

Rare-Event Fidelity in Simulated Thermodynamic Samplers

⌘ Farynth Research
September 29, 2026

Thermodynamic computing aims to reduce the cost of sampling by drawing on the natural fluctuations of physical systems. Realizing this promise requires understanding how the probabilities generated by a sampling process relate to those of the intended model, especially for rare events. We present a computational reference study of rare-event fidelity in simulated thermodynamic samplers. Combining a fixed GPU experiment with exploratory mathematical analysis, we calculate ideal stationary-law references for 133 of 140 conditions and examine how update rules, finite runtime, and diagnostic approval shape reported event probabilities. The resulting cases show how precise, repeatable estimates can describe an altered sampling distribution while missing the intended event probability. Finite-run calculations reveal how transient behavior can mask stationary bias, while independent-sampling calculations show how count-based selection can favor upward fluctuations and reduce interval coverage among approved estimates. These findings motivate evaluating the sampled distribution, the information in a finite run, and the effects of diagnostic selection separately. The accompanying reference probabilities and calculations provide test cases for future evaluation of probabilistic algorithms and thermodynamic hardware.

I. INTRODUCTION

Sampling systems estimate probabilities by generating states and counting how often an event appears. Thermodynamic computing seeks to reduce the cost of this process by using physical fluctuations to generate random states. The usefulness of those states depends on how their distribution relates to the model we intend to study. For a rare event, even a small absolute difference between distributions can change the estimated probability by a large factor.

Establishing that relationship requires more than agreement among repeated runs. Runs may consistently describe a distribution that differs from the intended target, or they may contain too few observations of the event to estimate its probability reliably. The distinction matters whenever a sampling system must provide a useful probability estimate within a finite computing budget.

Convergence diagnostics help assess consistency and sampling information for specified observables. Rank-normalized split $\hat{R}$ and effective sample size (ESS) include localized and tail diagnostics; applying them to an event’s indicator function focuses the assessment on that event [1, 3]. These checks characterize the process that generated the samples. Approximate algorithms and enhanced-sampling methods can satisfy internal checks while estimating a different target [2, 7], which is why accelerator evaluation distinguishes sampling quality, convergence, and goodness of fit [8].

In thermodynamic computing, the link between model and samples depends on couplings, biases, update schedules, and readout. Stochastic programming and thermodynamic compilation already distinguish errors in representing a model from errors in sampling it [9, 10]. Hardware implementations also address response variability, update timing, and randomness generation [18, 19]. Rare-event reference calculations offer a way to examine how such concerns can affect a particular probability, alongside broader measures of sampling behavior.

We examine these questions in small software models whose reference probabilities can be calculated independently. The original experiment compared a policy monitoring energy and magnetization with one monitoring a particular event. Exploratory analysis then connected their recorded outcomes to the probability laws generated by the update rules. This lets us distinguish three questions: which distribution the rule approaches, what a finite run can reveal about that distribution, and how diagnostic approval changes the set of reported estimates.

The resulting computational reference study builds on established theory of approximate sampling, finite-budget error, and selective inference [2, 8, 14–17, 20]. We calculate ideal stationary-law references for 133 of 140 experimental conditions and compare selected references with recorded estimates and intervals. For tractable cases, we also calculate finite-horizon moments and the effects of count-based selection. The contribution lies in these worked comparisons and reusable reference answers; the underlying principles are established, and the study does not validate a replacement diagnostic. The accompanying tables retain every condition, including unavailable laws and unfavorable comparisons.

The paper follows this progression from the intended model to the reported estimate. After defining the experiment and its probability references, Section III examines changes in the sampled distribution. Section IV asks how finite runtime changes the estimate, and Section V examines the further effect of selecting approved runs. The discussion considers how these reference cases could inform future assessment of probabilistic hardware.

II. PROBABILITY REFERENCES AND EXPERIMENTAL DESIGN

A. Using known answers to interpret diagnostic success

Known answers let us examine what a diagnostic pass actually establishes. We use small models whose intended event probabilities can be calculated independently of the sampler. Comparing estimates with those references reveals whether agreement between chains and narrow uncertainty intervals accompany an accurate answer.

The original experiment compared checks of energy and magnetization with checks of the event itself. Energy summarizes a state’s biases and interactions; magnetization measures the balance of positive and negative spins. The event indicator records whether the specified outcome occurred. Checking that indicator focuses the assessment on the desired probability, but consistency in those observations alone cannot establish that the sampler preserves the intended model. We developed the exploratory reference calculations after the experiment to examine that remaining question.

The four model families isolate different parts of this problem (Table I). Independent spins separate individual probabilities from joint occurrences. Coupled spins expose effects of update scheduling and initialization, while a synthetic graph translates joint edge states into connectivity. Their small size permits reference calculations that would often be unavailable in larger applications. These controlled examples do not estimate how frequently the problems occur in deployed systems.

TABLE I. Generated benchmarks. Reference coverage counts the conditions with a justified ideal stationary event probability in the exploratory analysis.

ID	Model, event, and reference coverage
B1	16 independent spins; conjunctions of 3, 4, or 5 specified positive spins, with programmed probabilities $10^{-3}, 10^{-4}, 10^{-5}$. References: 60/60 conditions.
B2	64-spin Curie-Weiss model, $J_{ij} = 1.35/64$ for $i \neq j$, $h_i = -0.04$. Magnetization thresholds 0.5 and 0.75. References: 40/40.
B3	16-spin frustrated Ising model; normalized overlap with a fixed sign pattern at least 0.75. References: 13/20.
B4	14-node, 20-edge synthetic graph; at least one of 5 sinks disconnected from the source. References: 20/20.

B. Separating the sources of an inaccurate estimate

An inaccurate estimate may reflect a changed sampling distribution, limited information in a finite run, or preferential approval of certain results. We separate these possibilities by first examining the update rule, then the length of the run, and finally the diagnostic decision. This connects the observed error to the stages that produced the reported estimate.

We begin with the update rule: how the sampler changes its state from one step to the next. Sequential updates (S1) let each site respond to the current states of its neighbors. Color-group updates (S2) change conditionally independent sites together. With independent redraw randomness, both preserve the corresponding static target distribution. In contrast, simultaneous updates of all sites (S3) or a random half of the sites (S4) allow interacting sites to change together using the preceding state. Comparing these rules asks whether the same local conditional probabilities still produce the same joint distribution when the update schedule changes.

The five settings distinguish changes to the model from changes to the sampling process: unperturbed parameters (P0), rounding to a 0.05 grid (P1), gain changes of -10% or $+10\%$ (P3), and shared random inputs (P4). Rounding and gain change the represented model or conditional response. P4 instead introduces latent-Gaussian correlation $\rho = 0.3$ between simultaneous redraws within each update call; the shared input is independent across chains and renewed at each call. Individual probabilities can remain unchanged while joint probabilities shift. The reference calculations separate these distributional changes from uncertainty in estimating the resulting probabilities.

Repeated runs reveal uncertainty that a single trajectory cannot show. Each replicate contains eight chains, and independent seeds separate replicates. Warmup reduces initialization effects, but the required duration depends on the dynamics. Because the experimental grid assigns different budgets to different conditions, it cannot isolate the effect of runtime. Our finite-horizon calculations instead vary runtime within a fixed mathematical rule.

Approval and accuracy answer different questions. The diagnostics assess the sampled observations; comparison with the known intended probability q_0 determines whether an estimate $\hat{q}$ is materially wrong. For positive estimates, the executed criterion classifies values below $q_0/2$ or above $2q_0$ as material errors. Appendix A 1 specifies the 35,840-run design, diagnostic thresholds, and separate treatment of zero estimates.

We report both false assurance and error among approved runs because withholding answers changes their interpretation. False assurance counts approved but inaccurate estimates as a fraction of all runs. Error among approved runs counts them as a fraction of accepted answers. A stricter policy can reduce false assurance simply by returning fewer answers, even if the answers it accepts are no more reliable.

III. PRECISE ESTIMATES CAN DESCRIBE AN ALTERED DISTRIBUTION

A. The stationary reference explains repeated agreement

The frustrated-spin model B3 shows why a passing estimate needs an independent reference. Under each of the four static simultaneous-update settings, every run passes both diagnostic policies and every estimate fails the intended-target factor-two criterion. Those outcomes establish false assurance. Calculating the distribution generated by the update rule helps explain what the apparently successful runs are estimating.

