Compute-Aware Randomized QMC for Deep ReLU Mean Estimation

ARC White-Box Estimation Challenge 2026 - Phase 1 Algorithmic Contribution

Submission #327792

Participant	Ömer Kiraz (AIcrowd: omer_kiraz)
Public adjusted score	3.327 × 10^-7
Public final-layer MSE	1.062 × 10^-6
Mean compute usage	31.35% of 2.72 × 10^11 FLOPs / MLP
Public grading	50 / 50 MLPs completed, 0 failures

A 28-submission experimental study of the compute-accuracy frontier

Abstract

This write-up describes Phase 1 submission #327792, a sampling-based estimator for the expected post-ReLU activations of depth-32, width-256 random ReLU MLPs under standard-normal inputs. The final estimator uses 20,000 points from a randomized rank-1 quasi-Monte-Carlo (RQMC) lattice. The generating vector is g_j = frac(sqrt(p_j)), where p_j is the j-th prime; an independent Cranley-Patterson shift is generated from the MLP seed, the resulting points are mapped through the Gaussian inverse CDF, and the network is evaluated exactly on those inputs. The first-layer mean is reported analytically, while all downstream activation means are ordinary RQMC sample means. The submission completed all 50 public MLPs with zero failures, obtaining an adjusted score of 3.327 × 10^-7, a raw final-layer MSE of 1.062 × 10^-6, and 31.35% mean compute usage. Across 28 Phase 1 submissions, the central empirical finding was that minimizing raw MSE is not the same as minimizing the competition objective: larger sample counts continued to reduce raw error but increased adjusted score because compute grew faster than accuracy improved. A sweep from 12k to 40k samples produced a local optimum near 20k. Antithetic pairing, an alternative Roberts generating vector, and an expected-gain control variate all worsened the official public score. The pure-RQMC core is not claimed as novel; it was motivated by public Phase 1 discussion. The contribution here is the controlled characterization of this estimator under the current deep-MLP grader and cost model, together with negative results that delimit the useful operating regime.

1. Problem setting and objective

WhestBench asks for an estimator of per-neuron post-ReLU means given the complete weights of a random MLP. **Only the final layer enters the primary ranking metric.** For Phase 1, the relevant per-MLP score is the final-layer MSE multiplied by a compute factor, with a floor at 0.1:

$$\text{score}_m = \text{MSE}_{\text{final},m} \times \max(0.1, C_m / B_m)$$

Here C_m is effective compute measured by flopscope (including penalized residual wall time) and B_m is the per-MLP budget. This objective creates a nontrivial accuracy-compute trade-off: spending more compute is beneficial only when the resulting reduction in raw MSE is sufficiently steep. Submission #327792 deliberately optimizes this product rather than raw MSE alone.

2. Final estimator: randomized rank-1 QMC

2.1 Point construction

Let d=256. Let p_1,...,p_d be the first d primes and define g_j = frac(sqrt(p_j)). For N=20,000 and a random shift $\Delta \sim \text{Uniform}([0,1)^d)$, generated deterministically from the MLP seed, the k-th lattice point is

$$u_k = \text{frac}(k\, g + \Delta), \quad k = 1,\ldots,N$$

Each coordinate is clipped only to avoid numerical infinities in the Gaussian inverse CDF. The input samples are then $x_k = \Phi^{-1}(u_k)$, computed in float64; the expensive network matrix multiplications are performed in float32. The same 20,000 samples are propagated through all 32 layers of the actual ReLU network.

2.2 Why the estimator is unbiased

For any fixed lattice index k, adding an independent uniform random shift modulo one makes u_k uniformly distributed on [0,1)^d. Applying the coordinate-wise Gaussian inverse CDF therefore gives x_k ~ N(0,I). The samples are dependent across k, which is precisely what gives RQMC its variance-reduction behavior, but each term has the correct marginal distribution. Consequently, the arithmetic mean of the network outputs is an unbiased estimator of the desired expectation. The random shift is re-created from the MLP seed, so the estimator is deterministic for a given grader instance while still retaining the random-shift construction across the evaluation distribution.

2.3 First-layer closed form

If the first-layer preactivation of neuron j is z_j = w_j^T X with X ~ N(0,I), then z_j ~ N(0, ||w_j||^2) and

$$\mathbf{E[ReLU(z_j)] = ||w_j|| / \sqrt{2\pi}.}$$

Submission #327792 reports this exact value for layer 1. Importantly, the sampled layer-1 activations are still used for all subsequent forward propagation, so this closed form improves the all-layer diagnostic but does not itself alter the final-layer sample estimator.

