Correcting a Budget-Matched Ising Critical-Temperature Comparison: A Bounded Negative and Diagnostic Study

Venue-independent general empirical research report; author and affiliation information await human confirmation.

Abstract

Computational estimates of a phase transition can be invalidated before any statistical comparison is made if a random-number call sequence restricts the support of the simulated updates. We audit and correct such a failure in a two-dimensional square-lattice Ising workflow, then compare a locked baseline Binder-crossing estimator with a nearest-anchor single-histogram reweighting candidate under a fixed development budget. The legacy Metropolis sequence is reproduced at full call order: for the audited even sizes $L \in \{16, 32, 64\}$, only L of L^2 coordinate pairs are reachable, and a proposed one-update pair-index repair remains parity constrained because the unconditional acceptance draw advances the same generator. Independent PCG64DXSM coordinate and decision streams pass support, replay, state-update, and exact $L = 2$ checks within the audited implementation. In the corrected eight-seed development block, each sampler completes all 440 cells, but neither route passes the joint accuracy, stability, coverage, and cost gate; the gate's MAE component is computed after estimates are frozen against the external scoring reference. The reweighting candidate reduces mean absolute error from 0.087335 to 0.032601 for Wolff, while its half-split coverage is incomplete; for Metropolis, the candidate changes mean absolute error from 0.056765 to 0.055961 and also fails coverage and stability clauses. No held-out seed is used because neither route reaches the predeclared gate. The result is therefore a bounded negative/diagnostic finding about the audited implementation, tested estimator pair, and development block, not a sampler recommendation or a general claim about random-number generators.

1 Introduction

Monte Carlo studies of critical phenomena combine several validity problems that are easy to conflate. The Markov update must explore the intended state space, the saved trace must retain enough support for any reweighting step, finite-size comparisons must be interpreted as diagnostics rather than hidden corrections, and serial dependence must not be mistaken for independent replication. A numerical estimate can look plausible while failing one of these checks. This is particularly consequential when a comparison is made under a fixed CPU budget: a faster or more accurate-looking route is not useful if its estimates are not defined for the same development cases.

The square-lattice Ising model provides a compact test bed for this problem. In zero field, its critical temperature is known exactly from the solution of the two-dimensional model [Onsager, 1944], so a frozen estimate can be scored against an external reference without using the reference to fit or tune the estimator. That frozen-reference score is nevertheless an input to the predeclared finalist gate; the reference is excluded from estimator construction and tuning, but it participates in gate classification and the decision about whether held-out confirmation is triggered. The model also supports both local Metropolis updates [Metropolis et al., 1953] and collective Wolff updates [Wolff, 1989], making it possible to distinguish a sampler implementation problem from a reweighting or finite-size problem. Histogram reweighting is powerful near a simulated anchor temperature [Ferrenberg and Swendsen, 1988], but its accuracy depends on finite energy support, statistical dependence, and the relationship between numerator and denominator fluctuations [Ferrenberg et al., 1995, Newman and Palmer, 1999].

This report answers a narrow question: within the audited implementation, does replacing a coupled legacy random-number sequence with independent coordinate and decision streams, and then comparing the locked baseline estimator with a nearest-anchor reweighting candidate, yield a route that passes a predeclared joint gate on eight development seeds? The comparison is deliberately bounded. The held-out seed block is reserved and unused because neither corrected route reaches finalist status; unresolved derivative-specific analyses are not silently reconstructed from historical traces.

The contributions are fourfold. First, we give a full-call-order diagnosis of the legacy Metropolis support defect and of a rejected intermediate repair. Second, we document deterministic, exact-small-system, and replay checks for the corrected streams. Third, we report the corrected development comparison as a joint validity decision, retaining invalid splits and negative trials rather than ranking methods by accuracy alone. Fourth, we separate the remaining axes—finite-size drift, serial correlation, reweighting support, seam/root validity, split stability, coverage, and cost—and state which downstream claims remain unresolved.

2 Related work

Ising criticality and finite-size crossings. The two-dimensional zero-field Ising model has an analytic critical point and is therefore a useful benchmark for simulation workflows [Onsager, 1944]. Binder's finite-size scaling work established the importance of distribution-based observables for locating critical behavior [Binder, 1981]. A reproducible computational study by Dolfi et al. [2014] estimates the critical temperature using Binder-cumulant crossings and finite-size scaling, while broader reviews place such crossing and scaling procedures in the context of Monte Carlo phase-transition practice [Binder et al., 2000]. Our estimator is deliberately more modest: the 48–64 crossing pair is a primary development diagnostic, not an extrapolation to the thermodynamic limit.

Monte Carlo dynamics. The Metropolis construction is a canonical local-update method [Metropolis et al., 1953]; the Wolff algorithm grows and flips same-spin clusters and was introduced to reduce critical slowing down near criticality [Wolff, 1989]. The present study does not assume that either algorithm is preferable. Instead, it audits the actual random-number calls and compares their corrected traces under the same lattice, temperature, seed, sampling, and gate rules.

