Complementary Errors Beat Either Branch

A Fixed Gain-Covariance and Whitened Antithetic Monte Carlo Blend for WhesTBench Phase 2

WhesTBench Phase 2 submission 328274

26 August 2026

Technical write-up for the ARC White-Box Estimation Challenge 2026
This report is mapped to AIcrowd submission 328274.

Public Mini result	3.0440×10^{-7}	Baseline reduction	24.85%
Compute use	9.345%	Failed MLPs	0 / 100

Submitted method	0.75 gain-covariance propagation + 0.25 whitened antithetic Monte Carlo
Evaluation cohort	Official `v2-phase2` public Mini split, 100 MLPs, width 1,024, depth 16
Code SHA-256	`5d747a91672861526f3407211de85803aef8b61932bd4fd31e56407b57948846`
Tarball SHA-256	`7687d9f8f65ec4bbab47cd18793faf91a5f5ff08fb743065642f4ef4ff9a069e`
Report date	26 August 2026

The algorithmic-contribution rules do not require placement in the score-based rankings. This report evaluates the frozen artifact through preregistered experiments, negative-result records, and measured comparisons; it makes no rank-based claim.

Abstract

This report studies a deliberately simple hybrid estimator for the expected post-ReLU activations of wide random multilayer perceptrons. One branch is the official full-covariance Gaussian closure: it propagates means and covariances analytically, applies exact marginal ReLU moments, and approximates off-diagonal post-ReLU covariance by a gain product. The second branch is a whitened, antithetic Monte Carlo estimator adapted from a public MIT-licensed implementation. The submitted estimator fixes their output weights at 0.75 and 0.25, respectively.

On the 100-MLP official Phase 2 public Mini split, the fixed blend reduced adjusted final-layer MSE from $4.050381542 \times 10^{-7}$ for covariance propagation to $3.044010045 \times 10^{-7}$, a **24.846% reduction**, with zero failures and mean compute utilisation of 9.345%. The local result was within 1.433% of the successfully evaluated live score of 3.001×10^{-7}. A post-submission branch decomposition found pooled residual correlation $r = 0.0382$. The closed-form MSE-minimising covariance weight was 0.7341 and the measured grid optimum was 0.735, while the preregistered shipped weight was 0.750 and cost only 0.118% relative MSE versus the measured optimum. The Monte Carlo branch was 2.652 times worse than covariance propagation alone, yet its low error alignment made a 25% mixture beneficial.

The development record includes one scientific negative and three operational failures directly relevant to the method. A full-strength second-order Gaussian Hermite correction worsened paired final-layer MSE by 1.204%. Two early full-split evaluations failed because their runners exceeded preregistered wall-time caps; neither observed a full result. A later sensitivity run was stopped before reading its curve when the loader silently selected the wrong seed protocol. A corrected loader reproduced the first-eight gate with 0.0000% drift. These distinctions matter: they separate a falsified mechanism from evidence invalidated by execution.

Contribution. Covariance propagation, whitening, antithetic sampling, and convex averaging are established techniques. The contribution is a hash-bound, falsifier-backed measurement that the covariance and Monte Carlo branches' Phase 2 errors are complementary enough for a fixed, untuned blend to remove 24.85% of the covariance baseline's adjusted MSE, together with a closed-form weight check and the negative results that delimit the claim.

1 Claim discipline

The experiment ledger uses three labels throughout this report.

- **Measured** means the value is copied from a finished registered run, a receipt-bound artifact, or the successfully evaluated AIcrowd entry.
- **Derived** means the value follows algebraically from measured aggregates. The derivation and inputs are shown.
- **Unvalidated explanation** means a plausible mechanism is offered to guide further work but the present evidence does not isolate it causally.

The word "mechanistic" is used narrowly. The quadratic benefit of combining two recorded residual fields is validated. The proposed explanation, that one field mainly carries Gaussian-closure bias while the other mainly carries sampling noise, is consistent with the design but was not isolated by a dedicated causal experiment and is therefore labeled unvalidated.

