \mathbb{R}^n \to \mathbb{R}^n_{\geq 0}$ (any depth prefix of it), is *positively homogeneous*: $M(cx) = c\,M(x)$ for $c > 0$. Writing $x = r\,u$ with $r = \|x\|$ and $u = x/\|x\|$, Gaussian x gives independent $r \sim \chi_n$ and $u \sim \mathrm{Unif}(\mathbb{S}^{n-1})$, hence *exactly*

$$\mathbb{E}_{x\sim\mathcal{N}(0,I_n)}[M(x)] = \mathbb{E}[\chi_n] \cdot \mathbb{E}_{u\sim\mathrm{Unif}(\mathbb{S}^{n-1})}[M(u)]\,, \qquad \mathbb{E}[\chi_n] = \sqrt{2}\,\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \overset{n=256}{=} 15.98438\ldots \quad (3)$$

The radial factor is a constant we know in closed form; all remaining error is the spherical cubature error. Empirically this factorisation alone (i.i.d. unit vectors with antipodal pairing instead of i.i.d. Gaussians) reduces the variance-per-point by 1.45× (Table 4). We implement it by placing the design points on the sphere of radius $\mathbb{E}[\chi_n]$ and never materialising r.

Antipodal pairing does more here than classical antithetic variates: the map $u \mapsto -u$ combined with $\mathrm{ReLU}(z) + \mathrm{ReLU}(-z) = |z|$ and $\mathrm{ReLU}(z) - \mathrm{ReLU}(-z) = z$ makes odd-degree spherical harmonics integrate to zero *exactly*, and (Section 3.5) lets the second layer of the negative half be recovered from the positive half at almost no cost.

3.2 The point set: all 129 real MUBs of $\mathbb{R}^{256}$

A cubature rule with equal weights is a *spherical t-design*: a finite $X \subset \mathbb{S}^{n-1}$ whose empirical average integrates every polynomial of degree $\leq t$ exactly. The quality of an equal-weight rule for an isotropic integrand ensemble is governed by its even moments $S_{2k}(X) = \sum_{u,v\in X} \langle u, v\rangle^{2k}$, which obey the Welch–Sidelnikov bound [6, 7]

$$S_{2k}(X) \;\geq\; |X|^2\,\frac{(2k-1)!!}{n(n+2)\cdots(n+2k-2)}, \quad (4)$$

with equality *iff* X is a spherical $2k$-design. Designs of high strength in high dimension are scarce; the He-initialised, isotropic weight ensemble makes the integrand ensemble isotropic, so (4) is exactly the right yardstick.

Our design is the union of all 129 real mutually unbiased bases of $\mathbb{R}^{256}$: the identity basis plus 128 "flat" bases of the form $H\,\mathrm{diag}(c)/16$, where H is the 256×256 Walsh–Hadamard matrix and $c \in \{\pm 1\}^{256}$ ranges over the Kerdock-code sign patterns (bent-function chirps) of the Calderbank–Cameron–Kantor–Seidel orthogonal-spread construction [3]; $129 = n/2 + 1$ is the maximum possible number of real MUBs in dimension $n = 4^s$ [4]. Every cross-basis inner product has modulus exactly $1/16 = 1/\sqrt{n}$. Closed under $u \mapsto -u$, the set has

$$k \;=\; 2 \times 129 \times 256 \;=\; 66{,}048 \text{ points} \qquad (33{,}024 \text{ antipodal pairs}).$$

The artifact stores it as the 128 sign vectors (`chirps.npz`, ± 1, 7 KB compressed); the flat structure is what later makes layer 1 nearly free (Section 3.4).

Measured moments. On the shipped point set, with the bound (4) evaluated at $|X| = 66{,}048$, $n = 256$:

degree $2k$	achieved S_{2k}	bound	ratio
2	$1.704\,038 \times 10^7$	$1.704\,038 \times 10^7$	1.00000000
4	$1.981\,440 \times 10^5$	$1.981\,440 \times 10^5$	1.00000000
6	$1.323\,540 \times 10^5$	$3.810\,462 \times 10^3$	34.734

Equality to eight decimals at degrees 2 and 4 certifies a *tight spherical 4-design*; antipodal symmetry upgrades it to strength 5. For scale, the Delsarte–Goethals–Seidel bound [5] requires any antipodal 5-design in $\mathbb{S}^{n-1}$ to contain at least $n(n+1)/2 = 32{,}896$ antipodal pairs; we use 33,024, i.e. the design is within 0.39% of the smallest conceivable point count for its strength. Random orthonormal frames of the same size, by contrast, exceed the degree-4 bound by 65.9% and are measured 1.25× worse in final MSE (Section 6).

The degree-6 row is not slack to be engineered away; it is structural:

Proposition 1 (Degree-6 floor for antipodal configurations)**.** *Let $X \subset \mathbb{S}^{n-1}$ be any antipodally closed set of k unit vectors. Then $S_6(X) \geq 2k$. Consequently, at $k = 66{,}048$, $n = 256$: every antipodal configuration has $S_6 \geq 132{,}096$, while the shipped design achieves $S_6 = 132{,}354$. No 66,048-point antipodal set can improve on it by more than $258/132{,}096 = 0.195\%$.*

Proof. The sum $S_6 = \sum_{u,v \in X} \langle u, v\rangle^6$ contains, for each $u \in X$, the diagonal term $\langle u, u\rangle^6 = 1$ and the antipode term $\langle u, -u\rangle^6 = 1$; all other terms are non-negative because the exponent is even. This forces $S_6 \geq 2k = 132{,}096$. For the Kerdock design, vectors within one orthonormal basis are orthogonal, and all cross-basis inner products are $\pm 1/16$, so the entire off-diagonal contribution is $k\,(k - 2n)\,(1/16)^6 = 258$ exactly, giving $S_6 = 132{,}096 + 258 = 132{,}354$. □

The Welch–Sidelnikov degree-6 bound (3,810.5) is therefore unreachable by a factor of 34.67 for *every* antipodal configuration of this size—the bound certifies that no 66k-point 6-design exists, and Proposition 1 says all such configurations pay essentially the same degree-6 energy. This is why random frames sit within 3% of Kerdock at degree 6 while being 66% worse at degree 4: degree 4 is where designs can differ, and there the Kerdock set is exactly optimal. A degree-6 design would need $\binom{258}{3} + \binom{257}{2} = 2{,}861{,}952$ points [5]—43× our point budget. **Conclusion: at this size, the point set is a solved sub-problem.**

3.3 Random orientation

A fixed design risks aligning with a particular network. We therefore apply a uniformly random *signed permutation* to the design's coordinates—equivalently (and this is how it is implemented, so no input batch is ever materialised) to the rows of W_0. A signed permutation is an orthogonal map that permutes the coordinate axes, and the design is invariant in distribution under it *as a design of the same strength*; the He weight distribution is likewise invariant. The orientation is drawn from `default_rng(seed ⊕ salt)` with the MLP's own seed, making the submission bit-reproducible per instance.