Histogram reweighting and uncertainty. Single-histogram reweighting uses samples drawn at one temperature to estimate observables at nearby temperatures [Ferrenberg and Swendsen, 1988]. Follow-up analyses show that statistical errors, correlations, and finite energy range can each limit the method [Ferrenberg et al., 1995, Newman and Palmer, 1999]. These concerns motivate an explicit effective-sample-size floor and support checks in the candidate estimator, plus separate serial-correlation diagnostics. They also motivate treating a missing root or invalid split as a failed analysis case rather than as a number to be imputed.

Reproducible computational practice. The study follows the broader reproducibility lesson that code, data, parameters, and analysis decisions must be linked rather than summarized only by a final number [Dolfi et al., 2014]. The evidence and consistency workflow used to prepare this report is procedural tooling rather than scientific evidence; its use is disclosed because the corresponding skill library is itself a citable methodological resource [Kassis et al., 2026].

3 Model and sampling methods

3.1 Physical model and observables

We study a ferromagnetic square lattice of side length L with periodic boundary conditions, coupling $J = 1$, Boltzmann constant $k_B = 1$, and zero external field. A configuration is $s = (s_i)_{i=1}^{L^2}$ with $s_i \in \{-1, +1\}$. The Hamiltonian is

$$H(s) = -\sum_{\langle i,j \rangle} s_i s_j, \tag{1}$$

where each undirected nearest-neighbor bond is counted once. We record the magnetization density and energy density,

$$m = \frac{1}{L^2} \sum_i s_i, \qquad e = \frac{H(s)}{L^2}. \tag{2}$$

The exact reference used for scoring after estimates are frozen is $T_c^{\mathrm{ref}} = 2/\log(1 + \sqrt{2}) = 2.2691853142\ldots$. It is not an input to the sampler, a repair target, a sampling stopping rule, or a query-grid selection criterion, and it is not used to fit or tune either estimator. The resulting frozen-estimate MAE is, however, a predeclared finalist-gate input, so the reference participates in gate classification and in the decision about whether held-out confirmation is triggered.

3.2 Update kernels and the call-order defect

For a Metropolis proposal, a coordinate stream draws (i, j), the local energy increment is

$$\Delta H = 2s_{ij}\left(s_{i-1,j} + s_{i+1,j} + s_{i,j-1} + s_{i,j+1}\right), \tag{3}$$

and a decision stream supplies one uniform variate for every proposal, including proposals with $\Delta H \leq 0$. The spin is flipped when $\Delta H \leq 0$ or when the decision is below $\exp(-\beta\Delta H)$, where $\beta = 1/T$.

The historical implementation used a 64-bit linear-congruential generator (LCG) for both coordinate and acceptance calls. Let $x_{r+1} = ax_r + c \pmod{2^{64}}$, with the saved constants $a = 6364136223846793005$ and $c = 1442695040888963407$. The full proposal call order was

$$x \xrightarrow{\text{mod } L} i, \qquad x' \xrightarrow{\text{mod } L} j, \qquad x'' \xrightarrow{\text{uniform}} u. \tag{4}$$

For the audited even sizes $L \in \{16, 32, 64\}$, the first two calls imply $j = (ai + c) \bmod L$; the coordinate support therefore contains only L ordered pairs rather than L^2. A proposed repair that drew one state modulo L^2 for the pair and then retained the unconditional acceptance draw was rejected. The next pair-index state is two LCG transitions later, and for even L the resulting pair index and second coordinate have fixed parity. The important point is that call order, not an isolated coordinate helper, determines the support.

The corrected implementation preserves the initial-state construction but creates independent PCG64DXSM generators for coordinate and decision channels. For each sampler, seed, size, and temperature index, the stream seed is a tuple

$$(400004,\ \text{seed},\ L,\ \text{temperature index},\ \text{sampler code},\ \text{channel code}), \tag{5}$$

with channel codes 101 for coordinates and 103 for decisions. Wolff updates use a coordinate stream for the cluster seed and consume one decision variate for each eligible same-spin neighbor in the fixed up/down/left/right order; the variable per-update consumption is recorded rather than hidden.

3.3 Implementation audits

The deterministic audit reproduces the historical support result, evaluates the rejected parity repair under its full call order, and tests the corrected independent streams. For each of $L = 16, 32, 64$, the PCG64DXSM pair audit uses 1,048,576 draws per stream, checks joint and marginal counts against their uniform expectations, and requires a maximum standardized deviation no larger than eight with byte-identical replay. Incremental energy and magnetization are recomputed from the complete spin array on small lattices $L = 2, 3, 4$; spin bounds, periodic neighbors, cluster membership, and replay determinism are checked at the same time.