2 Problem and score

For a bias-free ReLU MLP of depth L and width n, with standard-normal input,

$$h^{(0)} = x, \qquad x \sim \mathcal{N}(0, I_n), \tag{1}$$

$$z^{(\ell)} = (W^{(\ell)})^\top h^{(\ell-1)}, \tag{2}$$

$$h^{(\ell)} = \mathrm{ReLU}(z^{(\ell)}), \qquad \ell = 1, \dots, L. \tag{3}$$

The estimator must return the per-neuron means of every hidden layer. Phase 2 scores the final row against a high-precision Monte Carlo target. For MLP m, the adjusted score is

$$s_m = \mathrm{MSE}_m \max\left(0.1, \frac{C_m}{B_m}\right), \tag{4}$$

where Phase 2 sets $B_m = 2^{41}$ FLOPs and requires meaningful arithmetic to occur through `flopscope`. The submitted method used 9.345% of the budget on average, below the 10% multiplier floor. Within this observed operating point, both the covariance branch and the blend receive the same 0.1 multiplier; their adjusted-score ratio is therefore their raw-MSE ratio.

The official Phase 2 cohort used here contains 100 public Mini MLPs, each width 1,024 and depth 16. Ground-truth means in the dataset were baked from 10^9 samples per MLP. The private final suite was never available locally, so no result in this document estimates private-suite rank or award probability.

3 Submitted estimator

3.1 Branch A: gain-covariance propagation

The analytic branch starts with $\mu = 0$ and $\Sigma = I$. Each linear step is exact under the current approximate moments:

$$\mu^- = W^\top \mu, \qquad \Sigma^- = W^\top \Sigma W. \tag{5}$$

Let $\sigma_i = \sqrt{\Sigma^-_{ii}}$, $\alpha_i = \mu^-_i / \sigma_i$, and let ϕ and Φ denote the standard-normal density and distribution functions. The marginal ReLU mean and second moment are

$$\mu^+_i = \mu^-_i \Phi(\alpha_i) + \sigma_i \phi(\alpha_i), \tag{6}$$

$$q^+_i = \big((\mu^-_i)^2 + \sigma_i^2\big)\Phi(\alpha_i) + \mu^-_i \sigma_i \phi(\alpha_i). \tag{7}$$

Thus the diagonal variance $q^+_i - (\mu^+_i)^2$ is exact for the Gaussian marginal assumption. The off-diagonal closure is the gain approximation

$$\Sigma^+_{ij} \approx \Phi(\alpha_i)\Phi(\alpha_j)\Sigma^-_{ij}, \qquad i \neq j. \tag{8}$$

The implementation uses float32 arrays, symmetry-aware contraction, exact diagonal replacement, and symmetry revalidation after each write. It is adapted from the official starter-kit covariance example at the pinned source revision listed in Section 8.

3.2 Branch B: whitened antithetic Monte Carlo

The stochastic branch draws half of an ensemble and mirrors it:

$$X = \begin{bmatrix} Z \\ -Z \end{bmatrix}, \qquad Z_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{9}$$

With k rows, it forms $G = X^\top X/k$ and whitens the ensemble with $G^{-1/2}$. Rather than paying for $XG^{-1/2}$ directly, the code fuses the whitener into the first layer:

$$X(G^{-1/2}W^{(1)}). \tag{10}$$

The remaining layers are ordinary batched ReLU forward passes, and each layer's activation mean is the row average. At the Phase 2 shape, the archived branch-decomposition run records $k = 4{,}038$ samples. Antithetic pairing forces the empirical mean and every odd input moment to zero; whitening forces the empirical second moment to identity. These are properties of the input ensemble, not claims that the nonlinear network output is unbiased after finite-sample whitening.

The sample sizer reserves conservative upper bounds for both branches, targets 9.5% of B_m, rounds to an even count, and caps at 100,000 samples. The observed mean utilisation was 9.345%.

3.3 Fixed blend

The returned prediction is

$$\widehat{Y} = 0.75\,\widehat{Y}_{\text{cov}} + 0.25\,\widehat{Y}_{\text{wmc}}. \tag{11}$$

The coefficient was frozen before the full 100-MLP result and before any weight sweep. The later sweep is therefore a sensitivity analysis of a shipped choice, not a model-selection procedure. No empirically selected weight was submitted.

4 Development and evaluation protocol

4.1 Registered sequence

Nine WhesTBench experiments were registered before execution. Each record included a hypothesis, mechanism, required prize delta, cheap falsifier, untouched holdout statement, cost cap, success criteria, and stop conditions. Table 1 gives the report-relevant sequence.

Table 1: Registered development sequence. "Operational" means the measurement process failed; "scientific" means the preregistered mechanism gate failed.