Two measured consequences. First, on 8 MLPs × 32 orientations the estimator's squared bias is below the sampling floor of the bias estimate itself: the error is *pure orientation variance*, so MSE falls exactly as $1/k$ in design copies and nothing is removable "for free". Second, this implies the salt constant is not a tunable: for a fresh MLP every salt value has identical expected MSE. We verified the arithmetic of this the hard way—an internal candidate that "gained" 17.7% by selecting a salt over a fixed 50-MLP panel was fully explained by best-of-N selection noise ($\sqrt{2\ln N}\,\sigma$ with panel CV 10.6%; Section 8) and was rejected.

3.4 Layer 1 by fast Walsh–Hadamard transform

For each flat basis, the layer-1 pre-activations of its 256 points are $(H\,\mathrm{diag}(c))^\top$ applied to the columns of W_0, i.e. $H(\mathrm{diag}(c)W_0)$ by symmetry of H: one elementwise sign flip and one batched fast Walsh–Hadamard transform of 8 add/subtract stages in place of a dense $256 \times 256 \times 256$ matmul. Under the billing model the FWHT costs $8n^2$ additions per basis versus $n^2(2n-1)$ for the dense product—a 64× reduction that makes layer 1 ∼1% of the budget. Only the 33,024 positive representatives are materialised; $\mathrm{ReLU}(-z) = \mathrm{ReLU}(z) - z$ recovers the negative half where needed. The identity basis contributes W_0 itself, scaled by $\mathbb{E}[\chi_n]$; flat rows (norm 16 before normalisation) are scaled by $\mathbb{E}[\chi_{256}]/16 = 0.99902\ldots$, placing all points on the radius-$\mathbb{E}[\chi_n]$ sphere of (3).

3.5 Layer 2: the antipodal half for free, and the exact variance snap

Let $z = XW_0$ be the layer-1 pre-activations of the positive representatives (X never materialised; z from the FWHT above) and $h^+ = \mathrm{ReLU}(z)$. The layer-2 pre-activation of a *negative* representative is

$$\mathrm{ReLU}(-z)\,W_1 = (\mathrm{ReLU}(z) - z)\,W_1 = h^+W_1 - \underbrace{X\,(W_0W_1)}_{\text{FWHT of a composed } n\times\text{cut matrix}},$$

so the second layer of all 66,048 points costs one Strassen product of the 33,024 positive rows plus a single FWHT of the small composed matrix W_0W_1—the negative half is essentially free. This "bounded product reuse" is the second place the flat algebraic structure of the design pays off.

Exact moment corrections. Two closed-form facts about Gaussian ReLU moments are then imposed on the state:

- *Layer-1 mean snap.* $\mathbb{E}[\mathrm{ReLU}(w^\top x)] = \|w\|/\sqrt{2\pi}$ [9]. The design's layer-1 sample mean $\bar{h}_1$ is replaced by the exact μ_1; the correction is carried into layer 2 as the rank-one shift $(\mu_1 - \bar{h}_1)W_1$ rather than by touching the 66,048-row state.
- *Layer-2 variance snap.* The exact layer-1 second-moment matrix is the ReLU arc-cosine kernel [8] $\mathbb{E}[\mathrm{ReLU}(u^\top x)\,\mathrm{ReLU}(v^\top x)] = \frac{\|u\|\|v\|}{2\pi}\big(\sin\theta + (\pi-\theta)\cos\theta\big)$ with $\theta = \angle(u,v)$, computed for all column pairs of W_0 from one symmetric Gram matrix. From it, the exact layer-2 pre-activation variance v_j^{ex} of each retained neuron is formed (with the spherical-ensemble radius correction

$\mathbb{E}[\chi_n]^2/\mathbb{E}[\chi_n^2]$), and the centred cubature pre-activations are rescaled by $\sqrt{v_j^{\mathrm{ex}}/v_j^{\mathrm{emp}}}$ around their exact mean $\mu_1 W_1$.

The snap matters more than its size suggests, because depth transports coherent low-layer error *multiplicatively*: in a controlled experiment, a 0.195% error injected into v^{ex} (double-applying the radius factor) raised final MSE 4.2× (variance·k 0.0194 → 0.0813); a 0.1% coherent layer-2 error emerges as 0.19% at layer 32. Conversely the chain neither damps nor amplifies incoherent noise. Getting the low layers *exactly* right is the cheapest variance reduction in the whole system. (Extending the same idea further—full covariance snap at layer 2, exact per-neuron variance at layer 1—was measured a null; Table 5.)

3.6 Pilot-driven pruning of the propagation width

Deep-layer neurons fire with wildly unequal probabilities, and a neuron that fires on a few pilot points in 1,344 contributes almost nothing to downstream means. The estimator therefore maintains two coupled states (Figure 1):

- a **pilot** of $n_{\mathrm{pil}} = 1{,}344$ design rows (an evenly strided, antipodally balanced subset) propagated at *full width* n every layer; and
- the **main state** of the remaining 64,704 rows, propagated only through the retained ("cut") columns of each layer.

At each interior layer the pilot's pre-activations z yield per-neuron fire counts $c_j = \#\{z_{\cdot j} > 0\} \sim \mathrm{Binomial}(n_{\mathrm{pil}}, p_j)$. The *cut size* keeps every neuron with $c_j > p_{\min} = 5$, rounded up to a multiple of 16 (the Strassen block granularity, Section 4.2); typical cuts are ∼200–220 of 256. The *ranking* that decides which neurons occupy the retained block uses a lower-variance monotone proxy for p_j: the shifted count $\#\{z_{\cdot j} > -0.25\,\hat{\sigma}_j\}$ (a much larger effective sample for the same tail ordering) with the smooth tie-break $\frac{1}{2} + \frac{1}{2}\, t_j/(1 + |t_j|)$, $t_j = \hat{\mu}_j/\hat{\sigma}_j$. Both statistics reuse reductions the pilot pass already performs.

Dropped neurons are compensated, not ignored. A dropped neuron j still feeds layer $\ell{+}1$ through row j of $W_{\ell+1}$. Its pilot mean $\bar{h}_j$ (damped by a per-layer factor α_ℓ ramping $0.500 \to 0.625$ over depth) is propagated through the dropped rows as a rank-limited correction to the next layer's pre-activations, and the resulting deficit is carried forward as a running per-neuron δ, gated by each receiving neuron's fire rate p_j (a closed-ReLU passes no mean). Ablating this correction costs +0.8% MSE at zero FLOP savings; ablating pruning entirely *improves* MSE only 1.9% while costing 45% more compute (Table 5)—the cut sits almost exactly where the marginal column's value equals its cost, which is why sweeps of $p_{\min}$ in both directions lose (Section 6).