As a separate correctness check, all 2^{L^2} states are enumerated for $L = 2$ at $\beta \in \{0.25, 0.5\}$. Each sampler is run with 10,000 burn-in updates and 50,000 recorded states. The maximum absolute error between empirical and exact joint probabilities of (H, M) must be at most 0.03. This check tests the implementation on a tiny system; it is not a precision interval for the critical-temperature estimate.

4 Reweighting estimator and evaluation protocol

4.1 Binder observable and two estimators

For a saved trace of magnetizations, the Binder cumulant used in the comparison is

$$U_4(T) = 1 - \frac{\langle m^4\rangle_T}{3\langle m^2\rangle_T^2}. \tag{6}$$

The baseline, called *anchor Binder roots* below, evaluates this observable at the eleven simulated anchors and linearly interpolates each adjacent sign change between two sizes. Its primary estimate is the median of valid roots for the 48–64 pair.

The candidate, called *nearest-anchor reweighting* below, reuses the same saved samples. For an anchor temperature T_0, let n be the number of measurements and let $E_b = L^2 e_b$ be the total energy of sample b. For q query temperatures, the implemented log-weight matrix $W \in \mathbb{R}^{q\times n}$ is

$$W_{ab} = -\left(\frac{1}{T_a} - \frac{1}{T_0}\right) E_b, \qquad a = 1,\ldots,q, \quad b = 1,\ldots,n. \tag{7}$$

The row-normalized weights are computed stably after subtracting the maximum element of each row:

$$P_{ab} = \frac{\exp(W_{ab} - \max_{b'} W_{ab'})}{\sum_{b'=1}^{n} \exp(W_{ab'} - \max_{b''} W_{ab''})}, \qquad P \in \mathbb{R}^{q\times n}. \tag{8}$$

Thus $P(m^{\odot 2}) \in \mathbb{R}^{q\times 1}$ and $P(m^{\odot 4}) \in \mathbb{R}^{q\times 1}$ are the reweighted moments, and Eq. (6) is applied elementwise. The exponent uses total energy, not energy density, and the rows are query temperatures while the columns are samples.

The query grid has 1001 points from 2.0 to 2.5 in increments of 0.0005. Each query is assigned to its nearest anchor; an exact midpoint uses the lower anchor. A query is valid only when the normalized weights, moments, Binder value, and effective sample size are finite and

$$\mathrm{ESS} = \frac{1}{\sum_{b=1}^{n} P_{ab}^2} \geq 0.20n. \tag{9}$$

When the Binder curves for two lattice sizes cross, a candidate root is discarded if the crossing lies at an anchor seam or if either endpoint is invalid for either curve. Remaining roots are linearly interpolated and summarized by their median. This rule makes support and seam failures visible in the result rather than treating them as missing-at-random measurements.

4.2 Splits, diagnostics, and gate

Each estimator is evaluated on the full trace and on its first and second halves. For seed r, the half-split diagnostic is $d_r = |\widehat{T}_{cr,\mathrm{first}} - \widehat{T}_{cr,\mathrm{second}}|$. Absolute error is computed only after a primary estimate is valid and frozen. Cross-seed sample standard deviation, mean half-split shift, finite-size pair drift, and Geyer initial-positive-sequence diagnostics are descriptive summaries; none is relabeled as a confidence interval. The Geyer effective sample size is a serial-correlation diagnostic and is not the reweighting ESS in Eq. (9).

The finalist gate is joint for each sampler. The candidate must have mean absolute error no larger than 0.90 times the baseline and at least 0.02 lower, end-to-end process CPU per development seed no larger than 1.10 times the baseline, all eight development seeds valid for the full and both half splits, full-trace sample standard deviation no larger than 0.90 times baseline, and mean half-split shift no larger than 0.90 times baseline. A route must pass every clause; an accuracy gain cannot compensate for incomplete coverage.

4.3 Development design and resource boundary

The development block uses eight fixed seeds; four separately reserved seeds are not used. For each sampler, the design has five sizes $(16, 24, 32, 48, 64)$, eleven anchors $(2.00, 2.05, \ldots, 2.50)$, 500 saved samples per cell, 200 warm-up updates, and a measurement interval of four updates. This gives $8 \times 5 \times 11 = 440$ cells per sampler. The corrected comparison reuses the locked estimator rules and does not fit or tune them against the reference temperature; after estimates are frozen, reference-based MAE is computed for the predeclared gate.

The final corrected branch completes all development cells. The charged process-CPU accounting for the entire correction trial is 112.531728 seconds, including a recovered failed attempt of 29.158953 seconds and the final branch at 83.372775 seconds, within the single 300-second cap. These accounting records include failed and recovered work; no held-out scoring or paper-generation computation is included.