2.4 Pseudocode

```
g[j] = frac(sqrt(j-th prime))
shift = Uniform([0,1)^256; seed = mlp.seed)
for k = 1..20000:
    u[k] = frac(k * g + shift)
    x[k] = NormalPPF(u[k])
for layer l = 1..32:
    x = ReLU(x @ W[l])
    estimate[l] = mean_k x[k]
estimate[1] = exact Gaussian ReLU mean
return estimate
```

3. Development path: 28 official Phase 1 experiments

The final method was reached through 28 submissions in the Phase 1 submission history. The campaign began with analytic mean/variance propagation and full-covariance Gaussian closures, then moved through Hermite/Price covariance corrections and hybrid analytic-RQMC estimators. Those hybrids reduced error substantially but retained closure bias. A major transition occurred when the analytic closure was removed and the estimator became pure randomized QMC through the full network. From that point, the remaining experiments were organized as controlled ablations around the RQMC design and the compute-accuracy operating point.

The most useful lesson from the campaign is methodological: the official metric must be treated as an objective in its own right. Raw accuracy, all-layer accuracy, and adjusted score can move in different directions, so every added computation must justify its cost under the grader rather than only improve MSE.

4. Sample-count sweep and compute frontier

Submission	Samples	Compute	Final-layer MSE	Adjusted score
#327795	12,000	18.85%	2.123e-6	4.001e-7
#327810	18,500	29.00%	1.154e-6	3.345e-7
#327792	**20,000**	**31.35%**	**1.062e-6**	**3.327e-7**
#327803	21,200	33.22%	1.027e-6	3.411e-7
#327783	27,000	42.28%	8.491e-7	3.590e-7
#327787	40,000	62.58%	6.114e-7	3.827e-7

The sweep gives a direct counterexample to selecting by raw MSE. Moving from 20k to 40k samples reduced raw final-layer MSE by about 42%, but the adjusted score became about 15% worse because compute nearly doubled. Conversely, reducing to 12k cut compute substantially but nearly doubled raw MSE and worsened the adjusted score by about 20%. The neighborhood around 20k is therefore a genuine empirical operating point for this lattice under the Phase 1 grader, not simply the largest affordable sample count.