Because the pilot is itself a subset of the design, its rows after the layer-2 snap are *bit-identical* to the corresponding main-state rows; a fixed row permutation places them contiguously in front, so decision bytes and propagation bytes are the same memory and nothing is computed twice.

3.7 Final layer and the per-layer output rows

At the final layer the fire statistics separate three regimes:

- **Always-on neurons** (pilot fire count = n_{pil}, margin $\min z/\overline{z} \geq \beta$ with the preregistered $\beta = 0$): over the design, ReLU is the identity for them, so their mean is the mean *pre-activation*, computable from the running state mean $\bar{h}$ as $\bar{h}\, W_{d-1}$—one 256-vector product instead of a 64,704-row propagation. These occupy a 16-aligned leading block (the "sign fold").

Algorithm 1 `predict(mlp, budget)` — contest shape ($n = 256$, $d \geq 3$)

cast W_ℓ to float32; fold seeded signed permutation into W_0 ▷ Sections 3.3, 4.4
layer 1: batched FWHT of $\mathrm{diag}(c)W_0$ over 128 chirps + identity basis; ReLU ▷ Section 3.4
snap layer-1 mean to $\|w\|/\sqrt{2\pi}$; carry correction as rank-one shift into layer 2
layer 2: Strassen product of positive half; antipodal half via FWHT of W_0W_1; exact variance snap; ReLU ▷ Section 3.5
permute pilot rows (1,344) to the front; split state → (pilot, main)
for $\ell = 2, \ldots, d-2$ **do**
    pilot: full-width product; fire counts → cut_ℓ, order_ℓ; emit row ℓ mean
    main: retained columns via Strassen–Winograd L4/L5 (hot) + degree-bucket cold path ▷ Sections 4.2, 4.3
    add dropped-neuron mean correction, gated by fire rates; ReLU both lanes
end for
final layer: always-on fold from state mean; mixed block by full propagation; zeros below cut ▷ Section 3.7
return $d \times n$ rows (pilot means for interior layers; exact row 1; folded row $d-1$)

- **Mixed neurons**: full propagation as in the interior.
- **Rarely-firing neurons** (below the cut): predicted 0; their true means are $O(10^{-3})$ with negligible mass.

Interior output rows cost almost nothing: the pilot already holds a full-width post-ReLU state at every layer, so row ℓ is its column mean ($\sim 3.4 \times 10^5$ FLOPs per layer, 0.004% of budget in total; layer 1 emits the exact analytic μ_1). This leaves the ranked final row bit-identical while improving the published tie-break `all_layers_mse` from ~ 0.94 (broadcasting the final row, as most entries do) to 7.2×10^{-5}.

3.8 Off-nominal shapes and budget adaptation

The private re-evaluation uses "the same width and layer-count ranges and the same FLOP-budget calibration", but the estimator must not *assume* the nominal shape:

- For $n \neq 256$ or $d < 3$ (no Kerdock design exists off $n = 256$), a width-agnostic fallback runs antithetic i.i.d. Gaussian directions projected to the radius-$\mathbb{E}[\chi_n]$ sphere—the factorisation (3) and the layer-1 mean snap are dimension-generic—sized to $\sim$9% of the supplied budget.
- A shape-only guard prices the main propagation from the billing model before running it and halves the flat-basis count until the projection fits 75% of the *supplied* budget: by (2) the score is nearly flat in design size, so shrinking is score-neutral insurance against a deeper or tighter private calibration.
- The per-layer damping table clamps at its ends (a depth-33 instance once raised `IndexError` and zeroed itself; found in pre-submission fuzzing, fixed, and re-verified at depths 3–128).

An additional wall-clock degrade path exists in the artifact but is *disabled* (the `v73_development_no_clock` flag); we verified it is arithmetic- and billing-inert as shipped, and Section 7 explains why no such gate can defend against the one failure mode we observed.

4 Spending FLOPs where the meter looks

The contest bills compute analytically per operation, so the implementation target is not wall time but the exact cost model. This section describes the model, then the three mechanisms that put the estimator at $F/B = 0.373$: fast matrix multiplication, an allocation discipline that makes bookkeeping free, and an exact sparse path for rarely-firing neurons. It closes with the two cost-model discoveries that dominated everything else.

4.1 The billing model

`flopscope` 0.10.0 charges flop_cost × weight × dtype_rate: matmul $MN(2K-1)$; elementwise ops, reductions, *and data movement* (`reshape`, `concatenate`, `stack`, copies, fills) 1/element; gathers (`take`, fancy indexing) 4/element; sorts $\approx 4N\lceil \log_2 N \rceil$; transcendentals and `x**y` 16/element; symmetric `einsum` Gram products billed on the upper triangle only. Views, `empty`, `zeros`, and `transpose` are free, and `out=` writes charge only the arithmetic of the op being written. The dtype rate is 1.0 for ≤32-bit floats (float16 = float32: there is no sub-single-precision discount) and 2.0 for float64, following NumPy promotion. Residual wall time—anything between counted kernels—is converted at $\lambda = 10^{11}$ FLOP/s and added to C_m (1).

Three practical corollaries shaped the code: (i) never pay for data movement—write every combination directly into a preallocated view; (ii) never let a float64 operand promote an expression; (iii) minimise the *number* of instrumented calls, because residual turns out to be a per-call tax (Section 4.5).

4.2 Strassen–Winograd propagation at the analytical count

All large products are (rows × p) · ($p \times q$) with rows ∈ {64,704, 33,024} and p, q multiples of 16 (the cut granularity guarantees this). They are executed as a recursive 2×2 block Strassen scheme [10] of depth $L = 4$ (blocks 16×16), with a fifth, non-stationary outer level when both operand widths permit ($\min(p, q) \geq 160$, multiples of 32):

- **Winograd form inside.** The Winograd variant of Strassen trades the classic $(5, 5, 8)$ additions per level for $(4, 4, 7)$, with the B-side combinations landing on the *weight* matrix, which is negligible against a 64,704-row data side [11]. Billed cost at L=3: 0.7433× dense versus 0.7494 for classic Strassen.
- **Preregistered outer rule.** At the fifth level two rank-7 rules were implemented—classic Strassen and the Dumas–Pernet–Sedoglavic scheme [12]—and the choice (classic) was frozen on the development panel before any fresh-suite evaluation.
- **Copy elision.** The three copy-only operand slots at the deepest level remain strided views contracted by separate matmul calls; per the model, a copy would bill 1/element for nothing. Reconstruction writes the seven products straight into the $(g, g, \text{rows}/g, q/g)$ block grid the *next* layer consumes, so no billed re-blocking occurs between layers.
- **Arenas.** `setup()` reserves one product arena and one recursion-scratch arena as uninitialised backings (free) with views prepared for every shape the contest recursion can request; every view is fully overwritten before being read. Prediction-time allocation cost: zero billed elements, near-zero allocator traffic.