5 Results

5.1 Call-order diagnosis and implementation checks

The full legacy Metropolis audit reproduces the affine coordinate support at all three even sizes. The rejected pair-index repair also retains a fixed parity once its acceptance draw is included. This establishes a deterministic support

Table 1: Corrected eight-seed development comparison. MAE is mean absolute error against the frozen external reference; SD is the sample standard deviation of the eight full-trace primary estimates; shift is the mean absolute first–second-half difference. CPU is end-to-end process CPU per development seed. Values are saved summaries, not confidence intervals.

Sampler	Estimator	MAE	SD	Shift	Full	First	Second
Metropolis	baseline	0.056765	0.073141	0.071958	8/8	8/8	8/8
Metropolis	candidate	0.055961	0.072448	0.110986	7/8	7/8	5/8
Wolff	baseline	0.087335	0.098444	0.128408	8/8	8/8	8/8
Wolff	candidate	0.032601	0.065161	0.129242	8/8	8/8	7/8

defect in the historical call sequence, but it does not establish a universal property of LCGs or of Metropolis sampling. The corrected PCG64DXSM construction passes the pair-support, marginal, replay, dependency, incremental-state, periodic-boundary, and cluster-consumption checks within the audited implementation.

The exact $L = 2$ audit gives maximum absolute joint-probability errors of 0.013985 and 0.006577 for Metropolis at $\beta = 0.25$ and 0.5, respectively, and 0.001739 and 0.000537 for Wolff. All four values are below the predeclared 0.03 tolerance, and replay produces identical final states. These results support the narrow conclusion that the corrected update paths pass the declared low-level checks; they do not validate the uncomputed downstream derivative tables.

5.2 Corrected development comparison

Table 1 reports the saved gate metrics. The paired full-trace estimates in Fig. 1 show that the candidate can move individual seeds in both directions; the candidate series is not a uniformly shifted version of the baseline. The coverage panel retains invalid seeds rather than dropping them.

For Metropolis, the candidate's MAE decreases by only 0.000804 and its ratio to baseline is approximately 0.986, so it fails both the 0.90 relative and 0.02 absolute accuracy clauses. Its full-trace SD is not at most 0.90 times baseline, its mean half-split shift increases, and its coverage is incomplete: seed 2013 is invalid for the full and first-half estimates, while seeds 2019, 2033, and 2039 are invalid for the second half. The candidate is therefore not a finalist even though its completed development cells match the baseline CPU accounting at 3.443071 seconds per seed.

For Wolff, the candidate reduces MAE by 0.054735 and full-trace SD from 0.098444 to 0.065161, satisfying the accuracy and SD clauses. Its mean half-split shift is 0.129242, slightly above the baseline 0.128408, and seed 2033 lacks a valid second-half estimate. Thus the candidate fails coverage and split stability despite its qualified accuracy/dispersion improvement. Its end-to-end process CPU is 5.496142 seconds per development seed, within the cost rule. Neither sampler passes the joint gate, so the held-out confirmation trigger remains closed.

5.3 Separate diagnostic axes

Figure 2 shows diagnostics that should not be collapsed into a single explanation. Across 440 saved trace cells per sampler, the median Geyer integrated autocorrelation time for energy is 4.593 for Metropolis and 0.950 for Wolff; for m^2 it is 12.641 and 0.762, respectively. The corresponding median serial-correlation ESS values are 54.4 and 263.2 for energy, and 19.8 and 327.9 for m^2. The whiskers in the figure span the saved per-trace minimum and maximum, not a sampling confidence interval. These summaries are consistent with shorter serial correlation in the corrected Wolff traces, but they do not by themselves establish that correlation caused any candidate estimate to be invalid.

The baseline largest-minus-smallest pair drift has mean 0.0282 for Metropolis and 0.0813 for Wolff across eight seeds. Candidate pair-drift summaries are incomplete (three valid seed summaries for Metropolis and seven for Wolff), so they are displayed only as diagnostic points. No pair drift is applied as a finite-size correction. Likewise, the exact-small-system errors in Fig. 2 are implementation checks, not evidence that the 48–64 primary estimator has a formal uncertainty bound.