The independently calculated stationary law supplies that explanation. A stationary law is the distribution left unchanged by continued application of a sampling rule. For simultaneous heat-bath redraws with symmetric interactions, that law is known and generally differs from the programmed Gibbs distribution [14]. Applying it to this small model gives an event probability of 0.00129470 in the unperturbed condition. The observed mean, 0.00129950, is close to this value, while the intended probability is 0.00517166. Across all four static settings, the recorded means lie within 0.38% of their stationary references.

The estimates closely track the implemented rule's probability while missing the intended model's probability. Their repeated agreement is therefore consistent with accurate estimation of an altered distribution. The stationary reference explains the direction and scale of the discrepancy; passing the diagnostics alone does not establish preservation of the intended target. Appendix A 2 gives the stationary expression and case-level comparisons.

B. Correct individual probabilities can conceal incorrect joint events

The independent-spin model isolates a related problem at the component level. Each of five relevant spins is positive with probability 0.1. If the spins are independent, the probability that all five are positive together is 10^{-5} . The five individual probabilities alone therefore do not determine the rare-event probability; the product calculation additionally assumes independence.

Shared randomness changes how often the spins are positive together, even though each spin remains positive with probability 0.1. Under a full P4 redraw, the calculated five-spin event probability rises to 0.00135257, about 135 times its intended value. The change is smaller for three- and four-spin events and depends on which sites update together. Appendices A 2 and B give the specified correlation mechanism and its joint-probability calculations.

The recorded estimates reflect the changed joint law. In all nine B1 P4 conditions using S2, S3, or S4, the event policy approves every estimate despite intended-target material error. For the five-spin S3 condition, 243 of 256 nominal 95% intervals contain the calculated stationary probability, while none contains the intended probability. Narrow uncertainty about the sampled law can therefore coexist with a large discrepancy in the event the model was meant to describe.

The same distinction persists when spins interact. In B2, symmetry allows us to calculate the dynamics of the positive-spin count without enumerating every configuration. The positive and deep-positive events require at least 48 and 56 of the 64 spins to be positive. Under correlated S3 and S4 updates, their stationary probabilities are 10.9–20.9 times the intended values, and the recorded means broadly follow those shifts. Figure 1 compares the same stored intervals against both references: they frequently contain the sampled-law probability but never contain the intended probability.

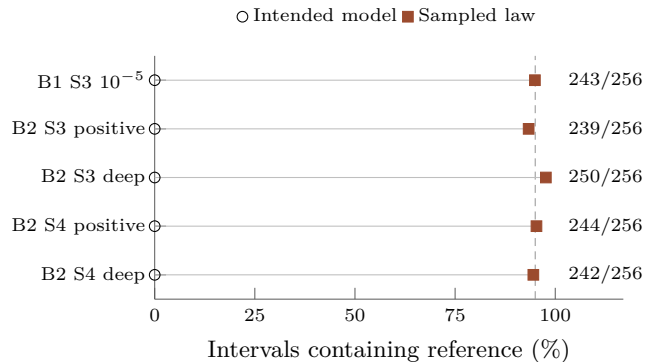


FIG. 1. The reference changes what coverage means. Each row uses the same 256 stored nominal 95% event intervals under correlated updates (P4). None contains the intended probability (open circles); 239–250 contain the independently calculated stationary kernel probability (filled squares; counts at right). The dashed line marks nominal 95% coverage. B1 uses the 10^{-5} conjunction; B2 uses positive and deep-positive basins. These empirical frequencies describe the five conditions, not a general calibration guarantee.

The synthetic graph gives a system-level version of this distinction. With full correlated redraws, its calculated probability of losing connectivity to at least one sink increases 4.96-fold, while the expected disconnected-sink fraction increases 6.10-fold. The first quantity asks whether any sink loses service; the second measures the average fraction affected. Both depend on joint edge states, so individual edge probabilities alone do not characterize either outcome. These calculations concern a generated graph of available and unavailable edges, not measured traffic, outage duration, or physical consequences.

C. Matching mean energy does not recover the rare events

Correlated B2 sampling might resemble the intended model at a different temperature. If that interpretation held, adjusting one parameter could account for the changed event probabilities. We test it within the Gibbs family, keeping the original energy function fixed and varying its dimensionless inverse-temperature parameter β .

We fit that parameter by matching mean energy. Within this restricted family, the resulting distribution is also the unique closest fit in forward Kullback–Leibler divergence, a measure of distributional discrepancy. This gives the temperature interpretation a stronger test than an arbitrary parameter adjustment (Appendix A 2).

Even this fitted distribution misses the rare-event probabilities. The S3 and S4 fits yield $\beta = 0.984798$ and $\beta = 0.969039$, yet the correlated-update event probabilities remain 8.8–20.4 times the fitted Gibbs values (Fig. 2). Matching mean energy therefore does not recover the frequency of the rare configurations. The result excludes exact representation by a scalar temperature adjustment within this fixed-energy family; richer effective models remain possible. Here β is a model parameter, not a measured physical temperature.

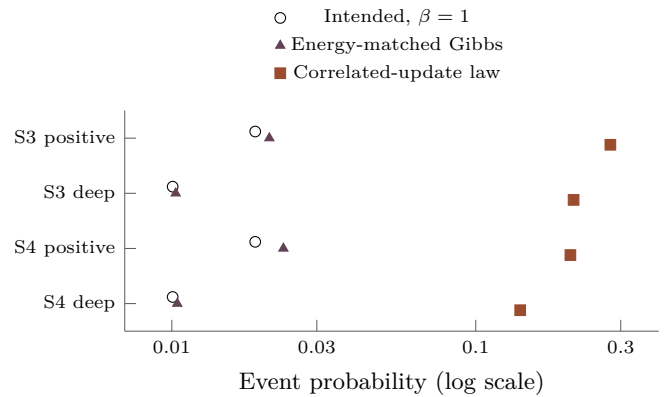


FIG. 2. Mean-energy calibration does not recover the event probabilities in this fixed-Hamiltonian family. For correlated B2 updates, matching mean energy changes the Gibbs inverse temperature from one to 0.984798 (S3) or 0.969039 (S4). The stationary kernel probabilities remain 8.80–20.38 times their energy-matched Gibbs counterparts. Each row compares the intended Gibbs law, the energy-matched Gibbs law, and the calculated P4 kernel law. The parameter β is dimensionless; this comparison neither measures physical temperature nor excludes models with changed biases or interactions.

Aggregate observables still contain useful information. In independent-spin full-update P4 controls, the variances of energy and magnetization change even though their means do not. Comparing those variances with known target values across repeated runs would use evidence beyond the per-run checks evaluated here. Our result supports testing joint events alongside aggregate observables; it does not imply that all aggregate checks must fail.

Together, the reference calculations cover 133 of the 140 conditions. Within that supported subset, 9,008 of the 9,009 event-policy approvals that miss the intended factor-two criterion are within a factor of two of their ideal stationary reference. This broad proximity supports examining the update law when interpreting false assurance, but is not a convergence certificate or a causal attribution of every error. The seven unresolved reference laws all belong to B3 and remain explicitly unavailable.

IV. FINITE RUNS CAN TEMPORARILY CONCEAL STATIONARY BIAS

A. A correct average can result from opposing errors

A finite estimate retains the history of the run. It averages over visited states, so early observations can still matter after the current state has largely lost its dependence on initialization. To distinguish that history from stationary behavior, we calculate the expected estimate at specified run lengths.

For the B2 synchronous sampler with +10% gain, initialization accounts for most of the gap between the observed mean and stationary reference. At the executed budget, the stationary positive-event probability is 0.00008053, whereas the expected retained estimate is 0.08680. The observed mean, 0.08388, lies 0.59 empirical standard errors below that finite-run expectation. Appendix A 3 gives the calculation and corresponding deep-event comparison.

Longer runs change the balance between initialization effects and stationary error. With the rule and warmup fixed, the expected positive-event estimate is 4.61 times the intended probability at 250,000 retained draws, 1.28 times at one million, and 0.132 times at ten million. It crosses the intended value between those horizons because initialization-driven overestimation offsets stationary underestimation. At the crossing, the mean is correct even though both sources of error remain.

That agreement is specific to the event. When the positive-event expectation matches its target, the deep-event expectation remains about 54% high. Choosing a runtime because one known answer looks correct can therefore give misleading reassurance about another. These calculations explain behavior under known references; they do not provide a stopping rule for an unknown probability.