For rows $\gg p = q$ the per-level cost ratio is $7/8 + \frac{5}{8q} + \frac{1}{p}$, so the fifth level is worth ∼2.3% at width 256 and becomes a loss below width ≈208; the eligibility threshold was swept (144–999;

total effect $\leq 0.25\%$ of C) and set to the measured optimum 160. Where L5 is *illegal*, forcing it by padding costs +4.5%—the guard, not the level, is the optimisation.

4.3 The exact cold path for rarely-firing neurons

After ReLU, the columns of the main state that belong to low-fire-rate neurons are mostly zeros. The retained columns are ordered by fire rate (Section 3.6), and the trailing 16-aligned block whose cumulative pilot hits fit a mass cap (the maximal suffix with $\sum c_j \leq 4\, n_{\text{pil}}$) is contracted *exactly* but sparsely: for each design row, the number q of non-zero cold entries is counted, rows are bucketed by exact degree $q \in \{1, \ldots, 6\}$, a padded degree-≤ 8 route, and a dense $q \geq 9$ route; each bucket gathers its q values and weight rows and contracts them at $O(q)$ per output. The break-even fire rate where the sparse route stops paying is $p^\star = \frac{1}{3}(7/8)^L \in [0.171, 0.195]$ (gather at 4/element + einsum versus the Strassen-discounted dense rate); the shipped boundary lands at 0.175. The path saves 14% of F while itself costing 4.5% of F; the outputs are exact, so predictions change only by float32 reassociation.

4.4 The float64 trap: a deterministic $2.00\times$

The grader hands `mlp.weights` as **float64**. Under NumPy promotion every downstream temporary inherits it, so the entire hot path bills at the $2\times$ dtype rate—and, more insidiously, prepared float32 arenas never match and silently fall back to fresh allocations. One cast per layer ($\sim 2 \times 10^6$ FLOPs total, 0.0007% of budget) moved the submission from $F/B = 0.740$ to 0.371 and, by re-activating the arenas, collapsed residual from 8.9 s to 0.79 s on the development host. Numerical cost, measured element-wise: RMS prediction change 3.4×10^{-6} against the estimator's own RMS error of 3.8×10^{-4} (+0.008% MSE). On the deterministic FLOP-only score this single fix is a $2.00\times$ improvement (adj(F): $2.083 \times 10^{-7} \to 1.040 \times 10^{-7}$, measured ratio 0.499)—larger than every algorithmic refinement after the initial design combined. We flag the generalisable lesson in Section 8.

A second allocator-level tuning (`mallopt` raising the glibc mmap/trim thresholds in `setup()`) removed a further 0.79 s→0.16 s of residual *locally*; cross-checking against a sibling implementation without it showed the grader's residual is unaffected (0.12 s both ways)—the local gain was an artefact of a contended development host. We report it as shipped-but-inert on the grader, and it illustrates why we do not trust single-host residual measurements (Section 5.1).

4.5 Residual is a per-call tax

Direct measurement of the residual bucket: 3.3 µs per instrumented call for tiny operands, 13.2 µs at 54 MB with `out=`, and 6250 µs when a fresh 54 MB output must be allocated (hence the arenas). The shipped estimator makes 10,402 instrumented calls per MLP at ∼11.4 µs: 0.118 s, matching the grader's measured residual exactly. Residual is therefore proportional to *call count* (with a mild size term), i.e. a ∼4.3% surcharge on F rather than an independent lever: every attempt to trade calls against FLOPs moved the total by ∼1:1 or worse (dense pilot −26% calls +1.5% F: net 0.4%, inside noise; Strassen L3 fewer calls: +8.7% F). This law also kills design-doubling (Section 6).

4.6 Where the FLOPs go

Exact per-op attribution of the shipped estimator's $F = 1.010 \times 10^{11}$ (one contest MLP):

operation family	FLOPs	share
Strassen–Winograd leaf matmuls	8.00×10^{10}	79.2%
Winograd A/C-side additions	1.39×10^{10}	13.7%
cold-path gathers (`take`, 4/elt)	3.14×10^{9}	3.1%
cold-path contractions (`einsum`)	9.9×10^{8}	1.0%
`reshape` (blocked↔dense)	4.4×10^{8}	0.4%
ReLU (`maximum`)	4.3×10^{8}	0.4%
cold-path merge (`concatenate`, scatter)	6.0×10^{8}	0.6%
pilot, decisions, layers 1–2, analytic snaps	—	∼4%

93% of billed compute is the irreducible arithmetic of the propagation itself, at the analytical Strassen count for these shapes; accounting overhead has been engineered to ∼1%.

5 Evaluation protocol and results

5.1 Protocol: fresh suites, paired statistics, deterministic ranking

Three disciplines governed every measurement:

1. **Fresh, untuned evaluation data.** Headline numbers come from `fresh100`: 100 width-256 depth-32 MLPs with seeds from `secrets.randbits(63)`, ground truth from $N = 2 \times 10^8$ GPU samples (three A100s, merged; ground-truth noise 2.4×10^{-10}, 0.09% of the estimator MSE). *None of this estimator's choices were tuned on this suite or on the public split.* Hyperparameter sweeps ran on separate development panels.
2. **Paired comparisons.** Per-MLP MSE has ∼75% relative SD, so candidate deltas are measured MLP-by-MLP (and where possible orientation-by-orientation) against the incumbent; the −0.71% of the final version is resolved at ±0.02% this way, invisible to unpaired means.
3. **Deterministic ranking metric.** Analytical FLOPs are exactly reproducible; residual is host-dependent (worker-to-worker spread ±0.1 s ≈ ±3.7% of C). Internal ranking therefore uses adj = MSE × $(F + \lambda R^{\text{grader}})/B$ with the grader's calibrated residual $R^{\text{grader}} = 0.118$ s, never single public readings. Public-panel readings are reported but are optimistic by ∼25–30% for everyone (panel luck; and see Section 8 on selection).

5.2 Headline numbers

Table 1 summarises the submission. Per-MLP adjusted scores on the fresh suite span 2.6×10^{-8} to 4.6×10^{-7} (heavy-tailed, as expected from orientation variance); the estimator spends 37% of the budget on instrumented FLOPs (41% effective compute) and half the wall-time cap, leaving 2.4× budget and ∼2× wall margins as failure insurance.