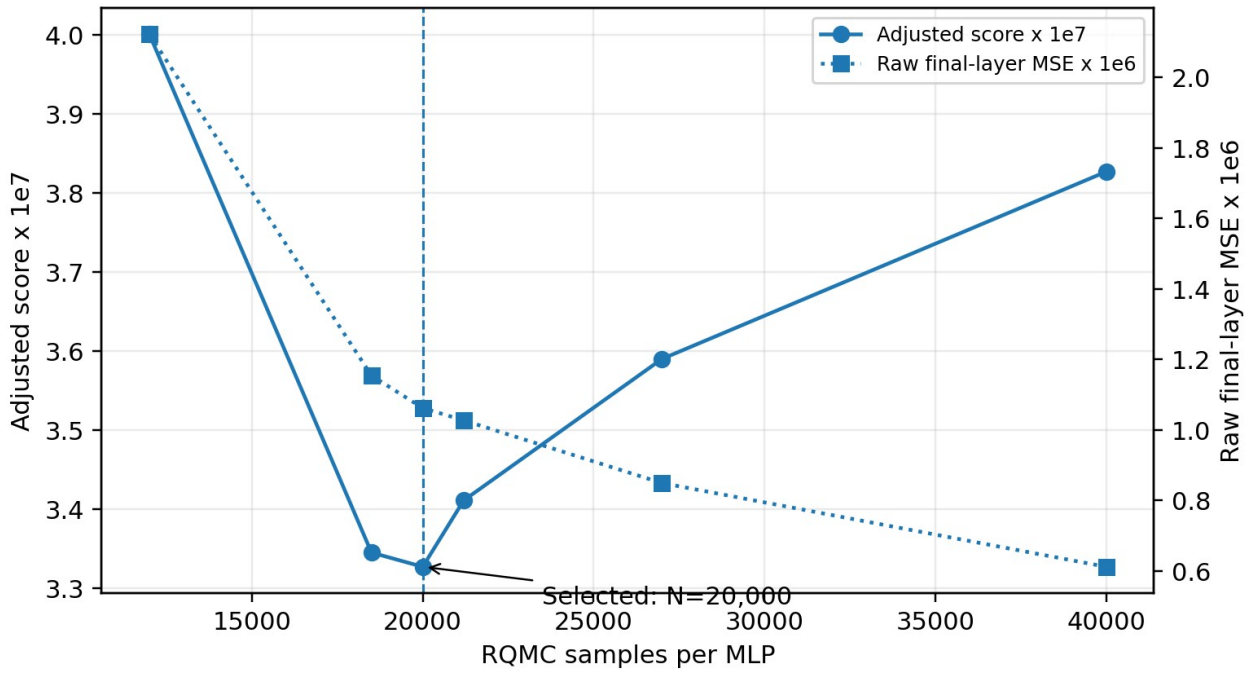

Figure 1. Sample-count sweep. Raw MSE continues to decrease with N, while adjusted score is minimized near 20,000 samples.

5. Structural ablations and negative results

Submission	Variant	Compute	Final-layer MSE	Adjusted score
#327792	**sqrt-prime RQMC, 20k (selected)**	**31.35%**	**1.062e-6**	**3.327e-7**
#327784	antithetic RQMC, 27k total	42.06%	1.091e-6	4.589e-7
#327812	Roberts generalized-golden-ratio vector, 20k	31.34%	2.087e-6	6.542e-7
#327815	layer-1 expected-gain control variate, 20k	31.49%	1.083e-6	3.412e-7
#327778	hybrid analytic/RQMC with Wick-Hermite covariance	10.81%	5.834e-6	6.308e-7

5.1 Antithetic pairing

At essentially the same compute as the 27k pure lattice, replacing half the points by exact Gaussian mirrors x and -x worsened raw MSE from 8.491e-7 to 1.091e-6 and adjusted score from 3.590e-7 to 4.589e-7. The odd-component cancellation that makes antithetic sampling attractive for many integrands did not compensate for the loss of distinct lattice points in this deep ReLU network.

5.2 Generating-vector sensitivity

At the same 20k sample count and essentially identical compute, a Roberts generalized-golden-ratio generating vector produced 2.087e-6 raw MSE and 6.542e-7 adjusted score - roughly twice the error of the sqrt-prime construction. This was one of the largest negative results in the campaign and shows that rank-1 lattice quality is strongly dependent on the generator at a fixed finite N.

5.3 Layer-1 control variate

A zero-mean layer-1 control variate was propagated through deterministic expected ReLU gains. It slightly increased compute and produced 3.412e-7 adjusted score, worse than the 3.327e-7 baseline. This suggests that the low-discrepancy sample already captures much of the first-layer structure relevant downstream, while approximate gain propagation introduces enough mismatch to erase the benefit.

5.4 Hybrid analytic closures

Earlier submissions used diagonal/full-covariance Gaussian propagation, Hermite/Price corrections, moment shrinkage, and analytic final-layer integration. These methods were far cheaper, but the best versions plateaued well above the pure-RQMC frontier. Adding an eighth-order Wick/Hermite covariance anchor increased compute beyond

the 10% score floor and also worsened raw MSE in our hybrid, illustrating that a locally more detailed covariance approximation does not necessarily reduce the downstream closure bias of a 32-layer ReLU network.

6. Interpretation

6.1 The useful frontier is an objective-level property

The final selection cannot be inferred from raw accuracy alone. Between 20k and 40k, additional RQMC points improve MSE but not quickly enough to pay for the linear growth in forward-pass compute. At lower N, however, the lattice enters a visibly worse finite-sample regime. The competition optimum is therefore produced by the intersection of two effects: diminishing error reduction at higher N and deterioration of low-discrepancy quality at lower N.

6.2 Pure sampling beat our biased white-box hybrids

Our initial expectation was that a hybrid white-box estimator would dominate full forward sampling. Under the deep Phase 1 network, the opposite occurred for our implementations. Removing the Gaussian closure bias and spending more compute on an unbiased RQMC estimator reduced adjusted score from the high-5e-7 range to 3.327e-7. This is not evidence that mechanistic estimation is intrinsically inferior; it is evidence that the particular closure approximations we tested accumulated enough depth-dependent bias that their compute advantage was insufficient.