Time (UTC)	Experiment	Outcome	What was learned
24 Aug 03:50	Official package baseline	Passed	Pinned environment, validator, package, and metadata checks passed before public Mini scoring.
24 Aug 03:56	Covariance wiring gate	Passed	One generated Phase 2 MLP satisfied contract, time, shape, and budget gates.
24 Aug 03:59	Covariance Mini	Passed	All 100 public Mini MLPs completed; adjusted score $4.050381542 \times 10^{-7}$.
24 Aug 04:12	Hermite-2 correction	**Scientific failure**	Full-strength second-order correction worsened paired final-layer MSE by 1.204%; the public candidate holdout stayed unopened.
24 Aug 04:24	Covariance-WMC blend	**Operational failure**	First eight improved by 32.246%, but the subprocess run emitted no full result before its 1,200 s cap.
24 Aug 04:51	Local holdout retry	**Operational failure**	MLP 0 matched exactly; the serial run emitted no aggregate result before its 360 s cap.
25 Aug 12:40	Full Mini blend	Passed	Corrected in-process runner completed 100/100 with adjusted score $3.044010045 \times 10^{-7}$.

Time (UTC)	Experiment	Outcome	What was learned
25 Aug 12:50	Weight response v1	**Operational failure**	A 1.5671% reconstruction drift tripped the falsifier before the curve was read; the wrong loader had selected seed protocol 2.0.
25 Aug 12:58	Weight response v2	Passed	Official loader restored 0.0000% first-eight drift; residual correlation and the full weight response were then measured.

4.2 Identity and leakage gates

The code inside the receipt-bound tarball is byte-identical to the locally reviewed candidate. The tarball manifest pins Python 3.10.20, `whestbench` 0.16.0, `flopscope` 0.12.0, and NumPy 2.2.6. The full-Mini run used the official `v2-phase2` revision and a preflight digest of `2c8a2c9d92d2bf18e84ac135c7525a2689d967583b3452e8110851778814a97b`.

The first eight MLPs were used as an identity falsifier. Their candidate-to-baseline ratio had to reproduce 0.6775 within 0.05 before the remaining result could be accepted. The completed run reproduced the first-eight adjusted candidate score bit-for-bit at $2.976850084 \times 10^{-7}$. The prior failed runners had not emitted the final 92 result values or a full aggregate. No coefficient, estimator logic, dataset revision, or seed protocol changed between the identity gate and the completed 100-MLP evaluation; only the runner and wall-time allowance changed.

The weight-response analysis had an additional gate: reconstructing the shipped 0.75 weight had to match the completed full-Mini score within 1%. Version 1 failed at 1.5671% drift and was discarded before any other weight was read. Version 2 loaded the split through `whestbench.load_dataset`, restoring the Phase 2 metadata side channel and matching the first-eight reference with 0.0000% drift.

5 Results

5.1 Primary comparison

Table 2: Full public Mini comparison. Lower adjusted final-layer MSE is better. Raw MSE is ten times the adjusted score because both methods remain below the 10% compute floor.

Method	Adjusted MSE	Raw final MSE	Mean utilisation	Failures
Gain covariance baseline	$4.050381542 \times 10^{-7}$	$4.050381542 \times 10^{-6}$	2.351%	0/100
Fixed 0.75/0.25 blend	$3.044010045 \times 10^{-7}$	$3.044010045 \times 10^{-6}$	9.345%	0/100
Candidate / baseline	0.751537		–	–
Reduction	**24.846%**		–	–

The paired comparison fixes the MLP cohort, ground truth, scoring rule, and software identity. It therefore measures the effect of replacing the covariance output with the frozen blend on this cohort. It does not establish the same percentage on a fresh private suite.

The successfully evaluated AIcrowd entry scored 3.001×10^{-7}. The local full-Mini result is 1.433% higher, or a local-to-live ratio of 1.0143. That agreement is a useful external consistency check, but it is one live evaluation and is not a confidence interval.

5.2 Small-split optimism

The first-eight gate produced a candidate-to-baseline ratio of 0.677538, or a 32.246% improvement. The full-100 ratio was 0.751537, or a 24.846% improvement. The first-eight result therefore overstated the absolute improvement by 7.400 percentage points, and its ratio was 9.846% lower than the full-split ratio. This is why the report uses the full-100 number as its headline and retains the first-eight result only as an identity gate.