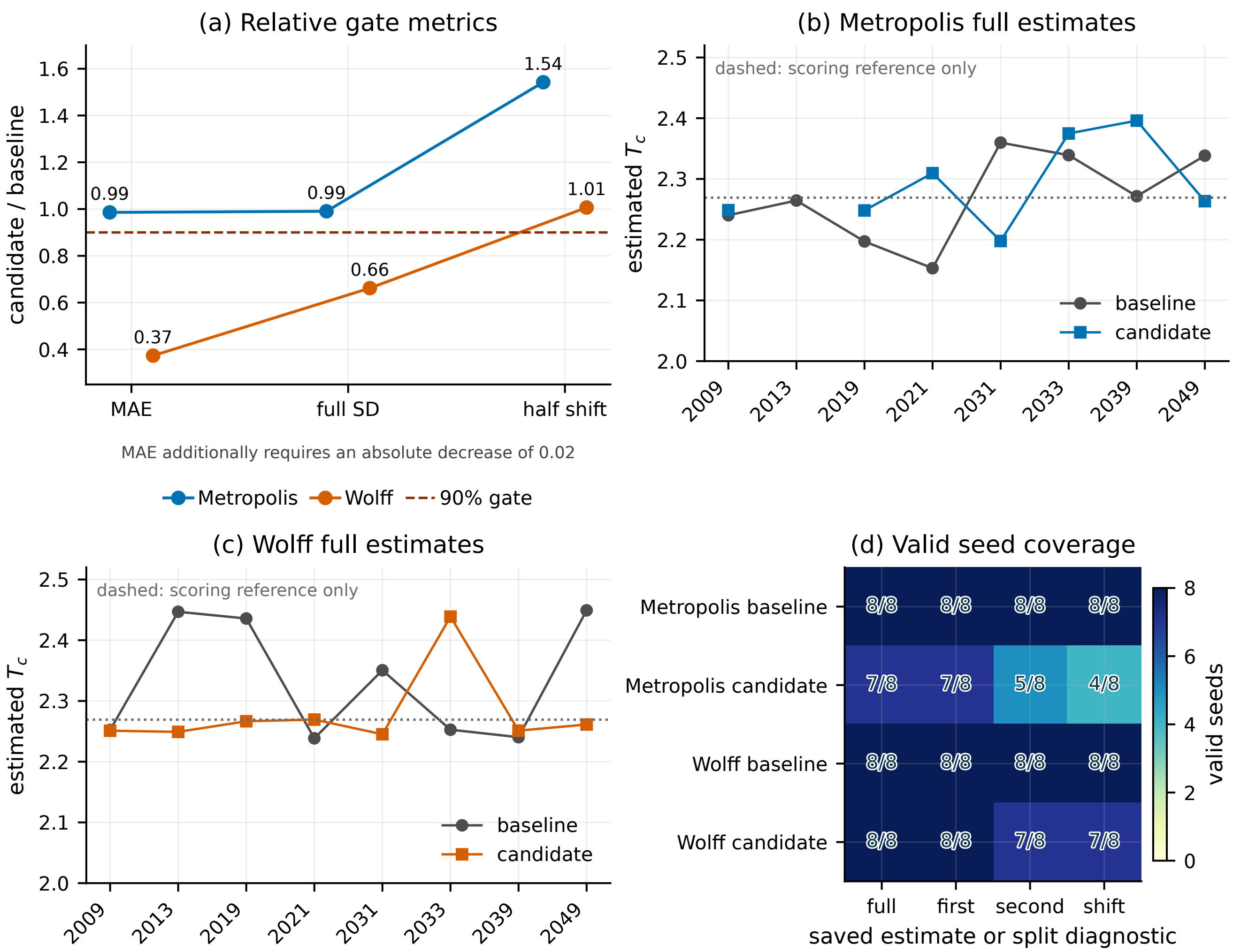

Figure 1: **Development comparison and validity.** (a) Candidate-to-baseline ratios for the three gate metrics; values below the dashed line satisfy the corresponding relative stability/accuracy component, while MAE also has a separate absolute-improvement requirement. (b,c) Paired full-trace primary estimates for the eight development seeds, with missing candidate estimates left missing and the dotted line showing the external scoring reference only. (d) Valid-seed counts for full, first-half, second-half, and half-split diagnostics. Legends and summary notes are kept in annotation strips outside the data axes. The figure is generated from the frozen corrected comparison artifact by the accompanying Python script; no new sampling or uncertainty resampling is used.