B. A matching mean can still leave large uncertainty

The expectation alone also leaves out differences between repeated runs. To assess the accuracy of an individual estimate, we combine its variance with its squared bias relative to the intended probability. Let $q_{W,T}$ be the expected estimate after warmup W and T retained draws per chain. Under eight ideal independent chains, if S_T is one chain's event count, the decomposition is

$$\begin{aligned} \text{Var}(\hat{q}) &= \frac{\text{Var}(S_T)}{8T^2}, \\ \text{MSE}_{q_0}(\hat{q}) &= \text{Var}(\hat{q}) + (q_{W,T} - q_0)^2. \end{aligned} \quad (1)$$

These moments include within-chain dependence and predict behavior of the mathematical kernels at the specified budgets. They allow runtime comparisons while holding the transition rule fixed. Standard Monte Carlo error theory concerns the sampling law and its observables; finite-budget bias-variance tradeoffs for approximate chains are established [15, 16, 20].

At the S3 positive-event mean-matching horizon, the predicted standard deviation is about 0.99 times the intended probability. At S4's corresponding horizons, standard deviations exceed twice the target probabilities. The average can therefore be correct while individual runs remain highly variable. Locating cancellation of bias alone does not identify a reliable estimate.

Including variance changes which runtime looks favorable (Fig. 3). For the synchronous positive event, mean squared error keeps falling after the mean crosses its target: decreasing variance initially outweighs increasing squared bias. It later rises as the estimate approaches the biased stationary answer. For the deep event under the same rule, mean squared error continues to fall over the evaluated horizons. The value of additional computation therefore depends on the event and error measure, consistent with established finite-budget analyses [15, 16, 20].

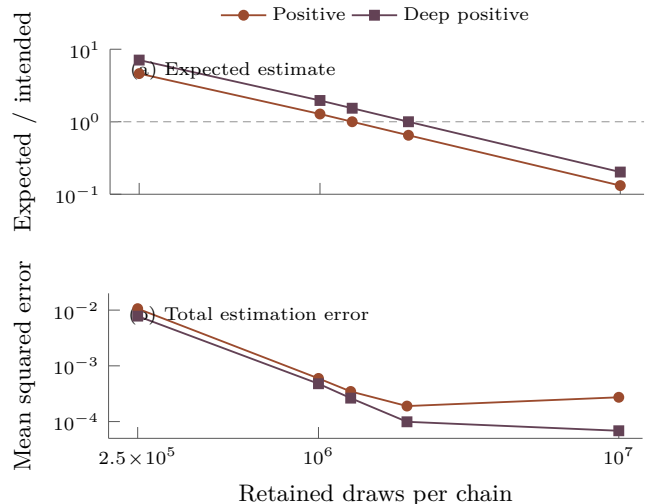


FIG. 3. A matching mean does not identify a useful stopping time. Calculated finite-horizon behavior of B2 synchronous updates with +10% gain, 25,000 warmup sweeps, fair-spin initialization and eight ideal independent chains. (a) Mean estimate divided by the intended probability; the dashed line is one. (b) Total MSE, including variance and squared bias. For the positive event MSE falls past the mean match, then rises at 10^7 draws; for the deep event it continues to fall over the evaluated grid. Points show model calculations, with connecting lines as visual guides. Both axes are logarithmic.

C. Observed variability leaves an unresolved discrepancy

The ideal calculations explain the mean more completely than the variability. For S3 with +10% gain, observed standard deviations agree with prediction within 1.9% and 0.8% for the two events. For the S4 positive event under the same gain change, the observed standard deviation is 28.9% higher than predicted. Agreement in the mean therefore leaves part of the recorded run-to-run behavior unexplained.

Two outcomes contribute heavily to that discrepancy. Their chains spend very different fractions of time in the event, and neither run earns event-policy approval. Appendix A 3 quantifies their influence, but its cause remains unresolved. We retain both outcomes because the available evidence gives no basis for excluding them.

V. APPROVAL CHANGES THE REPORTED SAMPLE

A. A sensible count requirement can select upward fluctuations

Diagnostic approval selects which estimates remain available to report. Requiring enough observations of a rare event addresses sparse information, but the same observed count also determines the probability estimate. Selecting runs by that count can therefore change the distribution of accepted estimates, even when the underlying sampling law is correct.

Quantized B2 isolates this selection effect under independent sampling. Rounding removes its interactions and changes its biases, giving the realized model a positive-event probability of about 6.91628×10^{-6} . Ideal S1–S3 full redraws produce independent samples from that model. We use its realized probability as the reference so that the comparison measures selection separately from quantization. Two million draws yield an expected 13.83 event occurrences, below the twenty-hit threshold in the simplified count gate we examine first.

The twenty-hit gate favors runs that happen to see the event unusually often. Among runs meeting that gate, the calculated mean estimate is 55.0% above the realized probability. This upward shift follows from selecting the upper tail of the count distribution; it requires neither dependent samples nor an incorrect stationary law. Appendix A 4 gives the binomial identity, and Appendix C adds the requirement that every chain observe the event.

Selection changes interval coverage for the same reason. Before selection, the ordinary Wilson intervals contain the realized probability with calculated frequency 95.85%. Retaining only counts of at least twenty lowers that frequency to 63.10%; also requiring a hit in every chain lowers it to 58.39% (Fig. 4). Intervals that would have contained the reference are disproportionately removed with the lower-count runs. A nominal 95% interval for an unselected sample need not retain that coverage among approved estimates [17].

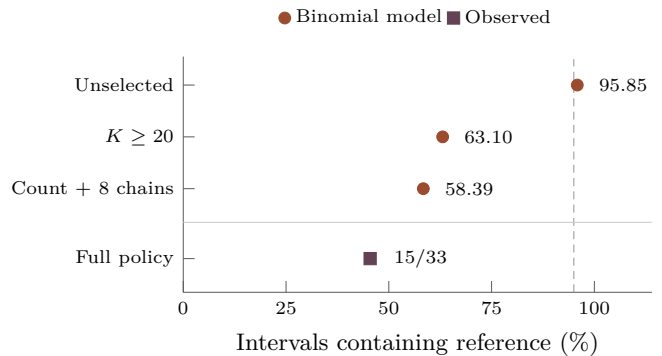


FIG. 4. Selection can reduce interval coverage even under independent redraws. For quantized B2, the finite-binomial calculation uses two million independent draws with $q \simeq 6.91628 \times 10^{-6}$. Ordinary Wilson coverage is 95.85% unselected, 63.10% after $K \geq 20$, and 58.39% after also requiring a hit in all eight chains. The separated square is empirical: 15 of 33 full-policy approvals across S1–S3 have stored intervals containing the actual probability (45.45%). The analytical count gates do not model the full policy, whose intervals use estimated ESS. The dashed line marks nominal 95%.

B. Selection also appears under the full policy

The full diagnostic policy adds ESS, R-hat, and other requirements to the count checks. In the retained S1–S3 quantized B2 conditions, its approved set happens to match the count-plus-chain set: 33 of 768 runs pass. Their mean estimate is 62.8% above the realized probability, and only 15 of the 33 stored intervals contain it. These observations make selection relevant to the empirical result, while leaving the analytical count rule and complete policy distinct.

Figure 4 separates the analytical and empirical results because they use different procedures. The calculated 58.39% describes independent-binomial samples passing the specified count gates. The observed 45.45% comes from 33 full-policy approvals with intervals based on estimated ESS. Their difference cannot be attributed to one cause by this comparison. The unperturbed B1 case shows a weaker selection effect; temporally dependent S4 falls outside the binomial derivation (Appendix A 4).

C. Fewer wrong approvals must be weighed against fewer answers

Event-specific checks reduce wrong approvals in the original experiment, but also withhold more answers. Under the executed factor-two criterion, conventional checks approve 35,344 runs, including 15,504 material errors; event-specific checks approve 25,750, including 9,265 errors. The event policy has a lower observed false-assurance rate in 43 conditions and a higher rate in none. To interpret that reduction, we also examine how many answers remain and how accurate they are.

Figure 5 shows where the reduction occurs. The event policy removes many wrong approvals in parts of B1 and B2, while the B3 simultaneous-update block and several B4 settings remain dark under both policies. Because the colors count wrong approvals across all runs, a pale cell can also reflect a policy withholding answers. The comparison below separates that reduction in output from the selection of more accurate estimates.