6 Why the blend works on the measured cohort

Let e_C and e_M be the final-layer residual vectors of covariance propagation and whitened Monte Carlo against the fixed ground truth, pooled over the measured cohort. For covariance weight w,

$$e(w) = we_C + (1 - w)e_M. \tag{12}$$

Define

$$A = \frac{\|e_C\|^2}{N}, \quad B = \frac{\|e_M\|^2}{N}, \quad C = \frac{\langle e_C, e_M \rangle}{N}. \tag{13}$$

The MSE is the exact quadratic

$$J(w) = w^2 A + (1 - w)^2 B + 2w(1 - w)C, \tag{14}$$

with minimiser

$$w^* = \frac{B - C}{A + B - 2C}. \tag{15}$$

This identity makes the blend effect easy to audit: it requires only the two branch residual fields, not a flexible fitted model.

Table 3: Weight-response measurements. Scores use the same 100 public Mini MLPs and the same corrected seed protocol.

Configuration	Covariance weight	Adjusted MSE	Interpretation
Whitened antithetic MC alone	0.000	1.074091×10^{-6}	2.652 times covariance MSE
Measured grid optimum	0.735	3.040423×10^{-7}	Post-submission sensitivity
Closed-form optimum	0.7341	$\approx 3.040412 \times 10^{-7}$	Derived from pooled residuals
Shipped fixed blend	0.750	3.044010×10^{-7}	Frozen before the sweep
Covariance alone	1.000	4.050382×10^{-7}	Official analytic baseline

The archived analysis reports residual correlation $r = 0.0382$. Although B is much larger than A, the cross term is small enough that a modest stochastic component reduces the total squared error. The closed-form optimum 0.7341 independently predicts the grid optimum 0.735. The shipped 0.750 weight lies 0.0159 away and is only 0.118% above the measured optimum. An oracle that chooses a separate weight per MLP, using ground truth unavailable to an estimator, improves only another 1.69%; on this cohort, most of the attainable benefit of this two-branch family is captured by one global weight.

Measured residual geometry and causal limit. The residual quadratic, the low recorded correlation, and the optimum-weight agreement establish error complementarity on the public Mini cohort. The proposed causal decomposition remains **Unvalidated explanation** because no intervention isolated it. The result does not establish that the correlation remains near zero on private MLPs or under another Monte Carlo seed.

7 Negative results and ablations

The ledger records scientific negatives and operational failures separately. Scientific negatives falsify a mechanism under its gate; operational failures invalidate the measurement process without falsifying estimator accuracy.

7.1 Scientific negative: full-strength Hermite-2 correction

The candidate augmented the first-order gain covariance term with the exact second-order Gaussian Hermite term, then restored the exact marginal variance. On one preregistered generated Phase 2 MLP with a shared 200,000-sample target, baseline final-layer MSE was $5.568854249 \times 10^{-6}$ and candidate MSE was $5.635910384 \times 10^{-6}$. The ratio was 1.012041: a 1.204% worsening instead of the required 10% improvement. The candidate was retired before public Mini scoring.

This negative is narrow. It falsifies the full-strength second-order Gaussian correction under that paired gate; it does not establish that every shrinkage coefficient, higher-order term, or non-Gaussian correction is harmful.

7.2 Scientific stop: partial Kerdock MUB design

A second mechanism was killed by a cheap public-evidence falsifier before candidate construction. A same-family Phase 1 ablation reported raw MSE 1.04775×10^{-6} even with 32 bases. The proposed Phase 2 arm would have only six total bases at 9.6939% measured utilisation, while the score-path gate then required more than a 12-fold adjusted-score improvement. Because the available ablation pointed in the opposite direction, the branch was retired without spending dataset compute. This was a scientific portfolio decision, not evidence about the submitted blend.

7.3 Operational failures and clearing evidence

Table 4: Operational failures and their clearing evidence. No failed run is counted as a negative accuracy result.

Failure	Stop evidence	Clearing evidence
Subprocess throughput	No full JSON before 1,200 s; process stopped at 1,202 s.	Same artifact and dataset completed 100/100 under the preregistered in-process runner and larger cap.
Local runner cap	MLP 0 matched; no aggregate JSON before 360 s; stopped at 387 s.	Full run passed after a bounded runner streamed compact results without changing the estimator.
Seed-protocol mismatch	Shipped-weight reconstruction drifted 1.5671%, above the 1% gate; no curve read.	Official loader registered Phase 2 seed metadata; the first-eight score then reproduced with 0.0000% drift.

The 24.85% result was accepted only after the earlier failures' clearing conditions were met. The Hermite result remains retired because a passing runner cannot reverse a scientific gate failure.