6.3 Negative results matter for method design

The antithetic, Roberts-vector, and expected-gain-CV experiments all had plausible theoretical motivations and comparable computational cost, yet each degraded the official score. Recording these failures is useful because it narrows the search space: future improvements should target either better low-discrepancy constructions that are demonstrably robust at the chosen N, or structural methods that reduce the cost of exact propagation rather than layering approximate corrections on top of already-effective RQMC.

7. Relation to prior work and limitations

The core idea of randomized QMC for this challenge was already public during Phase 1. In particular, the public write-up “Unbiased randomized-QMC + Rao-Blackwell for post-ReLU activation means” described random-shift rank-1 lattice sampling and discussed a sqrt-prime construction. I do not claim the RQMC core of #327792 as a new algorithm. The contribution of this submission is instead the independent implementation and the 28-submission experimental characterization under the current deep-MLP evaluator: the sample-count frontier, the generator sensitivity, and several negative control-variate and antithetic results.

The method is also sampling-heavy rather than mechanistic. It uses the network weights only through ordinary forward evaluation and does not exploit neuron-level dead/on/kink structure, path sparsity, or exact distributional identities beyond the first layer. Accordingly, it should be viewed as a strong compute-aware sampling baseline and a diagnostic frontier, not as a solution to the broader mechanistic-estimation goal. Public Phase 1 results from structure-aware methods demonstrate that substantially better scores are possible when network structure is used to reduce effective sampling cost.

8. Reproducibility details for submission #327792

Submission ID	#327792
Estimator build tag	OMER_WHest_V20_PURE_RQMC_20K_20260810_2330
Estimator SHA256	24b30a72209bd0e1b5ceafc422da99199275b033129bc36d4b4620f8b8198008
Samples	20,000 per MLP
Generating vector	g_j = frac(sqrt(p_j)), first 256 primes
Randomization	one independent 256-D Cranley-Patterson shift seeded by mlp.seed
Inverse transform	Gaussian inverse CDF in float64; clipped only for numerical tail safety
Network arithmetic	float32 matrix multiplications and ReLU operations through all 32 layers
Layer 1 output	exact Gaussian ReLU mean; sampled layer-1 activations still feed layer 2
Official public result	3.327e-7 adjusted; 1.062e-6 final-layer MSE; 31.35% mean compute; 50/50 completed; 0 failures

9. LLM usage disclosure

Development was extensively LLM-assisted. OpenAI ChatGPT was used to brainstorm estimator variants, generate and modify estimator.py implementations, reason about the compute/accuracy trade-off, interpret official grader outputs, and help draft this technical write-up. The human participant selected and submitted variants to AIcrowd and supplied the resulting grader measurements back into the development loop. All numerical results reported here are copied from official AIcrowd grader outputs; they are not LLM-generated estimates. Several LLM-proposed ideas failed in the grader (including antithetic pairing, the Roberts generating vector, higher-order covariance corrections, and a layer-1 control variate), and those failures are reported rather than omitted. The mathematical description of #327792 was checked against the final estimator code used to produce the submission.

10. Conclusion

Submission #327792 is a deliberately simple estimator: 20,000 random-shift sqrt-prime rank-1 QMC samples propagated through the full network, with an exact first-layer mean reported for the diagnostic row. Its value in this Phase 1 study is not a new mechanistic identity but a clear empirical boundary. Under the current cost model, 20k was better than both smaller and larger sample counts; raw MSE improvements beyond this point did not pay for their compute; and three plausible variance-reduction or lattice modifications failed. The resulting adjusted score of 3.327 × 10^-7 provides a reproducible compute-aware baseline and a set of concrete negative results against which more structural Phase 2 estimators can be evaluated.

References

[1] AIcrowd / Alignment Research Center. ARC White-Box Estimation Challenge 2026 - challenge page and Phase 1 evaluator documentation.

[2] AIcrowd Forum thread 18041. “Algorithmic contribution prize guidelines: How ARC judges these prizes - discretion, technical writeups & LLM usage.”

[3] AIcrowd Forum. “Unbiased randomized-QMC + Rao-Blackwell for post-ReLU activation means - method, unbiasedness proof, and where the frontier is.”

[4] Wilson Wu, Victor Lecomte, Michael Winer, George Robinson, Jacob Hilton, and Paul Christiano. “Estimating the expected output of wide random MLPs more efficiently than sampling.” arXiv:2605.05179, 2026.

[5] AIcrowd Forum thread 18106. “Phase I Submission: Structure-Aware Estimation for Random ReLU MLPs.”