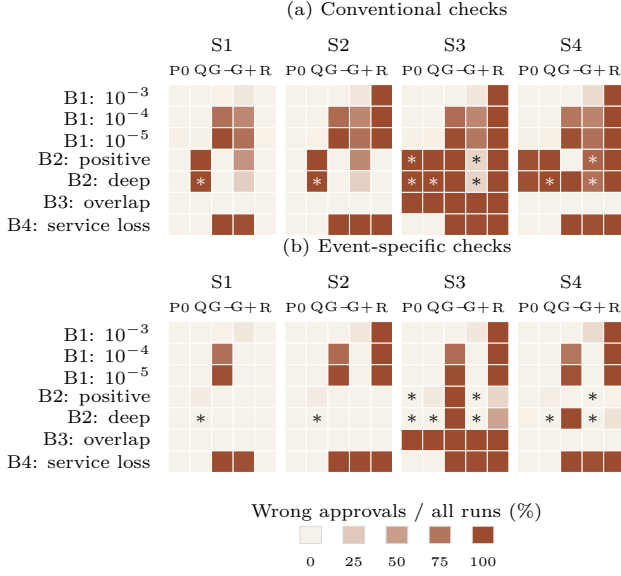


FIG. 5. Where each policy gives false assurance across the full experimental grid. Each cell shows approved estimates with executed factor-two material error as a percentage of all 256 runs in that condition. Both panels share a linear scale; darker copper means more wrong approvals. Rows show the seven events; columns group five settings under each update rule. P0 is unperturbed, Q quantized, G- and G+ the gain changes, and R correlated inputs. Asterisks flag the ten conditions with disputed zero-event failure labels. Pale cells may reflect withheld answers; observed zero counts do not establish zero risk. Table IV retains exact counts and abstentions.

We compare the event policy with a conventional policy that returns the same number of answers in each condition. Retaining conventional approvals uniformly at random to match those counts gives an expected 9,404.25 wrong approvals, compared with 9,265 under the event policy. This comparator already inherits the event policy’s allocation across conditions. It describes selection within that allocation and cannot identify the total value of event information.

The benefit depends on the event mix and error threshold. Outside B2, with a tenfold error tolerance, the event policy removes 3,487 approvals but only four such errors. Error among approved estimates rises from 4.98% to 5.74%. Fewer wrong approvals can therefore coexist with a less reliable set of accepted answers under this criterion. Reporting approval, abstention, and accuracy together exposes that tradeoff. Appendices A 4 and D, and Table IV, retain the full counts and zero-count qualifications.

VI. IMPLICATIONS, LIMITS, AND REUSE

A. The reference determines what successful sampling means

Repeated agreement, a correct average, and fewer wrong approvals each provide useful evidence, but establish different kinds of success. The stationary cases show estimates closely tracking an altered sampling law. The finite-run cases show opposing errors canceling in the mean while individual estimates remain uncertain. The selection cases show that accepting fewer estimates changes both their composition and interval coverage. Reading these outcomes together requires keeping their probability references distinct.

Each reference answers a specific question. The intended probability reveals whether an estimate answers the original problem. The stationary probability helps determine whether error reflects a changed sampling law. Finite-horizon moments describe expected initialization effects and variability, while selection calculations describe what approval changes. These established ideas [2, 8, 17] become concrete comparisons here because we can relate their predictions to retained observations.

The accompanying inventory provides stationary probabilities for 133 conditions and finite-horizon and selection calculations for tractable cases. It identifies the seven unavailable stationary references explicitly. Independent implementations can use these answers to examine their own sampling and reporting procedures; broader benchmark value depends on comparisons beyond this grid and independent reuse.

B. Implications for evaluating thermodynamic hardware

The software cases suggest how component calibration and event validation could complement each other on thermodynamic hardware. Calibrating individual responses characterizes each component’s response to its inputs; known joint events test how components behave together. Physical p-bit research already addresses calibration, randomness generation, and timing [18, 19], while thermodynamic compilation separates representation and sampling residuals [10]. Our cases connect those concerns to the accuracy of particular event probabilities.

A hardware follow-up could calibrate component probabilities, then measure known joint events under supported update schedules. A measured model of device behavior would help distinguish effects of model realization, update dynamics, finite runtime, and selection. Such a comparison must distinguish interacting simultaneous updates from conditionally independent group updates: the latter preserve the corresponding Gibbs target under independent redraw randomness [10, 11].

An accelerator’s energy benefit likewise depends on the probability-estimation task. Assessing that benefit requires a stated accuracy and uncertainty requirement, followed by accounting for the work needed to meet it. Preparation, warmup, readout, checking, and additional runs after rejection all contribute; existing thermodynamic performance models already include multiple workflow costs [13]. Our results motivate connecting those costs to event accuracy. They neither measure physical-device energy savings nor establish the loss of any particular advantage.

C. Evidence boundaries

The scope of these conclusions follows from the experiment’s design. The models are small and generated, the perturbations deliberately specified, and multiple events in one model do not represent independent application domains. The additional reference calculations, selected case studies, mean-crossing horizons, temperature projection, and variance audit were chosen after observing the data. They explain features of this corpus, but do not establish how prevalent those features are elsewhere. No new stopping rule, target-aware diagnostic, compiler, or selective interval procedure was prospectively evaluated.

The reference calculations also leave execution-level questions open. Observed interval containment remains distinct from interval availability and the factor-based point-error criterion; aggregate containment does not establish nominal coverage in every condition or after selection. In a B1 full-update condition, the observed energy and magnetization standard deviations are 11.6% and 15.4% below the ideal correlated-law predictions (Appendix A 3). Independently implemented quadrature, matrix, and binomial calculations strengthen the mathematical references, while the remaining empirical discrepancies require fuller reconstruction of the raw trajectories.

Quantization also changes the energy function evaluated in the trace. Subtracting the original model’s expected energy from that recorded mean compares different observables. We therefore exclude the 28 quantization conditions from the supplementary fixed-observable mean comparison. Appendix A 2 c defines the comparison and reports its results for the remaining 112 conditions, including the shared-noise cases with small first-moment discrepancies and inaccurate event estimates.

The zero-count classification and simultaneous-bootstrap analysis differ from the registered plan, limiting the interpretation of the original uncertainty estimates. Appendix D specifies those differences and the post hoc exact-bound and adjusted-comparison analyses. The latter provide retrospective sensitivity results under their stated assumptions. The study’s probability comparisons do not establish downstream decision-threshold performance, and its resource measurements do not quantify physical thermodynamic efficiency.

VII. CONCLUSION

Precise, repeatable estimates can describe a distribution that differs from the intended model. Our reference cases show how altered update laws produce that discrepancy, how initialization can temporarily conceal it, and how approval changes the estimates that remain. Assessing the sampled distribution, finite-run accuracy, and diagnostic selection separately makes these outcomes easier to interpret together. The retained calculations provide concrete answers for algorithm comparisons and a basis for investigating the same questions on physical probabilistic hardware.

ACKNOWLEDGMENTS

Extropic’s work on stochastic programming and thermodynamic computing helped motivate this study [9–11]. The experiment and custom implementation are the responsibility of Farynth Research.

FUNDING

Farynth Research self-funded this study and conducted the experiments on company-owned hardware.