8 Reproducibility and provenance

8.1 Artifact binding

AIcrowd submission	328274; successfully evaluated; live adjusted score 3.001×10^{-7}
Submission tarball	`submission-phase2-covariance-wmc-blend-v1-sanitized.tar.gz`
Tarball SHA-256	`7687d9f8f65ec4bbab47cd18793faf91a5f5ff08fb743065642f4ef4ff9a069e`
Estimator SHA-256	`5d747a91672861526f3407211de85803aef8b61932bd4fd31e56407b57948846`
Full-Mini experiment	`whestbench-phase2-blend-full-mini-v1`
Full-Mini run	`run-93e25793cb73ed414ef71e14`
Weight-response experiment	`whestbench-phase2-blend-weight-response-v2`
Weight-response run	`run-10a0c9f01dcf2490a287308d`
Dataset	`aicrowd/arc-whestbench-public-2026`, revision `v2-phase2`, split `mini`

The tarball contains exactly `estimator.py` plus its generated manifest. The manifest's estimator digest matches the reviewed candidate digest above. The receipt ledger binds that tarball digest to the one Phase 2 submission named in this report.

8.2 Code lineage

The gain-covariance branch adapts AIcrowd's MIT-licensed `examples/03_covariance_propagation.py` at starter-kit commit `35164f3535c2103742772dd11a68ce840464db0b`. The whitened antithetic Monte Carlo branch adapts Bin Yong Bong's MIT-licensed public `estimators/wmc.py` at commit `f58d8a9904c57c91a32ff5a5dc0588635aeb8128`. The submitted file retains both copyright notices and the MIT permission text.

The claimed contribution is limited to the frozen Phase 2 combination and its measured error geometry. The component algorithms and their implementation devices are excluded. The ARC companion paper supplies the broader mechanistic-estimation setting; this report does not reproduce that paper's cumulant estimator.

8.3 Minimal independent check

An independent practitioner with the official dataset and pinned packages can audit the central claim in four steps:

1. verify the tarball and estimator SHA-256 values above;
2. run the official gain-covariance example and the submitted estimator on the same 100 `v2-phase2/mini` MLPs through the official loader;
3. require zero failures and verify adjusted scores $4.050381542 \times 10^{-7}$ and $3.044010045 \times 10^{-7}$, within deterministic runner precision for the pinned seeds;
4. dump both final-layer residual fields once, reconstruct $J(0.75)$, and only then compute r, w^*, and the weight grid.

The seed-protocol incident shows why step 2 must use the competition loader. A generic dataset loader can load the correct weights and ground truth while silently deriving a different estimator seed.

9 LLM use disclosure

Disclosure. The entire development process was LLM-driven under a registered experiment ledger. LLMs were used to select and compare mechanisms, draft and adapt code, specify hypotheses and falsifiers, design preflights and stop conditions, diagnose runner and seed failures, analyse the branch residuals, and generate this technical write-up. The human owner set the prize objective, approved lifting the project Hold, and retained control of external submission. This was not limited to editing assistance: the substantive research and writing workflow was LLM-driven.

The numerical claims were not accepted from free-form model output. They were checked against finished structured records, exact file hashes, receipt binding, and registered run identifiers. The LLM had no access to the private evaluation suite or grader internals. No hidden result was inferred from the public leaderboard.

The explanation of why the residual fields are weakly aligned remains explicitly marked unvalidated. The quadratic blend identity, the measured correlation, the optimum-weight calculation, and the score comparisons are validated on the named public cohort. Readers should separate those facts from the causal narrative.

10 Limitations

- **Public-cohort reuse.** The full score and the weight response use the same 100 public Mini MLPs. The fixed 0.75 weight predates both measurements, but the 0.7341 optimum is descriptive and must not be treated as independently validated selection.
- **Private generalisation is unknown.** One live score agrees with local Mini within 1.433%, but neither result establishes private-suite performance, prize rank, or award probability.
- **One stochastic realisation.** The corrected weight curve uses the official seed protocol for the recorded artifact. The failed protocol-2 run moved the shipped-weight score by about 1.6%, showing that fine distinctions below that scale should not be generalised across seeds. The 0.118% optimum gap is exact for the recorded curve, not a claim of population precision.
- **Aggregate archive.** The durable run record retains aggregate scores and gates but not a per-MLP uncertainty interval. The 24.85% comparison is a paired fixed-cohort measurement, not an estimated confidence bound.
- **Component novelty.** Covariance propagation, antithetic pairing, whitening, and fused first-layer multiplication have public antecedents. The work does not claim them as new.
- **Causal explanation.** Low residual correlation validates complementarity on this cohort. It does not by itself prove which approximation error or distributional feature created that complementarity.