5.3 Calibration against other methods

Table 3 places the submission 3.7× ahead of the strongest published unbiased competitor and 16× ahead of the strongest published analytic chain. The randomised-QMC author independently concludes their method sits at "the practical accuracy-per-FLOP frontier among unbiased methods"; we read the residual gap between their frontier and ours as precisely the value of (i) the exact radial factorisation, (ii) tightness of the design at degrees 2–4, and (iii) the exact layer-1/2 moment snaps. Public-leaderboard entries far below all method-disclosed results are discussed—in aggregate only—in Section 8.

Table 1: Final submission on 100 fresh MLPs (never used for tuning) and on the grader. "Expected adjusted" pairs each MLP's MSE with its own F_m plus the grader-calibrated residual. The public reading is included for completeness; it is optimistic for all participants (Section 8). Public-column MSE and wall detail are the 100-MLP telemetry of the shared public panel (read via submission #324894); the v13/v15 public readings are indistinguishable (8.5529 vs 8.5530×10^{-8}).

	fresh-100 (untuned)	public grader (#325170)
final-layer MSE	2.599×10^{-7} (per-MLP SD 1.96×10^{-7})	2.048×10^{-7}
all-layers MSE (tie-break)	7.17×10^{-5}	—
mean F/B	0.3714 (SD 0.0133, max 0.4006)	0.374
mean multiplier C/B	0.414 (grader-calibrated)	0.418
wall time per MLP	—	29 s mean, 31.7 s max (cap 60 s)
failures	0/100	0/100
adjusted score	$\mathbf{1.068 \times 10^{-7}} \pm 7.9 \times 10^{-9}$	8.553×10^{-8}
FLOP-only adj(F), deterministic	9.583×10^{-8}	—

Table 2: Development lineage on fixed internal panels (dev50: 50 fresh MLPs; fresh100 as in Table 1). adj(F) is the deterministic FLOP-only score; "expected adj." adds the grader-calibrated residual. The v0→v2 step is the float64/arena fix (Section 4.4); v13 adds real per-layer rows, depth clamps and shape fallbacks; v15 re-tunes two cost constants and removes 3.2 ms of residual. v14/v15 predictions are bit-identical; v13→v14 differs only by float32 reassociation (MSE ratio 1.000025).

version	change	panel	result
v0	initial Kerdock cubature (float64 billing)	dev50	adj(F) $= 2.083 \times 10^{-7}$
v2/v5	float32 cast; arenas live; `mallopt`	dev50	adj(F) $= 1.040 \times 10^{-7}$ (2.00×)
v13	per-layer rows; depth/shape hardening	fresh100	expected adj. 1.0755×10^{-7}
v15 (final)	cold mass 3→4; L5 gate 192→160; array-op row permutation	fresh100	expected adj. $\mathbf{1.0680 \times 10^{-7}}$ paired v15/v13: 0.99287 ± 0.00024

6 Why this is the floor of its class

We spent the second half of Phase 1 trying to beat this estimator and failed in an unusually informative way: every lever in (2) is measurably at its optimum. This section is the evidence; we consider it the write-up's most useful content for future work.

6.1 Variance per point V: at the information-theoretic floor

Table 4 decomposes the gain over Monte Carlo. The remaining $V = 0.0185$ is not "the best design we found"—it is pinned on three sides:

- *Bias:* zero to measurement precision (Section 3.3); MSE is pure orientation variance, falling exactly as $1/k$.
- *Degrees* ≤ 5*:* the design integrates them exactly, and tightness (equality in (4)) certifies no 66,048-point set does better.
- *Degree* ≥ 6*:* carries 64% of the integrand's variance (only ∼36% is annihilable: $0.02900 \to 0.01847$ against an exact-through-5 rule), and Proposition 1 bounds every antipodal competitor within 0.195% of our degree-6 moment. Exactness at degree 6 needs 2.86×10^6 points—43× the budget.

Table 3: Where this sits. Baselines from the starter kit at the same budget; published competitor numbers from Phase-1 forum method disclosures (the only entries with disclosed, reproducible methodology); ARC's cumulant-propagation method re-implemented and run by us at the contest shape against baked ground truth. Adjusted scores where available; raw final-layer MSE otherwise.

method	source	final-layer MSE	adjusted
zeros baseline	starter kit	0.91	9.1×10^{-2}
mean propagation	starter kit	9.5×10^{-4}	9.5×10^{-5}
covariance propagation	starter kit	8.4×10^{-5}	8.4×10^{-6}
cumulant propagation, $k_{\max}$=2	[1], our run	6.8×10^{-5}	($F = 0.008B$)
cumulant propagation, $k_{\max}$=3	[1], our run	7.2×10^{-6}	($F = 3.04B$: infeasible)
trajectory-calibrated cumulant chain	forum topic 18097	5.9×10^{-6}	1.77×10^{-6}
randomised QMC + Rao–Blackwell	forum topic 18053	—	4.10×10^{-7}
this work	fresh-100	$\mathbf{2.60 \times 10^{-7}}$	$\mathbf{1.07 \times 10^{-7}}$

Table 4: Variance ladder: variance$\times k$ of the final-layer estimate, matched panels (30 MLPs $\times$ 6 orientations, everything at $k = 66{,}048$). Each rung is the measured value of one exact structural feature.

sampler	var $\cdot k$	gain
i.i.d. Gaussian Monte Carlo	0.04198	—
i.i.d. unit vectors, antipodal (exact radial factor)	0.02900	1.45×
Kerdock tight 5-design (shipped)	0.01847	1.57× further

A sharper corollary we verified the expensive way: **subsets of a tight design are not tight**. A 97-basis sub-design (natural for buying budget margin) has degree-4 moment 1.219× the bound, and its measured MSE penalty is 1.438 ± 0.026—the 1.330 predicted by pure $1/k$ scaling times 1.083 for the lost tightness. The score-flatness of (2) holds only *along families of equally good designs*, and Kerdock has no smaller tight member at $n = 256$.

6.2 FLOPs per point c and residual R: jointly at their floor

Every discrete knob was swept on paired panels; the shipped values are optima, mostly with the two gradients cancelling almost exactly:

- **Cut threshold** $p_{\min}$: paired sweep over 26 MLPs, values 2/3/4/5/7 give adj(F) ratios 1.0041/1.0025/0.9999/1.0000/1.0110 ($\pm\sim$0.01): the marginal column is worth almost exactly what it costs. (A GPU replica had suggested a 1% basin at $p_{\min} \in [7, 10]$; it does not reproduce on the real estimator—a caution against surrogate tuning.)
- **Cold-mass cap**: F/B across mass 1/2/3/5/8 is 0.3815/0.3745/0.3693/0.3706/0.3848 at identical MSE; the final re-sweep put the optimum at 4 (together with the L5-gate re-tune, worth -0.41%).
- **Strassen depth**: L4→L5 saves 2–2.3% where legal; forcing it by padding costs +4.5%; the eligibility gate moves total C by $\leq 0.25\%$ over its whole range.
- **Residual**: proportional to call count (Section 4.5); every call-vs-FLOP trade tested lands within noise of 1:1.