-
- [1] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter and P.-C. Bürkner. Rank-Normalization, Folding, and Localization: An Improved R-hat for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis* 16(2), 2021. <https://doi.org/10.1214/20-BA1221>
 - [2] B. Kang, J. Hughes and M. Haran. Measuring Sample Quality in Algorithms for Intractable Normalizing Function Problems. *Journal of Machine Learning Research* 25(286), 1-32, 2024. <https://jmlr.org/papers/v25/23-0810.html>
 - [3] A. E. Raftery and S. M. Lewis. How many Iterations in the Gibbs Sampler? *Bayesian Statistics* 4, 763-773, 1992. <https://doi.org/10.1093/oso/9780198522669.003.0053>
 - [4] C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* 26, 404-413, 1934. <https://doi.org/10.1093/biomet/26.4.404>
 - [5] S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 65-70, 1979. <https://www.jstor.org/stable/4615733>
 - [6] D. W. K. Andrews. Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space. *Econometrica* 68, 399-405, 2000. <https://doi.org/10.1111/1468-0262.00114>
 - [7] C. Kang et al. Convergence is not correctness: context-dependent performance of enhanced-sampling methods across biological complexity. *Nature Communications* 17, 6245, 2026. <https://doi.org/10.1038/s41467-026-74165-y>
 - [8] X. Zhang et al. Beyond Application End-Point Results: Quantifying Statistical Robustness of MCMC Accelerators. *arXiv:2003.04223*, 2020. <https://arxiv.org/abs/2003.04223>
 - [9] G. Verdon et al. A Framework for Stochastic Differentiable Programming. *arXiv:2608.01612v2*, 2026. <https://arxiv.org/abs/2608.01612v2>
 - [10] M. Amico et al. Thermalizing Stochastic Programs. *arXiv:2608.01615v2*, 2026. <https://arxiv.org/abs/2608.01615v2>
 - [11] Extropic. THRML: Thermodynamic Hypergraphical Model Library in JAX. Software, version 0.1.4. <https://github.com/extropic-ai/thrml>
 - [12] J. Bradbury et al. JAX: composable transformations of Python+NumPy programs. Software, 2018. <https://github.com/google/jax>
 - [13] D. Melanson et al. Thermodynamic Computing System for AI Applications. *arXiv:2312.04836v1*, 2023. <https://arxiv.org/abs/2312.04836v1>
 - [14] N. S. Skantzios and A. C. C. Coolen. Random field Ising chains with synchronous dynamics. *Journal of Physics A: Mathematical and General* 33, 1841-1855 (2000). <https://www.mbfys.ru.nl/~acoolen/2000/skantzios2000a.pdf>
 - [15] J. M. Flegal and G. L. Jones. Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics* 38(2), 1034-1070 (2010). <https://doi.org/10.1214/09-AOS735>
 - [16] A. Korattikara, Y. Chen and M. Welling. Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget. *Proceedings of the 31st International Conference on Machine Learning*, PMLR 32(1), 181-189 (2014). <https://proceedings.mlr.press/v32/korattikara14.html>
 - [17] W. Fithian, D. Sun and J. Taylor. Optimal Inference After Model Selection. *arXiv:1410.2597v4* (2017). <https://arxiv.org/abs/1410.2597v4>
 - [18] A. Z. Pervaz, B. M. Sutton, L. A. Ghashtasala and K. Y. Camsari. Weighted p-bits for FPGA implementation of probabilistic circuits. *arXiv:1712.04166v3* (2018). <https://arxiv.org/abs/1712.04166v3>
 - [19] J. Kaiser, W. A. Borders, K. Y. Camsari, S. Fukami, H. Ohno and S. Datta. Hardware-Aware In Situ Learning Based on Stochastic Magnetic Tunnel Junctions. *Physical Review Applied* 17, 014016 (2022). <https://doi.org/10.1103/PhysRevApplied.17.014016>
 - [20] J. E. Johndrow, J. C. Mattingly, S. Mukherjee and D. Dunson. Optimal approximating Markov chains for Bayesian inference. *arXiv:1508.03387v3* (2017). <https://arxiv.org/abs/1508.03387v3>

Appendix A: Experimental specification and extended comparisons

This appendix gives the detailed settings, derivations, and case-level comparisons supporting the interpretations in the main text. The original condition counts remain in Table IV; the reference inventory preserves unavailable laws explicitly.

1. Experimental specification

a. Which probability is being estimated?

We distinguish the intended model from the process used to sample it. For binary states $x_i \in \{-1, +1\}$, biases h_i , symmetric couplings J_{ij} with zero diagonal, and inverse temperature β , the programmed target is

$$\begin{aligned} E_0(x) &= - \sum_i h_i x_i - \sum_{i < j} J_{ij} x_i x_j, \\ p_0(x) &= Z^{-1} e^{-\beta E_0(x)}, \quad q_0 = \sum_{x \in A} p_0(x). \end{aligned} \quad (\text{A1})$$

Here A is the specified event, Z normalizes the distribution, and q_0 is the intended event probability. The programmed value $\beta = 1$ fixes a dimensionless model scale; it is not a measured device temperature. In the reliability benchmark, binary variables instead represent independently available edges, with a connectivity predicate defining A .

The intended event probability can change before finite sampling error enters the picture. Altering model parameters can produce a realized static target with probability q_R ; the update rule may then approach a different stationary probability q_∞ . We call the invariant distribution of the mathematical update rule its *ideal stationary law*. That law uses the rule specified by the inspected implementation and ideal random inputs. It is a reference for evaluating execution, rather than a claim that every executed random stream or saved trajectory has been independently reconstructed.

A finite run adds another distinction: its expected estimate can differ from the stationary probability while initialization effects persist. For warmup W and T retained observations per chain, $q_{W,T}$ denotes the expected estimate under the specified initial-state distribution and ideal transition law. The observed estimate is $\hat{q}$. Where these intermediate references are available, their signed differences give

$$\begin{aligned} \hat{q} - q_0 &= (q_R - q_0) + (q_\infty - q_R) \\ &\quad + (q_{W,T} - q_\infty) + (\hat{q} - q_{W,T}). \end{aligned} \quad (\text{A2})$$

The terms distinguish static realization, update-law displacement, finite-run expectation, and the remaining observed discrepancy. This is an accounting identity. The final term can include sampling variation and departures from the ideal execution model; it does not isolate a causal mechanism by itself. Dynamic-noise arms have no saved static realized reference, and we keep their separately calculated stationary references distinct.

b. Models, events, and budgets

The frozen design contains seven events across four generated benchmark families (Table I). Crossing each event with four kernels and five settings gives 140 conditions, each with 256 completed replicates. Independent seed derivation separates replicates; each replicate contains eight chains. The replicate, rather than an individual correlated draw, is the statistical unit in the reported empirical comparisons.

Normalized magnetization is $m = d^{-1} \sum_i x_i$ for d spins. The B2 events require at least 48 or 56 positive spins. B3's overlap event requires at least 14 of 16 signs to match its registered pattern. B4 evaluates connectivity in a generated graph of up/down edges; it includes no traffic, physical infrastructure, or measured outages. Exact programmed references use factorization, magnetization multiplicities, or complete enumeration.

The medium budget retains 50,000 draws per chain after 10,000 warmup sweeps; the long budget retains 250,000 after 25,000. The design assigns 80 conditions to the medium budget and 60 to the long budget, with thinning one. These fixed budgets are confounded with condition choice, so comparisons across them do not identify a runtime effect. The finite-horizon calculations below instead vary the budget within a specified mathematical kernel.

c. Update rules and perturbations

S1 updates every site once in a random permutation; its registered name is random-scan Gibbs. S2 updates conditionally independent graph-color groups. S3 redraws every site simultaneously from the old state. S4 selects a fixed-size random subset, $\lceil d/2 \rceil$ sites per chain, and redraws that subset simultaneously. A retained S4 sweep therefore attempts half as many site updates as a full sweep for these even-sized models; equal retained-state budgets do not imply equal computational work.

The single-site conditional is $\Pr(x_i = +1 \mid x_{-i}) = \sigma[2\beta(h_i + \sum_{j \neq i} J_{ij} x_j)]$, with $\sigma(u) = (1 + e^{-u})^{-1}$. With independent redraw randomness, sequential or correctly colored Gibbs updates preserve the corresponding static target. Simultaneous redraws of interacting sites need not do so. These controls and deliberately approximate rules serve different roles in the experiment.

The five settings are unperturbed (P0), coefficient rounding to a 0.05 grid (P1), gain changes that multiply β by 0.9 or 1.1 (P3), and correlated random inputs (P4). P4 uses $U_i = \Phi(\sqrt{\rho}Z + \sqrt{1-\rho}Z_i)$ within an update call, with $\rho = 0.3$ and ideal independent standard normals. The shared factor is separate for each chain and renewed at each call. The parameter is latent-Gaussian correlation, not generally correlation between spin values, event indicators, or successive retained draws.

Every initial spin is a fair independent sign. This is broad in spin space but does not deliberately place chains in rare basins. S1 and S2 share a random site or color order across chains; their uniform draws remain separate. S3 uses chain-specific random arrays, and S4 also uses independently selected subsets. The eight-chain finite-horizon variance calculations concern S3 and S4 under the stated ideal independence assumptions.

d. Diagnostics and outcome definitions

The conventional policy monitors energy and magnetization. The event policy monitors the binary event indicator. Table II states their requirements. An event-policy pass is defined separately from conventional passage. If its requirements fail, a zero estimate or effective event count below twenty triggers abstention; otherwise it records a failed check.