11 Conclusion

A weak component can improve a strong component when their errors point in different directions. In this Phase 2 artifact, whitened antithetic Monte Carlo alone had 2.652 times the adjusted MSE of gain-covariance propagation, yet a fixed 25% share reduced the analytic baseline's full-Mini error by 24.85%. The result survived the move from an optimistic eight-MLP gate to all 100 MLPs, stayed below the compute floor with zero failures, and matched one live evaluation within 1.43%.

The blend score rests on the chain around it: a weight fixed before measurement, a completed 100-MLP paired comparison, nearly orthogonal recorded residuals, a closed-form optimum that matches the grid optimum, an operational falsifier that caught the wrong seed protocol before the curve was read, and a scientific negative that was not relabelled as a runner problem. That chain supports a limited conclusion: complementary analytic

and stochastic errors are a measurable resource for this estimator family. Whether the same geometry persists on fresh private MLPs remains open.

References

[1] AIcrowd, "Phase 2 of the ARC White-Box Estimation Challenge is live," 23 August 2026. Phase 2 dimensions, compute rules, dataset revision, and 24 October write-up deadline. Official Phase 2 launch post.

[2] Alignment Research Center and AIcrowd, "Algorithmic contribution prize guidelines: How ARC judges these prizes - discretion, technical writeups and LLM usage," updated 6 July 2026. Official algorithmic-contribution guidelines.

[3] AIcrowd, "ARC White-Box Estimation Challenge 2026 - Challenge Rules." Official challenge rules.

[4] AIcrowd, `whest-starterkit`, commit `35164f3`; official covariance-propagation example and Phase 2 harness. Pinned starter-kit source.

[5] W. Wu, V. Lecomte, M. Winer, G. Robinson, J. Hilton, and P. Christiano, "Estimating the expected output of wide random MLPs more efficiently than sampling," arXiv:2605.05179, 2026. arXiv record.

[6] B. Y. Bong, `Oishi1029/arc-whestbench-2026`, commit `f58d8a9`; public MIT-licensed whitened Monte Carlo implementation. Pinned public source.

[7] S. Nygaard, `ARC-Whitebox`, commit `73d43d4`; public same-family basis-count evidence used for the Kerdock stop decision. Pinned public evidence.

[8] AIcrowd, `arc-whestbench-public-2026`, revision `v2-phase2`. Official Phase 2 dataset.

A Complete experiment and negative-result map

All identifiers in the first column below share the prefix `whestbench-phase2-` unless stated otherwise.

Registry ID	Final state	Report role
`official-baseline-v1`	Passed	Package and environment gate.
`covariance-gate-v1`	Passed	Generated-MLP wiring gate.
`covariance-mini-v1`	Passed	Full covariance baseline on 100 public Mini MLPs.
`hermite2-gate-v1`	Scientific failure	Paired Hermite-2 negative.
`covariance-wmc-blend-v1`	Operational failure	First-eight pass; subprocess cap before full output.
`covariance-wmc-blend-local-holdout-v1`	Operational failure	Exact MLP-0 parity; local cap before aggregate output.
`blend-full-mini-v1`	Passed	Accepted 100-MLP candidate result.
`blend-weight-response-v1`	Operational failure	Seed mismatch caught before curve access.
`blend-weight-response-v2`	Passed	Corrected sensitivity analysis and residual geometry.

The negative-results registry contains five WhesTBench entries: the Hermite-2 scientific negative, two runner-throughput operational failures, the partial-Kerdock scientific stop, and the seed-protocol operational failure. The passed retries do not erase those records; they explain which failure conditions were cleared.

B Submission mapping checklist

Requirement	Binding in this write-up
One PDF technical write-up	This document.
Exactly one Phase 2 submission ID	328274.
Successfully evaluated submission	Receipt-bound live result 3.001×10^{-7}.
Algorithmic approach	Section 3.
How it was developed	Section 4 and Table 1.
Evidence of actual impact	Sections 5–6 and Tables 2–3.
Negative results and ablations	Sections 6 and 7.
LLM disclosure	Section 9.
Uncertainty labels	Sections 1, 6, and 10.