6.3 Design growth: capped by the budget, correctly costed

Doubling the design verifiably halves MSE (paired ratio 0.545 ± 0.053, consistent with exact $1/k$). Writing $C = X + gV'$ with g design copies and X the g-independent part, the best possible gain from amortising X is $1 - V'/(X + V') = 13.8\%$ at $g \to \infty$; the budget caps $g \leq 2.5$; at $g = 2$ the naive gain is 6.7%, but per-call residual grows with operand size, collapsing the real gain to 1–2%—while pushing wall time from ∼30 s toward the 60 s cap. We bought failure margin instead.

6.4 The analytic and hybrid routes are closed

The only exit from the cubature class is to blend in an independent biased estimator. Optimally combining an unbiased estimator of variance V with a biased one of MSE M yields $VM/(V+M)$, so a 10% gain requires

$$M \;\leq\; 9V \;\approx\; 1.85 \times 10^{-6} \qquad \text{at negligible FLOP cost.} \tag{5}$$

Measured against (5):

- **Cumulant/moment propagation** [1] at the contest shape: $k_{\max}$=2 costs $0.008B$ but delivers $M = 6.8 \times 10^{-5}$ ($M = 262V$: blend worth 0.4%, costing 1.9%); $k_{\max}$=3 reaches $M = 7.2 \times 10^{-6} = 27V$ at $F = 3.04B$—three times the *entire* budget. Two sibling investigations measured the same closure ($M/V \approx 280$) independently. Consistent with the source paper, whose factorised estimators fall behind sampling by 8 hidden layers—here $d = 32$.
- **Learned free-feature corrector**: 400 fresh MLPs, 309 analytic features (per-neuron trajectories, per-layer summaries), correcting the Gaussian-closure output. Requirement from (5): held-out $R^2 \geq 0.978$. Achieved: ridge 0.586, tuned MLP 0.612 (then overfits). The gap is a factor ∼17 in residual variance, not a tuning margin. This independently reproduces a forum-reported 2.3× corrector ceiling with a stronger model.
- **Best published probe-fed chain** (forum #18097, $M = 5.9 \times 10^{-6}$ at $0.17B$): as a blend partner, *net negative* (−36%): the probe cost swamps the variance it removes. It would need to be ∼3× more accurate and ∼5× cheaper simultaneously to turn positive.

The two halves close each other: driving M below $\sim 10^{-5}$ demonstrably requires spending FLOPs on samples—and once FLOPs go to samples, Section 3.2 says the tight design is the optimal way to spend them. A genuinely better estimator at this shape must reduce *per-point variance at degree* ≥6 by other means; we believe that is the open problem, and state it as such.

6.5 Negative results with mechanisms

Table 5 catalogues what did not work and why. We highlight one calibration warning for anyone building a successor: the layer-2 variance snap is *exquisitely* sensitive—a 0.195% coherent error in the exact variance raised final MSE 4.2× (Section 3.5). Exactness of the analytic constants is load-bearing; approximate "improvements" upstream of depth are how this class of estimator quietly loses.

7 Robustness and tail risk

A zeroed MLP scores $\sim 10^7\times$ our per-MLP mean (1); at any realistic private-suite size, one failure outweighs every refinement in this document. The submission's engineering posture reflects that asymmetry.

Table 5: Ideas measured and rejected, with the mechanism of failure. Paired panels of ≥ 24 MLPs unless noted; "ratio" is candidate/incumbent on the stated metric (< 1 better). Entries marked † are cost-model interactions rather than statistics.

idea	result	mechanism
no pruning at all	score $\times 1.380$	MSE -1.9% for $+45\%$ cost: the cut is priced correctly
harder pruning ($p_{\min}$=30)	score $\times 2.03$	dropped mass is no longer negligible
full-covariance layer-2 snap	MSE $\times 1.017$	diagonal snap already captures the coherent part
exact layer-1 (co)variance snaps	null ($\pm 0.7\%$)	layer-1 sample moments already tight at k=66k
dropped-neuron correction off	MSE $\times 1.008$	the δ chain carries real mass
final-layer fold off	cost $\times 1.009$	fold is exact; removing it only re-buys FLOPs
fold margin $\beta > 0$	exactly null	no pilot-boundary always-on errors exist at k=1344
multilevel control variate (full-width twin)	score $\times 1.035$	twin variance $\approx$ its cost; no cheap coupling
low-rank state, r=64	MSE $\times 39{,}000$	truncation error compounds over 29 layers; 99% energy needs rank 118
rank columns by σ, $\sigma\lvert w\rvert$, energy, ...	0.99–1.01	fire-count ranking is already sufficient
per-layer $p_{\min}$ ramps	0.994–1.021	flat optimum; see paired sweep
pilot rows 672 / 2688	$+6.7\%$ / -0.7% MSE	binomial cut noise vs pilot cost; 1344 optimal
float16 storage†	no lever	billed = float32 by the model
design doubling $k \to 2k^\dagger$	$+1$–2% only	residual grows with operand size; wall $\to$ cap
97-basis sub-design (margin buy)	MSE $\times 1.438$	subsets of a tight design are not tight
salt/orientation selection	rejected by policy	pure $\sqrt{2 \ln N}$ selection noise (Section 8)

Margins and envelopes. Grader telemetry (100-MLP detail of submission #324843; consistent across this lineage's four gradings): C/B mean 0.418, SD 4.9%, max 0.520; wall mean 29.5 s, max 33.0 s against a 60 s cap; extreme-value scaling puts $\mathbb{E}[\max C/B]$ at 0.498 even for a 2,000-MLP suite. The per-MLP tail contributes essentially nothing to bust risk: the spread is dominated by the data-dependent cut sizes, bounded by construction at $F/B \leq 0.425$ (measured max over 1,200 local evaluations: 0.4250).

Off-nominal inputs. Depths 1/2/3/8/32/33/40/64/96/128, widths 17/64/128/255/256/257/512: all return finite (d, n) predictions within budget (Section 3.8). Grader-stack parity (Python 3.10.20 / NumPy 2.2.6 vs our 3.12 / 2.4.6): MSE agreement to 0.04%, zero failures. Subprocess isolation with `RLIMIT_AS` 65 GB: peak RSS 5.7 GB; `setup()` 0.023 s against a 5 s allowance.