TABLE II. Executed pass requirements. ESS is effective sample size; MCSE is Monte Carlo standard error. All criteria within a policy must hold.

Policy	Criteria
Conventional	For energy and magnetization: split $\hat{R} \leq 1.01$ after rank normalization; bulk and tail ESS at least 400 (5% and 95% tails).
Event-specific	Event bulk ESS at least 400; $\hat{R} \leq 1.01$; effective event count at least 20; at least one hit per chain; event-rate range at most 0.01; relative MCSE at most 0.5.

Effective event count is $N_{\text{eff}}\hat{q}$. The reported MCSE is $[\hat{q}(1 - \hat{q})/N_{\text{eff}}]^{1/2}$, and the nominal 95% interval is Wilson’s score interval with estimated ESS as its effective sample size. These are implemented diagnostic quantities; their nominal coverage is assessed rather than assumed. Effective count at least twenty already implies relative MCSE at most $\sqrt{1/20} \simeq 0.224$, making the separate 0.5 cutoff redundant once that count requirement holds.

An estimate has executed material error when $\hat{q} < q_0/2$ or $\hat{q} > 2q_0$ for positive estimates. The implementation classifies a zero estimate against a positive reference as failure. False assurance is the joint event of approval and material error, divided by all 256 replicates in a condition. Error among approved runs uses a different denominator. Appendix D explains the planned zero-count rule, its implementation discrepancy, and the separate post hoc uncertainty analysis.

e. Computational methods and analytical scope

The simulations used custom double-precision JAX kernels on one NVIDIA GeForce RTX 3090 Ti under Linux [12]. The computational environment comprised Python 3.12.3, JAX 0.11.1, NumPy 2.5.2, and SciPy 1.18.1.

The analysis includes all 35,840 planned unique replicates. The reference-law, finite-horizon, and selection analyses were developed after observing the simulation results. They provide exploratory explanations of those observations; their agreement with the data requires separate evaluation in independently specified experiments before generalization.

2. Stationary references and additional comparisons

a. Synchronous and correlated-update laws

For symmetric J and independent simultaneous heat-bath redraws, the known stationary law is [14]

$$\pi_{\text{sync}}(x) \propto e^{\beta h^\top x} \prod_i 2 \cosh[\beta(h_i + (Jx)_i)]. \quad (\text{A3})$$

Enumeration of all 65,536 B3 states evaluates this expression for the four static S3 settings. Their observed means are within 0.38% of the stationary references, while all runs pass both policies and fail the intended-target factor-two criterion.

For a full P4 redraw of k relevant B1 spins, conditional independence given the shared normal gives

$$q_k = \int \phi(z) \left[\Phi\left(\frac{\Phi^{-1}(0.1) - \sqrt{\rho}z}{\sqrt{1-\rho}}\right) \right]^k dz. \quad (\text{A4})$$

Here ϕ and Φ denote the standard-normal density and distribution function. At $\rho = 0.3$, the stationary-to-intended probability ratios for three, four, and five spins are 6.94, 28.31, and 135.26. Partial redraws attenuate the shifts; singleton redraws preserve the intended law under this within-call mechanism.

Across the nine B1 P4 conditions using S2–S4, all runs receive event-policy approval and fail against the intended target. Per condition, 237–249 of the 256 saved intervals contain the calculated stationary probability; none contains the intended probability. The five-spin S3 condition has median relative MCSE 1.92% and stationary-reference containment 243/256. The supporting table `b1-correlated.csv` gives all twelve B1 P4 reference predictions, including S1 controls.

B2’s exchangeability permits a 65-state count kernel (Appendix B). Its P4 stationary probabilities for positive and deep events are 0.277284 and 0.209869 under S3, and 0.204855 and 0.140052 under S4. In that order, observed means differ by +0.65, −0.76, −2.03, and −0.81 empirical standard errors of the replicate mean. Energy and magnetization mean residuals have magnitude at most 1.91 such standard errors. These are descriptive residuals, without multiplicity adjustment.

For the same four B2 conditions, stationary-reference interval containment is 239, 250, 244, and 242 out of 256; intended-reference containment is zero throughout. All 1,024 runs pass conventional checks with intended-target material error; event-policy approvals are 43, 121, 1, and 26. With the executed warmup, the ideal P4 kernels’ calculated initialization effects fall below numerical resolution. Figure 1 presents the paired-reference comparison.

b. Projection onto the fixed-Hamiltonian Gibbs family

For a calculated stationary count distribution P and $Q_\beta(k) \propto \binom{64}{k} e^{-\beta E_0(k)}$, the derivative of forward KL divergence is $E_P[E_0] - E_{Q_\beta}[E_0]$ and its second derivative is $\text{Var}_{Q_\beta}(E_0) > 0$. Thus, matching mean energy minimizes divergence within this one-parameter family. The fitted values are $\beta = 0.984798$ for S3 and 0.969039 for S4. Event probabilities remain 8.8–20.4 times the matched-Gibbs values; magnetization and count distributions also differ. The construction concerns this fixed Hamiltonian and does not exclude a richer effective model or measure physical temperature.

c. Inventory coverage and observable comparisons

The stationary inventory covers 133 conditions. The seven unavailable laws are B3’s four static S4 settings and its P4 settings for S2–S4. These contain 1,792 event-policy approvals, including 256 intended-target failures whose stationary proximity remains unclassified. The supported subset contains 23,958 approvals and 9,009 intended-target failures; 9,008 of those failures are within a factor of two of the ideal stationary reference. Every condition retains its eligibility status in `reference-inventory.csv`.

The supplementary first-moment comparison uses the same energy and magnetization functions on both sides. For each observable, we divide the absolute difference between the mean of 256 recorded replicate means and the programmed expectation by the empirical SD of those replicate means. Both discrepancies are within one SD in 43 of 112 eligible nonquantization conditions, and within two or three SDs in 44. At least half the event estimates are materially wrong in twelve conditions at each cutoff. Among the 57 eligible conditions with a majority of wrong estimates, 45 exceed three SDs for at least one observable; twelve do not. The table `global-observable-eligibility.csv` retains all 140 rows and marks the 28 excluded quantization comparisons explicitly.

This descriptive screen is less demanding than a test based on the standard error: one replicate SD equals sixteen standard errors of the mean for 256 independent replicates. It does not establish equality of expectations. For full-update B1 P4, marginal preservation establishes unchanged first moments analytically, while the energy and magnetization variances increase by factors 1.516 and 3.454. Detecting altered dispersion uses known target variances and repeated runs, an information budget beyond the evaluated per-run policy. Correct first and second moments still do not determine a rare joint probability.

3. Finite-horizon calculations and variance discrepancies

a. Retained averages and mean-matching horizons

For a row-stochastic count kernel P , initial count distribution v , and event-indicator column f , the finite-horizon expectation is

$$q_{W,T} = \frac{1}{T} v P^{W+1} \left(\sum_{t=0}^{T-1} P^t \right) f. \quad (\text{A5})$$

The first retained state follows warmup plus one update. Matrix powers and geometric sums evaluate the expectation without new trajectories. The endpoint probability $v P^{W+T} f$ concerns only the final state.

The B2 gain-plus calculations use fair-spin initialization and $W = 25,000$. At the executed $T = 250,000$, S3’s observed positive- and deep-event means differ from their finite-horizon predictions by −0.59 and +0.14 empirical standard errors. At ten million draws, S3’s positive-event endpoint is near stationarity, but its expected retained average remains about 31 times the stationary probability. These comparisons distinguish the recorded average from the final-state distribution.

Mean matching occurs at different horizons for each event and update rule. For S3, the positive event reaches q_0 near 1,280,156 retained draws, with predicted eight-chain SD 0.0186241 ($0.990q_0$); the deep event reaches it near 1,971,255, with SD 0.00995354 ($0.989q_0$). For S4, the corresponding horizons are 24,729 and 43,910, with SDs 0.0389939 ($2.073q_0$) and 0.0216982 ($2.156q_0$). These numerically located crossings use the known target and supply no stopping rule for an unknown probability. The supporting finite-horizon table reports the full horizon set, squared biases, variances, and MSEs.

b. Comparison with retained replicate variability

At the executed budget, S3 gain-plus observed and predicted SDs agree within 1.9% for the positive event and 0.8% for the deep event. For S4's positive event, observed SD is 0.00874203 versus prediction 0.00678180, giving variance ratio 1.662. The ideal fourth moment places this sample variance 3.23 sampling-SD units above prediction. Two of the 256 outcomes account for 35.8% of squared deviations about the empirical mean; both have unequal chain occupancies and fail event-policy approval. Retained-record checks found no duplicate trace hashes, mixed targets, inconsistent normalization, or recorded resumption explaining the difference. All outcomes remain included; its cause requires reconstruction of the original trajectories.