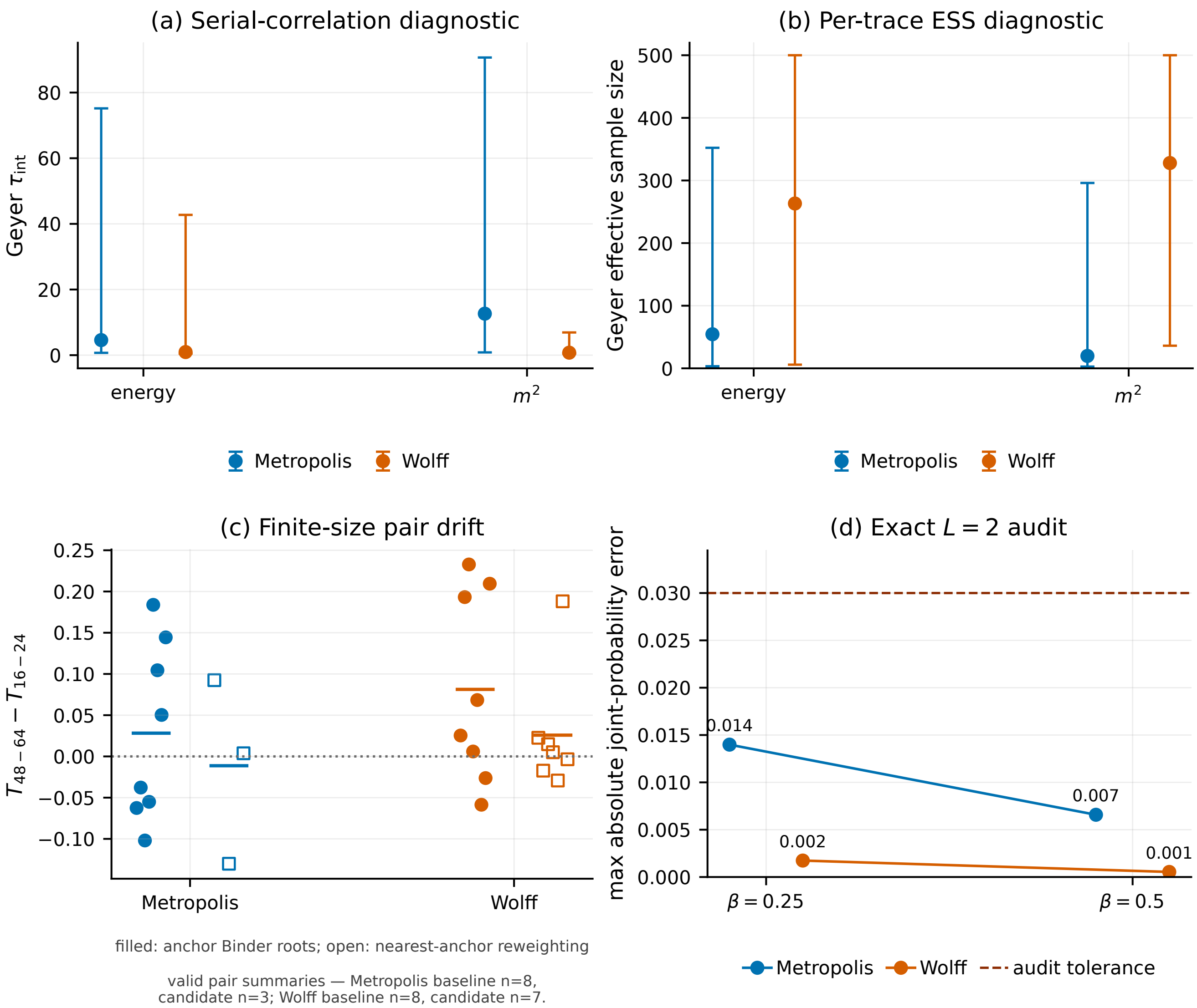

Figure 2: **Implementation and diagnostic axes.** (a) Per-trace Geyer integrated autocorrelation-time medians with saved minimum–maximum ranges for energy and m^2 across 440 cells per sampler. (b) The corresponding serial-correlation effective sample sizes; these are not the reweighting ESS. (c) Per-seed finite-size drift between the 16–24 and 48–64 Binder-root pairs, retained as a diagnostic rather than a correction; open candidate points are shown only when both pair summaries are valid. (d) Maximum absolute error of the exact $L = 2$ joint (H, M) distribution after 50,000 recorded states and 10,000 burn-in updates at each beta; the dashed line is the saved 0.03 audit tolerance. Legends, marker notes, and valid-count summaries are outside the data axes, while point annotations are read from the frozen artifacts.

5.4 Accounting, failed work, and historical status

The corrected comparison includes recovered computation rather than only the successful terminal branch. A prior estimator-recovery failure and partial attempts are preserved in the accounting record; the final report charges the recovered failure once and does not double-count copied prior ledgers. Because the exact technical cause of an unexplained termination is not established by the saved evidence, this report does not assign one. The reproducible scientific result is the final saved comparison and its explicit validity status.

The correction changes the evidentiary status of earlier results. Historical Metropolis/Wolff traces produced by the defective call sequence are not used as current sampler evidence. The earlier sampler comparison is replaced by the corrected development comparison in this report; an affected synthesis is withdrawn; and Trial 5/7 derivative-specific seam, root, support, and split-validity tables, along with synchronized downstream claims, remain unresolved because they were not recomputed in this snapshot. This is a provenance decision, not an assertion that every historical number is numerically wrong for every purpose.

6 Discussion

The central answer is negative but bounded: after the identified call-order defect is corrected and low-level checks pass, neither of the two locked estimator routes reaches the joint finalist gate on the eight-seed development block. The result is stronger than an accuracy-only negative because it retains coverage and split validity as requirements. It is weaker than a general algorithmic conclusion because it concerns one audited implementation, one estimator family, a finite size/temperature design, and eight development seeds.