The 1000-MLP stress run, and the one failure we cannot guard. On a deliberately contended host we evaluated 1,000 public MLPs: 999 clean (aggregate over valid MLPs 1.127×10^{-7}), zero exceptions, zero non-finite outputs, zero FLOP or wall overruns—and exactly one zeroed MLP, of a kind worth the organizers' attention. Its FLOPs ($F/B = 0.378$) and wall clock (21.9 s, 64th percentile) were unremarkable, but 2.3 s of time that is normally attributed to instrumented kernels

landed in the *residual* bucket ($R = 2.45$ s vs. median 0.169 s). At $\lambda = 10^{11}$ FLOP/s that reattribution alone is $0.90B$: $C/B = 1.28$, prediction zeroed, per-MLP score 0.437—a 3,875× inflation of the 1000-MLP aggregate from an accounting event with *no* wall-clock signature ($\mathrm{corr}(R, \mathrm{wall}) = 0.057$ across the run).

Both conceivable in-process defences are quantitatively ruled out. A wall-clock gate cannot see it (the wall was normal). Reading the residual from inside `predict` requires `budget_summary_dict()`, whose measured cost is ∼0.5 µs *per instrumented op of process history* and does not reset between MLPs: by the end of a 100-MLP grading run one probe costs ∼0.5 s, i.e. 18% of a budget per look—two orders of magnitude more than the event it watches for. Headroom arithmetic is equally unforgiving: surviving a spike of S seconds needs $C/B \le 1 - \lambda S/B$, i.e. $(1 - C/B) \times 2.72$ s of tolerance—1.54 s at our utilisation, 2.72 s even for a *free* estimator, versus the 2.45 s observed. The event appeared only under heavy co-tenancy (0 in ∼400 dedicated-core grader evaluations; 0 in 200 low-contention local ones), so we assess it as a grader-environment tail risk rather than a defect of this submission—but it is unguardable from inside, and we return to it in Section 8.

8 Observations on the evaluation methodology

We offer these as constructive findings from a team that benefited from the cost model being taken literally; each is measured, not speculative.

1. The float64 hand-off is a trap, not an incentive. Estimators receive float64 weights while the model prices float64 at 2× and NumPy promotion silently spreads it (Section 4.4). The fix is one cast, but nothing in the harness surfaces it, and it is worth a deterministic 2× on the ranked metric—more than any algorithmic choice we made after the initial design. Recommendation: hand estimators float32 weights, or document the cast prominently in the starter kit.

2. λR prices attribution, not work. Residual wall-time is a per-call floor (Section 4.5) plus whatever the timer attributes there; Section 7 shows a 2.3 s attribution shift with no wall-clock signature consuming 85% of a budget, with the only in-process instrument priced at 18% of budget per reading. Recommendations: reset the summary accumulator per MLP (making a cheap self-check affordable); score residual against a per-run calibration or cap its share of C; and/or treat $C_m > B$ caused solely by residual-time anomalies on an otherwise-normal wall clock as re-gradeable. We reported the probe-cost issue during Phase 1 (forum topics 18092/18093/18101).

3. Un-instrumented compute is under-priced 10–100×**.** $\lambda = 10^{11}$ FLOP/s converts residual seconds to FLOPs, but a modern vectorised host sustains well above that off-meter, so work executed outside instrumented kernels is systematically under-billed (community measurements: forum topics 18099/18108). The public leaderboard's extreme low scores are consistent with this channel (instrumented shares < 0.001 measured by participants, versus 0.897 for this submission) and with public-panel selection; the organizers have stated both will be addressed at private re-evaluation (topics 18118/18132), and the single-core sandbox change already narrows the arbitrage. We endorse the stated principle that residual time must not function as a second compute lane, and note that our submission's score composition (90% instrumented FLOPs, residual = measured call-count floor) is designed to be invariant to any such tightening.

4. Public-panel readings reward entry count. Per-MLP MSE has ∼75% relative SD, so a 50-MLP panel mean has CV ∼10.6% and best-of-N selection buys $\sqrt{2 \ln N}\,\sigma$: 33–39% of apparent

MSE for teams with hundreds of entries—no modelling, just order statistics. This fully explains several public gaps ahead of us—and the arithmetic applies to our own account as much as to anyone's: the team ran many submissions across parallel development lanes, and this artifact's public reading (8.55×10^{-8}) is itself ∼25% optimistic against its fresh-100 measurement. That is exactly why no promotion decision in this artifact's lineage was made on a public reading, and why this write-up quotes fresh-suite numbers as primary. The organizers' fresh-suite re-evaluation design is exactly right; we simply document the magnitude of the effect it removes.

9 Limitations and outlook

Scope. The estimator is specialised to the contest ensemble: bias-free ReLU MLPs at He initialisation. The exact factorisation (3) needs positive homogeneity (biases, LayerNorm, or smooth activations break it, though analogous Gauss–Hermite radial factorisations exist); the Kerdock design needs $n = 4^s$ (other widths fall back to antithetic spherical sampling); and random weights make the integrand ensemble isotropic, which is what our optimality yardstick (4) assumes. Trained networks have anisotropic, correlated weights: the design would survive, but the variance ladder and the pruning statistics would need re-measurement, and mean behaviour is no longer the interesting functional. We view the value of this work for ARC's agenda as (i) a measured map of where the sampling frontier actually is at depth, and (ii) an existence proof that *exact structural facts—homogeneity, algebraic point sets, closed-form low-layer moments—compound into a* 3.7× *lead over generic sampling within a fully-audited compute model.*

What would beat this submission. In order of leverage: a variance-reduction method for spherical degree ≥6 at depth (the open problem; Section 6.4); an analytic partner with $M \leq 9V$ at negligible cost (5); a larger FLOP budget, which re-opens design growth along tight families ($1/k$ verified); or a cost-model change repricing the components (Section 8). Within the current rules and budget, we assess the remaining headroom above this artifact at ∼0.3% (measured; Table 2 banked it) plus whatever noise our ±0.02% paired resolution conceals.

10 Reproducibility

Artifact. The final submission is AIcrowd submission **#325170** (accepted, graded clean), archive `02_opus5_wb_v15__expected_adj_1.068e-07.tar.gz`, SHA-256 `dc3caef5...` (full hashes in Appendix A). It contains three files—`estimator.py` (the complete method of Sections 3–4, ∼1,700 lines), `sparse_kernel.py` (the cold path), `chirps.npz` (the 128 Kerdock sign vectors)—plus the generated manifest. No network access, no external data, no learned parameters; the only randomness is the orientation, seeded from `mlp.seed`, so predictions are bit-reproducible per instance.