A separate discrepancy appears in B1's three-spin P4 S3 condition. Table III compares the SD of energy and magnetization means across 256 replicates, each using eight chains and 50,000 retained draws per chain. Both SDs fall below the correlated-law prediction despite exceeding the independent-target prediction. The ideal full-redraw calculation uses single-state variance divided by 400,000; fourth moments give the sampling variance of the empirical variance (Appendix B). The standardized residuals describe scale, not Gaussian significance or multiplicity-adjusted evidence. Energy and magnetization are dependent comparisons.

TABLE III. B1 three-spin P4 S3: observed and ideal correlated-law SDs of replicate means. The final column standardizes the variance difference by its model-derived sampling SD.

Observable	Observed SD	Predicted SD	Residual
Energy	0.00253644	0.00286973	-2.47
Magnetization	0.000555974	0.000657075	-3.21

Their mean residuals are +1.99 and +2.24 predicted standard errors. Across the other audited P4 full-update B1/B4 conditions, variance residuals have magnitude at most 1.03 sampling SDs. The P0 controls have observed-to-predicted SD ratios 0.9573–1.0837 and maximum absolute variance residual 1.97. The supporting `observable-information.csv` supplies every condition, observable, denominator, and residual. These comparisons preserve the mismatch while bounding which part of the reference calculation agrees with execution.

4. Selection identities and additional policy comparisons

a. Count-conditioned means and intervals

Quantization rounds B2's coupling $1.35/64 = 0.02109375$ to zero and its field from -0.04 to -0.05 . Its positive-event probability changes from $q_0 = 0.0188131$ to $q_R = 6.91628 \times 10^{-6}$. Under ideal independent full redraws (S1–S3), the eight-chain pooled count has law $K \sim \text{Bin}(N, q_R)$ with $N = 2,000,000$. For the simplified raw-count gate $A_m = \{K \geq m\}$,

$$\frac{\mathbb{E}_q[\hat{q} \mid A_m]}{q} = \frac{\Pr\{\text{Bin}(N-1, q) \geq m-1\}}{\Pr\{\text{Bin}(N, q) \geq m\}}. \quad (\text{A6})$$

At $m = 20$, the selected mean is 55.0% high. This identity describes the count-only rule; selection-aware error control must account for the relevant selection event [17].

The Wilson interval using known N contains q when $(K - Nq)^2 \leq z_{0.975}^2 Nq(1 - q)$. For this B2 reference, the containing counts are seven through twenty-one, of which the gate retains twenty and twenty-one. Exact binomial summation gives containment 95.85% before selection and 63.10% afterward. The summation is exact under the reference model; the Wilson interval has nominal 95% coverage. Requiring a hit in every chain gives a selected mean 56.4% high and coverage 58.39%; for the unperturbed B1 $q = 10^{-5}$ control, selected coverage is 91.55%. Appendix C gives the chain-conditioned calculation.

For the retained S1–S3 quantized B2 runs, the full policy's approved set happens to equal the count-plus-chain set: 33/768 approve, with mean 62.8% above q_R and interval containment 15/33 (45.45%). Saved intervals use estimated ESS; their containment classifications coincide with the known- N classification for these selected records only. B1's corresponding control has 273/768 approvals and containment 252/273 (92.31%). S4 has temporal dependence and different accepted sets, so it is excluded from the independent-binomial derivation.

b. Approval allocation and error denominators

Under the original executed factor-two rule, at least one wrong approval occurs in 81 conventional-policy conditions and 50 event-policy conditions. Every event-policy approval also passes the conventional checks in this corpus, although the policy definitions are separate. Uniform random retention of 25,750 conventional approvals across the whole corpus gives 11,295.50 expected errors. Matching counts separately within each condition gives 9,404.25, exceeding the observed event-policy count by 139.25. The latter comparator inherits the event policy’s allocation across conditions; neither comparison provides a unique causal attribution.

Outside B2 at a tenfold error tolerance, the conventional policy has 1,274 errors among 25,600 approvals, and the event policy has 1,270 among 22,113. Thus, 3,487 fewer approvals remove four such errors while increasing the conditional error fraction from 4.98% to 5.74%. The complete grid remains retained. Appendix F identifies the condition-level approval counts needed alongside Table IV to reconstruct these denominators.

Appendix B: Reference calculations and numerical checks

1. Correlated count kernels

For B2, write $a = 1.35/64$, $h = -0.04$, and let k count positive spins. The positive-redraw probabilities for old positive and negative sites are $p_+(k) = \sigma[2\beta(h + a(2k - 65))]$ and $p_-(k) = \sigma[2\beta(h + a(2k - 63))]$. Conditional on shared normal z , replace each probability p by $u(p, z) = \Phi[(\Phi^{-1}(p) - \sqrt{\rho}z)/\sqrt{1 - \rho}]$.

For S3, the next positive count is the sum of independent conditional binomials with sizes k and $64 - k$ and probabilities $u(p_+, z)$ and $u(p_-, z)$. Integrating over z gives a transition row. For S4, the number r of old positive sites in the selected 32-site subset is hypergeometric; the next count is $k - r$ plus the analogous binomials of sizes r and $32 - r$. Solving $\pi P = \pi$, $\sum_k \pi_k = 1$, gives the ideal stationary count law.

Increasing Gaussian quadrature order from 320 to 480 changes transition entries by less than 7×10^{-16} . An independent composite-Legendre construction, separate Bernoulli convolution, and power iteration reproduce kernels within 5.6×10^{-16} and stationary masses within 1.2×10^{-14} . A four-spin labeled-state calculation separately checks the fixed-subset reduction. These are numerical consistency checks, not certified bounds on physical or pseudorandom execution.

2. Factorized and network references

For partial updates in zero-coupling models, stationary conjunction moments obey a finite subset recursion. Let $M(S)$ be the probability that a specified set of bits has its required values and $F(T)$ the corresponding fresh correlated-redraw probability. Define $H(S, T)$ as the probability that the full refreshed set intersects S exactly in T ; sites outside S may also refresh. Independence of fresh draws from the old state gives $M(S) = \sum_{T \subseteq S} H(S, T)M(S \setminus T)F(T)$. Isolating the $T = \emptyset$ term permits evaluation in increasing subset size.

For B4, complete enumeration of 2^{20} edge states supplies the connectivity predicate and disconnected-sink fraction. Grouping by equal edge probabilities and applying a Boolean polynomial expansion permits substitution of the conjunction moments. Independent graph traversal, polynomial inversion, and quadrature reproduce the event and expected-severity predictions within 4×10^{-18} . The predicted histogram of the number of disconnected sinks is model-derived; only event probability and mean severity have corresponding retained scalar measurements.

3. Finite-horizon moments

For a block of transitions, let A contain transition probabilities, M the conditional first reward-sum moment times terminal-state indicator, and V the analogous second moment. Concatenating blocks (A, M, V) and (B, L, U) gives $(AB, MB + AL, VB + 2ML + AU)$. A one-step block is (P, PD, PD^2) , where D is diagonal in the post-transition event reward. Binary composition evaluates long horizons without simulated trajectories.

Centered rewards reduce subtraction loss when calculating variance. Independent block matrices and exact rational enumeration of small chains check the moment arithmetic. Fourth moments give the sampling variance of the usual unbiased variance estimate across $n = 256$ independent replicates: $[\mu_4 - (n - 3)\sigma^4/(n - 1)]/n$. Standardized empirical variance residuals in this paper use that scale descriptively, without a normal-distribution or simultaneous-inference claim.