The Wolff candidate provides a useful partial signal. Its MAE and full-trace dispersion improve substantially relative to the baseline under the same corrected traces, and its serial-correlation diagnostics are smaller in the saved development cells. However, the candidate has an invalid second-half estimate for one seed and fails the mean half-split stability clause. The appropriate interpretation is therefore a qualified partial improvement that motivates a future authorized reanalysis, not a recommendation to use Wolff or nearest-anchor reweighting. The Metropolis candidate provides the complementary warning: a small change in aggregate MAE can coexist with incomplete support/validity and worse split stability.

The call-order finding is methodologically important because a random-number repair can be correct in isolation and still be wrong in context. The unconditional acceptance draw is part of the state-transition schedule; retaining it on a coupled generator changes the next coordinate support. Independent streams remove that particular coupling and make variable Wolff bond consumption observable. The audit does not show that PCG64DXSM is universally superior, nor that the corrected code is free of all possible implementation bugs. It supports only the narrower claim that the declared support, replay, dependency, state-update, and exact-small-system checks passed.

The diagnostic separation also matters. Finite-size pair drift, serial correlation, reweighting support, seam/root validity, split stability, accuracy error, and CPU cost are not interchangeable proxies. The saved evidence does not identify a unique common bottleneck, and the absent corrected Trial 5/7 derivative tables prevent a complete downstream attribution. In particular, no claim is made that a single change in autocorrelation, finite-size drift, or support would explain the invalid candidate cases.

7 Limitations and reproducibility

The development design is intentionally small and fixed: eight seeds, five lattice sizes, eleven anchors, and 500 samples per cell. No held-out confirmation is available because no corrected route passes the finalist gate. The 48–64 pair is a primary comparison diagnostic, not a thermodynamic-limit estimate. Cross-seed SD, half-split shifts, root spreads, and Geyer diagnostics are descriptive summaries; no formal confidence interval, bootstrap, or independent-chain uncertainty calculation is available in the saved evidence.

The most material scientific gap is that corrected Trial 5/7 derivative-specific seam, support, root, and split-validity tables were not recomputed. The report therefore does not claim a complete reclassification of every downstream diagnostic or a single causal explanation for invalid estimates. A statistical-physics or MCMC specialist has not independently reviewed the exact implementation or project-specific thresholds. Finally, the defect was identified by the human owner and several execution/recovery attempts required intervention; this is not presented as unattended autonomous discovery.

All figures in this report are generated from saved numeric artifacts by a relative-path Python script, exported as vector PDF and PNG, and checked for visible renderer bounds and annotation collisions before inclusion. The exact source revision, inputs, quantitative claim mappings, and validation reports are retained with the manuscript package. Rebuilding the document does not rerun sampling; it only regenerates figures, bibliography output, and the PDF.

Declarations pending human confirmation

This anonymous draft intentionally omits authors and affiliations. Human authorship, author contributions, corresponding-author identity, and accountability have not been supplied and await confirmation; the writing agent is not an author. Funding and sponsor-role information were not supplied and await confirmation. A competing-interests declaration awaits confirmation. Ethics and consent applicability/status were not supplied and await human confirmation. Data and code release terms, access restrictions, and any public repository statement await confirmation. AI-assisted drafting, evidence auditing, and figure scripting were used in preparing this draft; human authors must verify the manuscript and confirm the required disclosure and venue-specific wording. Human verification and submission approval remain pending.

8 Conclusion

A full-call-order audit changes the status of an earlier Ising sampler comparison: at the audited sizes $L \in \{16, 32, 64\}$, the legacy Metropolis sequence has restricted coordinate support, and the rejected one-update repair remains parity constrained. Within the audited implementation, independent PCG64DXSM streams pass the declared low-level checks, but the corrected eight-seed development comparison produces no jointly valid finalist under a gate whose accuracy component scores frozen estimates against the external reference. In this development block, Wolff shows a qualified accuracy/dispersion improvement yet fails coverage and split stability, while Metropolis fails accuracy, stability, and coverage clauses. The evidence supports a careful bounded negative/diagnostic conclusion and a reproducibility lesson about call-order-aware validation. It does not support a universal sampler ranking, a unique technical cause, a held-out claim, or a submission-readiness assertion.

A Dimensional and notation checks

The matrix orientation in the reweighting estimator is part of the implementation specification. For one cell, $m, e \in \mathbb{R}^{n\times 1}$ are column vectors, $E = L^2 e \in \mathbb{R}^{n\times 1}$, and $W \in \mathbb{R}^{q\times n}$ has query-temperature rows and sample columns. Row normalization produces $P \in \mathbb{R}^{q\times n}$ with $P\mathbf{1}_n = \mathbf{1}_q$. Therefore

$$P(m \odot m),\ P(m^{\odot 4}) \in \mathbb{R}^{q\times 1}, \qquad U_4 \in \mathbb{R}^{q\times 1}. \tag{10}$$

No transpose is hidden in Eq. (8). The coefficients are row-wise normalized weights, not a column vector of regression coefficients. The Binder ratio is elementwise; the crossing operation compares two $q \times 1$ curves. A candidate root is a scalar temperature after linear interpolation of two adjacent query entries.