Environment. `whestbench` 0.14.0, `flopscope` 0.10.0; validated on Python 3.12 (NumPy 2.2.6 and 2.4.6) and on the grader stack (Python 3.10.20 / NumPy 2.2.6); cross-stack final-layer MSE agreement 0.04%, zero failures. Evaluation: `whest run --estimator estimator.py --dataset <suite> --runner subprocess`. The fresh-100 suite is reconstructible from its published per-MLP seed list (Appendix A) with ground truth at $N = 2 \times 10^8$; all sweep panels, per-MLP JSON evidence (including the 1000-MLP stress run), and the exact Welch-moment/DGS verification scripts are preserved in the development repository and available to the organizers on request.

Process discipline. This artifact's lineage was graded four times (submissions #324843, #324885, #324894, #325170) and none of those readings fed back into its development; the team's wider submission history across parallel lanes played no role in any decision reported here; all hyperparameters were frozen on development panels before fresh-suite evaluation; the outer Strassen

rule and the fold margin were preregistered; a code freeze preceded the deadline, with changes admitted only after the full off-nominal shape sweep plus a 100-MLP evaluation.

Disclosure. In accordance with the challenge guidelines on tool use: the estimator, the experiments, and this document were developed with substantial assistance from AI coding agents (LLMs) directed and reviewed by the team; all reported measurements were produced by the challenge harness (`whest/flopscope`) or by the scripts preserved alongside the artifact.

Acknowledgements

We thank the organizers for an unusually well-instrumented challenge—the analytical cost model is what makes results in this report exactly reproducible—and the forum participants whose method disclosures (topics 18053, 18097) made honest external calibration possible.

References

[1] W. Wu, V. Lecomte, M. Winer, G. Robinson, J. Hilton, P. Christiano. *Estimating the expected output of wide random MLPs more efficiently than sampling.* arXiv:2605.05179, 2026. (Alignment Research Center.)

[2] Alignment Research Center. *Competing with sampling*, blog post, https://www.alignment.org/blog/competing-with-sampling/.

[3] A. R. Calderbank, P. J. Cameron, W. M. Kantor, J. J. Seidel. $\mathbb{Z}_4$-Kerdock codes, orthogonal spreads, and extremal Euclidean line-sets. *Proc. London Math. Soc.* (3) 75(2):436–480, 1997.

[4] P. O. Boykin, M. Sitharam, M. Tarifi, P. Wocjan. Real mutually unbiased bases. arXiv:quant-ph/0502024, 2005.

[5] P. Delsarte, J. M. Goethals, J. J. Seidel. Spherical codes and designs. *Geometriae Dedicata* 6:363–388, 1977.

[6] L. R. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Trans. Inform. Theory* 20(3):397–399, 1974.

[7] V. M. Sidelnikov. New bounds for densest packing of spheres in n-dimensional Euclidean space. *Math. USSR-Sbornik* 24(1):147–157, 1974.

[8] Y. Cho, L. K. Saul. Kernel methods for deep learning. *Advances in Neural Information Processing Systems* 22, 2009.

[9] B. J. Frey, G. E. Hinton. Variational learning in nonlinear Gaussian belief networks. *Neural Computation* 11(1):193–213, 1999.

[10] V. Strassen. Gaussian elimination is not optimal. *Numerische Mathematik* 13:354–356, 1969.

[11] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*, 2nd ed., Chapter 23 (fast matrix multiplication). SIAM, 2002.

[12] J.-G. Dumas, C. Pernet, A. Sedoglavic. Strassen's algorithm is not optimally accurate. *Proc. ISSAC*, 2024. arXiv:2402.05630.

[13] S. S. Schoenholz, J. Gilmer, S. Ganguli, J. Sohl-Dickstein. Deep information propagation. *ICLR*, 2017. arXiv:1611.01232.

[14] AIcrowd. *ARC White-Box Estimation Challenge 2026*: challenge page https://www.aicrowd.com/challenges/arc-white-box-estimation-challenge-2026; starter kit https://github.com/AIcrowd/whest-starterkit; harness https://github.com/AIcrowd/whestbench; cost model https://github.com/AIcrowd/flopscope; public dataset `aicrowd/arc-whestbench-public-2026` (Hugging Face).

[15] AIcrowd challenge discussion forum, Phase 1: method disclosures, topics 18053, 18097; cost-model reports and organizer statements, topics 18092, 18093, 18099, 18101, 18105, 18108, 18118, 18129, 18132.

A Artifact manifest and configuration

Submitted archive

file	SHA-256	role
archive (`.tar.gz`, 29,579 bytes)	`dc3caef52889c236...aecd5575908e`	submission #325170
`estimator.py`	`a89281783cfc977c...698b0b5dcef86a`	method
`sparse_kernel.py`	`cc6b6a06fca77766...9526c49971cd4b0`	cold path
`chirps.npz`	`193bfedd56c89b13...90222541ec26055`	128 Kerdock sign vectors

Manifest pins: `whestbench` 0.14.0, `flopscope` 0.10.0, Python $\geq$ 3.12 (grader 3.10 verified), NumPy 2.2.6; entry point `estimator.Estimator`; packaged 2026-08-06, promoted 2026-08-07.

Frozen configuration (contest shape)

constant	value	meaning
flat bases / points	128+1 / 66,048	complete real-MUB union, antipodal closure
n_{pil}	1,344	pilot rows (antipodally balanced, strided)
$p_{\min}$	5	pilot fire-count cut threshold
cut granularity	16	guarantees Strassen L=4 on every layer
cold-mass cap	4	hot/cold boundary: max suffix with $\sum c_j \leq 4n_{\text{pil}}$
ranking shift	$0.25\,\hat{\sigma}$	shifted-count ordering statistic
α_ℓ	$0.500 \to 0.625$	dropped-mean damping ramp over interior layers
L5 eligibility	width ≥ 160, divisible by 32	non-stationary outer Strassen level
outer L5 rule	classic (preregistered)	vs. Dumas–Pernet–Sedoglavic alternative
fold margin β	0 (preregistered)	final-layer always-on gate
budget guard	75% of supplied B	halves basis count if projected cost exceeds it

Evaluation suites referenced

- **fresh100** (headline): 100 MLPs, seeds from `secrets.randbits(63)` (list preserved with the artifact), GT $N = 2 \times 10^8$ (3×A100, merged); generated 2026-08-06; never used for tuning.
- **dev50 and sweep panels**: development-only; all sweeps and preregistrations happened here.
- **public 1000-MLP split** (`arc-whestbench-public-2026@v1-phase1`, `full`): stress run of Section 7.
- **Grader**: submissions #324843, #324885, #324894 (development lineage), #325170 (final); zero failures in all four.