Appendix C: Selection-conditioned reference calculations

Let the eight independent reference chain counts be $K_j \sim \text{Bin}(L, q)$, with $L = 250,000$. The probability of total count k and a hit in every chain is the coefficient of t^k in $[(1 - q + qt)^L - (1 - q)^L]^8$. Summing coefficients for $k \geq 20$ gives the selection probability. Weighting those coefficients by $k/(8L)$ or the Wilson-containment indicator, summing over selected counts, and dividing by the selection probability gives the conditional mean or coverage.

Positive-coefficient convolution and an independent 80-digit inclusion–exclusion calculation agree on the reported B2 and B1 selected coverages within 2×10^{-16} . The count-only conditional means also satisfy the binomial tail identity in Eq. (A6). The numerical implementation retains truncation bounds and all selection denominators. Exact calculation under this independent reference does not establish the operating characteristics of the full policy with estimated ESS or of partially refreshed dependent chains.

Appendix D: Original analysis deviations and resource scope

The planned zero-event handling used interval bounds. The executed material-error path instead assigned infinite log-ratio error whenever the estimate was zero and the reference positive. This affects 1,593 replicates across ten conditions. No event interval is supplied for a constant zero indicator. We retain these observations and executed labels; the post hoc sensitivity envelope treats their disputed material-failure classifications as unknown.

The registration specified simultaneous intervals for conventional-policy rates. Execution instead constructed a bootstrap family for policy differences and marginal intervals for the rates. The original bootstrap’s nominal coverage remains unvalidated. A later conservative supplement constructs exact Clopper–Pearson bounds for 280 policy-condition rates with Bonferroni allocation at family error 0.05, then forms 140 conventional-minus-event differences. Under independent, identically distributed replicate outcomes within each condition, this supplies conservative simultaneous 95% coverage. Thirty-two difference intervals exclude zero under executed labels and 26 do so under the unknown-zero envelope. These results are post hoc and do not replace the omitted planned analysis [4–6].

The original Holm-adjusted treatment-versus-P0 analysis was omitted from execution. The later supplement contains 364 Holm-adjusted tests: 140 paired McNemar policy comparisons and 224 Fisher P0-versus-treatment comparisons. This expanded family and its test choices were specified post hoc. Ninety-two treatment comparisons have adjusted $p < 0.05$; eleven involve disputed zero-event labels, for which inferential interpretation is withheld. The full tables retain all outcomes and flags. Empirical pairing of policy decisions on the same replicate differs from treatment comparisons, whose seeds depend on different condition identities.

Resource accounting reconciles available GPU telemetry across fifteen collection sessions but retains gaps and channels without historical sensor identities. GPU board-power integration over observed intervals is neither wall-socket energy nor a complete active-compute total. No resource measurement represents a physical thermodynamic sampler. The retained companion documents coverage and exclusions independently of the probability findings.

Appendix E: Complete condition results

Table IV retains all 140 conditions as integer counts out of 256 replicates. Asterisks mark the ten conditions with disputed zero-event failure labels. False-assurance counts use executed labels; abstention is a separate policy outcome. S1–S4 follow Appendix A 1. P0 = unperturbed; Q0.05 = quantization; G-10% and G+10% = gain changes; R0.30 = configured random-number correlation. Table I defines the events.

TABLE IV. Complete condition results. All entries are integer counts out of 256 replicates. False-assurance counts retain executed labels; abstention is a separate outcome. Asterisks flag the ten conditions with disputed zero-event failure labels. Observed zeros do not establish zero underlying risk.

Event	Kernel	Conventional false assurance					Event-specific false assurance					Event-specific abstention				
		P0	Q	G−	G+	R	P0	Q	G−	G+	R	P0	Q	G−	G+	R
B1, $q = 10^{-3}$	S1	0	0	1	16	0	0	0	1	16	0	0	0	0	0	0
	S2	0	0	0	18	256	0	0	0	18	256	0	0	0	0	0
	S3	0	0	0	17	256	0	0	0	17	256	0	0	0	0	0
	S4	0	0	0	36	256	0	0	0	36	256	0	0	0	0	0
B1, $q = 10^{-4}$	S1	0	0	199	163	0	0	0	199	0	0	0	0	0	163	0
	S2	0	0	207	164	256	0	0	207	0	256	0	0	0	164	0
	S3	0	0	206	164	256	0	0	206	0	256	0	0	0	164	0
	S4	0	0	189	169	256	0	0	189	0	256	1	0	0	214	0
B1, $q = 10^{-5}$	S1	4	2	251	198	1	0	0	248	0	0	121	126	0	256	132
	S2	0	1	253	200	256	0	0	251	0	256	124	120	0	256	0
	S3	3	3	252	189	256	0	0	252	0	256	117	119	0	256	0
	S4	2	3	249	197	256	0	0	245	0	256	152	149	0	256	0
B2, positive basin	S1	0	256	0	144	0	0	9	0	0	0	0	238	0	255	0
	S2	0	256	0	161	0	0	10	0	0	0	0	235	0	256	0
	S3	256*	256	256	45*	256	0*	14	256	0*	43	256*	235	0	256*	0
	S4	248	256	0	217*	256	0	8	0	0*	1	256	240	0	256*	0
B2, deep-positive basin	S1	0	256*	0	55	0	0	0*	0	0	0	0	256*	0	256	0
	S2	0	256*	0	52	0	0	0*	0	0	0	0	256*	0	255	0
	S3	256*	256*	256	46*	256	0*	0*	256	0*	121	256*	256*	0	256*	0
	S4	248	256*	256	200*	256	1	0*	256	0*	26	255	256*	0	256*	0
B3, overlap	S1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	S2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	S3	256	256	256	256	256	256	256	256	256	256	0	0	0	0	0
	S4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B4, service loss	S1	0	0	256	249	0	0	0	256	249	0	0	0	0	0	0
	S2	0	0	256	254	256	0	0	256	254	256	0	0	0	0	0
	S3	0	0	256	251	256	0	0	256	251	256	0	0	0	0	0
	S4	0	0	256	249	256	0	0	256	249	256	0	0	0	0	0

Appendix F: Evidence and reproduction guide

The supporting archive separates the original derived observations and registered settings from exploratory references and calculations. Its original data tables remain unchanged. The additional reference inventory contains all 140 rows, numerical values where available, their derivation class, and explicit unavailable status. Selected numerical scripts accept a source-directory argument; their accompanying instructions distinguish original observations from ideal-model calculations.

TABLE V. Evidence locations within Supporting Evidence.zip. The archive contains derived records and selected scientific source; the full raw-chain archive is separate.

Path	Contents
<code>data/replicates.jsonl</code>	All 35,840 original derived replicate records.
<code>data/cells.json</code> , <code>registries/</code>	Original 140-condition summaries, model/event definitions and settings.
<code>data/supplement.json</code>	Post hoc exact-bound and adjusted-comparison supplement.
<code>exploratory/</code> <code>reference-inventory.csv</code>	Programmed, static realized, ideal stationary and available finite-horizon references; eligibility for every condition.
<code>exploratory/tables/</code>	Reference comparisons, finite-run moments, selection calculations and retained-data comparisons.
<code>exploratory/code/</code>	Selected deterministic derivations and numerical source; original sampling code remains separately preserved.
<code>exploratory/README.md</code>	Calculation scope, assumptions, numerical evidence, limitations and reproduction entry points.

Table IV reports false-assurance and abstention counts. For approval denominators, `exploratory/tables/policy/condition_policy_metrics.csv` provides `accepted`, `bad_executed`, and `replicate_count`, indexed by `cell_id` and `policy`. The `conventional` and `augmented` policy rows denote the two policies studied here; other rows describe exploratory ablations. In `data/cells.json`, `conventional_pass_rate`, `augmented_pass_rate`, and `augmented_abstention_rate` give the original policy outcomes. Event-policy failed-check counts equal the replicate count multiplied by one minus its pass and abstention rates; these differ from failed experiment executions.

The additional variance comparisons are in `exploratory/tables/observable-information.csv`; the qualified mean screen is in `exploratory/tables/global-observable-eligibility.csv`. Mean-matching horizons and their uncertainty appear in `exploratory/tables/finite-horizon/finite_horizon_uncertainty.csv`. These paths locate the full supporting values for Appendix A without replacing its assumptions or explanatory definitions.

The shared-noise references assume the specified ideal Gaussian construction and independent update calls. Exact-state reductions and quadrature determine the mathematical reference quantities; the simulation observations provide the empirical comparisons. The supporting material identifies the assumptions and numerical checks for each calculation, including the unresolved differences between predicted and observed variability.