For the lattice state, $s \in \mathbb{R}^{L^2\times 1}$, $H \in \mathbb{R}$, and $m \in \mathbb{R}$. The local neighbor sum in Eq. (3) is scalar, so $\Delta H \in \mathbb{R}$. The Geyer diagnostic accepts one scalar trace $x \in \mathbb{R}^{n\times 1}$ and returns scalar τ_{int} and scalar $n/(2\tau_{\text{int}})$. These checks reconcile the equations with the source implementation and do not introduce a textbook estimator in place of the saved one.

B Protocol and evidence boundary

The evidence boundary is part of the result. The corrected comparison is current evidence for the audited implementation and development block. Historical defective traces remain provenance for the diagnosis but are not current sampler evidence. The absent derivative-specific reanalysis is disclosed rather than backfilled, and no later held-out computation is inferred from the existence of reserved seeds.

Table 2: Locked development protocol and current evidence boundary.

Item	Saved protocol or status
Development seeds	2009, 2013, 2019, 2021, 2031, 2033, 2039, 2049
Reserved seeds	2051, 2061, 2063, 2069; unused
Lattice sizes	16, 24, 32, 48, 64
Anchor temperatures	2.00 through 2.50 in increments of 0.05
Samples per cell	500 after 200 warm-up updates; interval 4
Cells per sampler	440
Primary pair	48–64; median of valid crossing roots
Candidate query grid	1001 points; spacing 0.0005
Candidate ESS floor	0.20 of the split sample count
Finalist gate	Accuracy, cost, coverage, full-trace SD, and half-split stability
Held-out confirmation	Not performed; trigger closed after no finalist
Unresolved	Corrected Trial 5/7 derivative-specific tables

References

K. Binder. Finite-size scaling analysis of ising model block distribution functions. *Zeitschrift für Physik B Condensed Matter*, 43(2):119–140, 1981. doi: 10.1007/BF01293604. URL https://doi.org/10.1007/BF01293604.

Kurt Binder, Erik Luijten, Marcus Müller, Nigel B. Wilding, and Henk W. J. Blöte. Monte carlo investigations of phase transitions: Status and perspectives. *Physica A: Statistical Mechanics and its Applications*, 281(1-4): 112–128, 2000. doi: 10.1016/S0378-4371(00)00025-X. URL https://doi.org/10.1016/S0378-4371(00)00025-X.

M. Dolfi, J. Gukelberger, A. Hehn, J. Imriška, K. Pakrouski, T. F. Rønnow, M. Troyer, I. Zintchenko, F. Chirigati, J. Freire, and D. Shasha. A model project for reproducible papers: Critical temperature for the ising model on a square lattice. arXiv:1401.2000, 2014. URL https://arxiv.org/abs/1401.2000. Preprint.

Alan M. Ferrenberg and Robert H. Swendsen. New monte carlo technique for studying phase transitions. *Physical Review Letters*, 61(23):2635–2638, 1988. doi: 10.1103/PhysRevLett.61.2635. URL https://doi.org/10.1103/PhysRevLett.61.2635.

Alan M. Ferrenberg, D. P. Landau, and Robert H. Swendsen. Statistical errors in histogram reweighting. *Physical Review E*, 51(5):5092–5100, 1995. doi: 10.1103/PhysRevE.51.5092. URL https://doi.org/10.1103/PhysRevE.51.5092.

Timothy Kassis, Vinayak Agarwal, Yuhuan He, Darshil Patel, and Aubrey M. Brueckner. Scientific agent skills: A library of procedural knowledge for research agents. arXiv:2609.00065, 2026. URL https://arxiv.org/abs/2609.00065. Preprint.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114. URL https://doi.org/10.1063/1.1699114.

M. E. J. Newman and R. G. Palmer. Error estimation in the histogram monte carlo method. *Journal of Statistical Physics*, 97(5-6):1011–1026, 1999. doi: 10.1023/A:1004614130865. URL https://doi.org/10.1023/A:1004614130865.

Lars Onsager. Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Physical Review*, 65 (3-4):117–149, 1944. doi: 10.1103/PhysRev.65.117. URL https://doi.org/10.1103/PhysRev.65.117.

Ulli Wolff. Collective monte carlo updating for spin systems. *Physical Review Letters*, 62(4):361–364, 1989. doi: 10.1103/PhysRevLett.62.361. URL https://doi.org/10.1103/PhysRevLett.62.361.